Ferdowsi University of Mashhad

Iranian Journal of Numerical Analysis and Optimization

Volume 15, Number 2

June 2025

Serial Number: 33

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 2, June 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

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 2, June 2025

Ferdowsi University of Mashhad - Iran

Iranian Journal of Numerical Analysis and Optimization

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

* Full Professor

** Associate Professor

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

Letter from the Editor–in–Chief

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

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

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

Ali R. Soheili

Editor-in-Chief

Contents

Iranian Journal of Numerical Analysis and Optimization
Vol. 15, No. 2, 2025, pp 424–456
https://doi.org/10.22067/ijnao.2024.86859.1389
https://ijnao.um.ac.ir/

Research Article

A parallel hybrid variable neighborhood descent algorithm for nonlinear optimal control problems

M. Salimi, A.H. Borzabadi*, H.H. Mehne and A. Heydari

Abstract

In this paper, a numerical method for solving bounded continuous-time nonlinear optimal control problems (NOCPs) that based on variable neighborhood descent (VND) algorithm is proposed. First, the genetic algorithm

*Corresponding author

Received 13 February 2024; revised 13 August 2024; accepted 19 August 2024

Mitra Salimi
Department of Mathematics, Payam Noor University, Tehran, Iran. e-mail: salimi2012a@gmail.com

Akbar Hashemi Borzabadi
Department of Applied Mathematics, University of Science and Technology of Mazandaran, Behshahr, Iran. e-mail: borzabadi@mazust.ac.ir

Seyed Hamed Hashemi Mehne
Khayyam Institute, Tehran, Iran. e-mail: hmehne@ari.ac.ir

Aghileh Heydari
Department of Mathematics, Payam Noor University, Tehran, Iran. e-mail: a_heidari@pnu.ac.ir

How to cite this article
Salimi, M., Borzabadi, A.H., Mehne, H.H. and Heydari, A., A parallel hybrid variable neighborhood descent algorithm for nonlinear optimal control problems. *Iran. J. Numer. Anal. Optim.*, 2025; 15(2): 424-456. https://doi.org/10.22067/ijnao.2024.86859.1389

(GA) is combined with an improved VND that uses efficient neighborhood interchange. Then, to improve the efficiency of the algorithm for practical and large-scale problems, the parallel processing approach is implemented for discrete form of NOCP. It performs the required complex computations in parallel. The resulting parallel algorithm is applied to a benchmark of nine practical problems such as Van Der Pol problem and chemical reactor problem. For large-scale problems, the parallel hybrid variable neighborhood descent algorithm (PHVND) is capable of obtaining optimal control values effectively. Our experimentation shows that PHVND outperforms the best-known heuristics in terms of both solution quality and computational efficiency. In addition, computational results indicate that PHVND produces superior results compared to sequential quadratic programming or GA.

AMS subject classifications (2020): Primary 49M25; Secondary 65Y05, 68W50.

Keywords: Parallel algorithm; Neighborhood descent algorithm; Optimal control problem.

1 Introduction

The main goal of optimal control is to find a control for a linear or nonlinear system that will makes the system satisfy physical constraints while also minimizing or maximizing some performance measure. This involves determining the best control inputs over a given time horizon to achieve the desired system behavior [6]. Optimal control problems (OCPs) are utilized to provide an explanation for problems in various industrial processes, such as the control of robots, aircraft, and electrical systems [4]. An optimal control strategy was proposed for dynamic transfer from a simple half-circular legged monopod model in [27]. The development of the nonlinear optimal control problems (NOCP) method for reverse osmosis desalination devices with an induction motor (IM) to minimize energy dissipation and reduce the operating cost of the units is reviewed in [29]. In [28], a nonlinear optimal control approach is proposed for the dynamic model of a gas centrifugal compressor driven by an IM. One of the other applications of NOCPs to human health is

reported in [2], where NOCP is used to determine the dynamics of hepatitis E diseases.

The study of numerical methods has provided an attractive field for researchers in mathematical sciences who have seen the emergence of different numerical calculation methods and efficient algorithms to solve OCPs [25]. Nowadays, meta-heuristic approaches have gained popularity in solving NOCPs. Researchers applied the particle swarm optimization (PSO), for example, in [12], to optimize the parameters of a proportional integral derivative (PID) controller. The application of ant colony optimization (ACO) is another meta-heuristic for numerically solving NOCPs. The method uses the discrete form of the control values of the set. ACO solved the problems that have transformed into quasi-assignment problems [5, 19]. The genetic algorithm (GA) method is a population-based meta-heuristic. It was widely used to solve many optimization problems. In recent decades, GA has been also proposed to solve NOCP effectively [21, 30]. Continuous GA and improved GA to solve NOCP can be found in [1, 33].

To improve the precision of the solutions and reduce the computational time, hybrid methods were introduced [3, 20]. In [23], a modified hybrid GA has been proposed to solve NOCP numerically. A hybrid improved GA using a simple local search method is also provided to solve NOCP [34].

One of the advantages of the application of meta-heuristics for solving NOCP is their independence to near-optimal initial guesses. So, one approach to solving an NOCP with meta-heuristics is changing the NOCP into a nonlinear programming (NLP) problem with finite dimension. It is performed by discretizing the space-time and control, which leads to a large-scale problem.

After the discretization of NOCP, a new initiative is proposed to solve the problem more easily and faster. Due to the structural similarity of the discretization problem with the uncapacitated single allocation p-hub median problem (USApHMP), we will use this method to solve NOCP [32]. Due to the simple nature of USApHMP, the NOCP optimization procedure is easily programmed and can be solved effectively. It has many benefits, such as ease of programming, short computation time, self-starting, and convergence to the global minimum.

The large dimension of large-scale NLP is a limitation that makes it difficult to find accurate solutions, especially with population-based algorithms. Parallelization algorithms are used to speed up a wide range of algorithms, for example, the ACO method, which is applied to solve NOCP, [19]. GA uses hybridization, strategic search [31], and parallelization to reach high-quality solutions in a low computational time. In [36, 37, 38], a parallel GA for optimization is described. To reduce the computational time in solving large-scale problems, parallelism is proposed in this paper.

One of the effective methods to solve USApHMP is variable neighborhood descent (VND) [17]. It has been applied to solve the maximum clique problem [16], scheduling problem [7], and hub location [26].

To find a promising region of search space, we use four local searches in sequential VND (SVND). Then, to increase the quality of solutions obtained from the phase search, the number of control partitions in NOCP is changed as hubs in USApHMP [32]. Finally, we combine SVND with the GA in mutations and reduce computational time with parallel processing.

The parallel hybrid variable neighborhood descent algorithm (PHVND) is then evaluated by implementing some real-world examples. The results show that the proposed method obtains more accurate solutions compared with methods that search for control values in discrete form.

The remainder of this paper is organized as follows: Section 2 provides a brief overview of the formulation problem. Section 3 describes the basics of the PHVND algorithm. In Section 4, we examine some benchmark examples to compare the results of the proposed algorithm with the other existing methods. Finally, Section 5 presents more analysis of the results and concludes the paper.

2 Mathematical programming formulation

The goal of NOCP is to control a dynamical system from a given initial state to a final optimal state in a desirable way. Find an optimal controller $u^*(t)$ and state variable $x^*(t)$ that minimize the cost function $J(x, u)$, where $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$ are the state variable and the control variable respectively. Suppose that $g : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ and $f : \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ are both

continuously differentiable functions in x and u, t_0, t_f are the initial and final times, respectively, and that $\phi : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ is a continuous scaler function. The Bolza problem is minimizing

$$J(x(\cdot), u(\cdot)) = \phi(x(t_f), t_f) + \int_{t_0}^{t_f} g(t, x(t), u(t))dt, \tag{1}$$

under the following constraints and boundary conditions:

$$\dot{x} = f(t, x(t), u(t)), \tag{2}$$
$$c(t, x, u) = 0, \tag{3}$$
$$d(t, x, u) \leq 0, \tag{4}$$
$$\psi(t_f, x(t_f)) = 0, \tag{5}$$
$$x(t_0) = x_0, \tag{6}$$
$$u_l \leq u \leq u_b, \tag{7}$$

where all functions are assumed to be smooth enough in suitable open sets. The cost function (1) should be minimized according to the dynamics (2), control and restrictions of state equality (3) and control inequality and state inequality constraints (4), final state constraints (5), initial condition (6), and limitation of the control variable (7).

Minimum time problems, linear quadratic regulator (LQR) problem, tracking problem, minimum energy, and terminal control problem are another special case of NOCPs.

3 Parallel hybrid algorithm

Representation scheme. At first, the problem of minimizing (1) subject to (2)–(7) is changed to a discrete form. To do this, the time interval is divided to a number of equidistant nodes $\{t_0, t_1, \ldots, t_{N_t-1}\}$. For any interval $[t_{i-1}, t_i], i = 1, 2, \ldots, N_t - 1$, the interval $[u_l, u_b]$ is divided into m subinterval $[u_{j-1}, u_j], j = 1, 2, \ldots, m$ with the same length. Then, p values are determined as hub $(u_0, u_1, \ldots, u_p)$ in USApHMP from m subinterval. Therefore, in direct form of the problem, we have to find an assignment

$$\{t_0, t_1, \ldots\} \to \{u_0, u_1, \ldots\} \tag{8}$$

with minimum performance index. The new problem is an assignment type of optimization problems with a finite dimension. Since the control interval is divided into m partitions and the time interval is divided into $N_t - 1$ sections, and one control must be selected for each time. Then, this problem can be extended to USApHMP, where every time is assigned to only one hub from the set of $u_0, u_2, \ldots, u_p$. (see details in [32])

Fitness function. Now, a piecewise control function is constructed as follows:

$$u(t) = \sum_{j=1}^{N_t - 1} \chi[t_{j-1}, t_j](t) u_j, \tag{9}$$

where χ is the indicator function:

$$\chi[t_{j-1}, t_j](t) = \begin{cases} 1, & t \in [t_{j-1}, t_j], \\ 0, & otherwise. \end{cases} \tag{10}$$

The time nodes are defined by

$$t_j = t_0 + jh, \qquad j = 0, \ldots, N_t - 1, \tag{11}$$

where $h = (t_f - t_0)/N_t - 1$. This is a finite-dimensional optimization problem with $u_1, \ldots, u_p$ as unknowns.
Let $\psi = [\psi_1, \psi_2, \ldots, \psi_{n_\psi}]$ is the vector form of constraints, defined in (5). Then satisfaction of constraints can be relaxed as:

$$\varphi_f = \|\psi\| < \epsilon, \tag{12}$$

where $\|\cdot\|$ is the Euclidean norm and ϵ is a given small number indicating the accuracy. The term integral of (1) can be calculated by a numerical integration such as Runge-Kutta or Newton algorithms for each control input $u(t_j)$, and the corresponding state variable $x(t_j), j = 0, 1, \ldots, N_t - 1$, which satisfy in (2). The notations $x(t_j)$ and $u(t_j)$ are briefly indicated by x_j and u_j respectively.

Then, the discrete form of the performance index (1) is as follows:

$$\text{minimize} \quad J(x(\cdot),u(\cdot)) = \sum_{j=0}^{N_t-1} w_j g(t_j, x(t_j), u(t_j)) + M\varphi_f, \tag{13}$$

where $w_j, j = 0,\ldots,N_t-1$ are weights and M is a large enough positive number. In the case of equality or inequality constraints (3) or (4), penalty constants $M_1, M_2, M_{31}, \ldots, M_{3n_\psi}$ are added to the fitness function:

$$\begin{aligned} I = \tilde{J} &+ \sum_{j=0}^{N_t-1} M_1 \max\{0, d(x_j,u_j,t_j)\} + \sum_{j=0}^{N_t-1} M_2 c^2(x_j,u_j,t_j) \\ &+ \sum_{k=1}^{n_\psi} M_{3k}\psi_k^2(x_{N_t-1}, t_{N_t-1}), \end{aligned} \tag{14}$$

where $\tilde{J}$ is an approximation of J.

In the following, the GA is introduced in the first subsection and the SVND algorithm in the next subsection. Then, the structure of the hybrid PHVND algorithm is explained and then the parallelization process is mentioned.

3.1 Genetic algorithm

Initialization. The initial population is chosen randomly from candidate solutions in the search space. Then, p control input values are selected from the m control subintervals at random. This can be done in the following stages. In our experiment, the population size is fixed at a value of 300 ($N_{pop} = 300$). Then, each time $t_j, j = 0,\ldots,N_t-1$ is randomly assigned to the selected control values. With this selection pattern, initial chromosomes are produced. When this repeats N_{pop} times, then N_{pop} feasible chromosomes $\{C_1, C_2, \ldots, C_{N_{pop}}\}$ are generated.

Selection. For NOCP, the better chromosome is one with a smaller fitness value. In the first, a certain number of individuals in the population are randomly selected. Using this method of selection, we will have a great chance to find better random chromosomes to produce children. This phase can be also performed in parallel.

Elitism. The incumbent from generation to generation is maintained. If in crossover or mutations, the best chromosome (individual) within the population is lower than the existing solution, the better answer is replaced by the worse one in the current population.

Crossover. The crossover is a combination of two or more parental solutions to provide better solutions. There are several ways to achieve this, and good performance depends on a well-designed recombination mechanism [35]. Here, the standard one-point crossover operator is proposed, where a crossover operator creates a child solution for selected parents. The crossover parents are selected from $\{C_1, C_2, \ldots, C_{N_{pop}}\}$ and random j from $\{0, \ldots, N_t - 1\}$.

To construct a child solution, its first j genes are taken from the first parent, and the remaining $N_t - 1 - j$ genes are taken from the second parent. After applying the crossover on the arrays, if the number of assignments control is more than p in the offsprings, we keep the feasible one, then for the remaining gens if a time node t_i was assigned to another control from $\{u_{p+1}, \ldots, u_{N_u}\}$, then the node t_i is randomly assigned to a selected $\{u_1, \ldots, u_p\}$.

Mutation. The modification of the mutation method preserves the entire population. Repeat the selection to the mutation process until a termination condition is met.

3.2 Local searches

The fundamental neighborhood structures can also be extended for mutations in GA exchanges and node interpolation moves. Local search methods are as follows:

1- First, we randomly choose one person in the population, $pop(i)$ and randomly choose gen position $j_1 \in \{1, \ldots, N_t - 1\}$ and $j_2 \in \{j_1, \ldots, N_t - 1\}$. Then, exchange position of j_1 and j_2 and replace if the new offspring is better than the worst person in the population. This means that we select

(t_{j_1}, u_{j_1}) and (t_{j_2}, u_{j_2}) then exchange u_{j_1} with u_{j_2} and vice versa.

2- The second, randomly choosing one person, $pop(i)$, then the local search method is as follows. This local modification does not change the location of the selected controls $(u_1, \ldots, u_p)$. For all time nodes, we try to change their membership. This means that we have to process every selected control and try to replace the assignment time node with a new assignment. This does not change other elements.

3- The location of selected controls u_j (which in control array $\{u_1, \ldots, u_p\}$) is changed in $pop(i)$ that choose randomly and move to this neighborhood solution if the cost function is reduced. First, a control u_k that is not in control array, is selected. From the representation of the current assignment, all nodes times that assignment to u_j can generate and then change its assignment to u_k. But if the $pop(i)$ has only one control, do nothing. Otherwise, the change in the cost function for all populations is calculated. We developed four neighborhood structures independently, although one of them is essentially the same as the second local search method (Greassign and Greplace) in [39] and one and the second are the same as local search in [17], but these neighborhoods are applied in a different order.

4- The one parent $pop(i)$, u_j from the control array and u_k that is not in the control array was selected, randomly. The u_j was deleted and u_k was replaced and randomly reassigned all time nodes in $pop(i)$ to new control that includes u_k. Thus all assignments in $pop(i)$ are changed.

3.3 SVND

Among available strategies for the exploration of several neighborhoods within deterministic local search, the sequential strategy is more common. In VND, the neighborhood structures are explored one by one in the given sequence that is specified in Algorithm 1 (SVND) [14, 15]. SVND uses all $l_{\max}$ neighborhood structures within VND. Then it performs local searches with them, considering their orders. SVND stops whenever no better solution in the last neighborhood exits.

Algorithm 1 Sequential variable neighborhood descent

Require: Make an order of all $l_{\max} \geq 2$ neighborhoods that will be used in the search, select the set of neighborhood structures $N_l, l = 1, \ldots, l_{\max}$, an initial solution x.

$x_{opt} = x$, $f_{opt} = f(x)$, $l = 1$
repeat
  $i = 0$
  repeat
    $i = i + 1$
    find best neighbor $x^{'} = argmin\{f(x), f(x_i)\}, x_i \in N_l(x)$
  until all solutions from neighborhood l are visited set
  if solution $x^{'}$ thus obtained is better than x **then**
    $x = x^{'}$, $x_{opt} = x$, $f_{opt} = f(x)$ and $l = 1$
  else
    $l = l + 1$
  end if
until $l > l_{\max}$
return x

The last obtained solution is a local minimum with considering all $l_{\max}$ neighborhood structures. After construction or update of the population, the fitness values of the candidate solutions are evaluated using (14). Here, four local search neighborhoods are used for each generated individual to refine the allocation of time nodes to the control array. It will be explained in the next subsection.

3.4 Parallelization

The size of the population is an important factor that affects the scalability and performance of the genetic algorithm. However, increasing the size of the population will increase the computational time. To solve this problem, a parallel implementation is proposed. It will reduce the time and improves the ability of solving complex problems. Due to independent operations in

a large amount of data, population-based optimization algorithms can run in parallel. The first stages of discretization and population generation are serial. Then, the iteration of GA including mutation, fitness evaluation and crossover are implemented as loop parallelization.

The process is then repeated until the maximum allowed number of iterations is reached. The parallel algorithm proposed by PHVND is effective in solving the test criteria for the large-scale NOCP algorithm. The algorithm of PHVND to solve NOCP is given as Algorithm 2.

In PHVND, GA is combined with the SVND method in the mutation step as local search to improve the solutions.

Algorithm 2 The algorithm of the PHVND method

Require: Select the number of time steps, N_t, the number of control functions, N_u, the size of the population, N_{pop}, the maximum number of iterations $Maxiter$, the mutation factor, p_m, the crossover constant, p_c, and the number of the available processors, N_p.

Generate a random initial population P_A and distribute it among processors by master processors.

let $iter = 0$

while $(iter < Maxiter)$ **do**

Selection: Each processors receives its population, P_B, and calculates the fitness function of its population.

Elitism: Each processors selects the best individual among P_B and send them the master processor.

The master processor determines the best global individuals, send them to processors, and updates the P_A.

Each processor receives the global best individuals and replace them with the worst individual in the P_B.

Crossover: Each processors performs a crossover for P_B

Mutation: Each processors executes SVND for P_B

let $iter = iter + 1$.

end while

Master processor returns the best individual in the final population (P_A) as an approximate solution of NOCP.

4 Computational results for NOCP

To examine the efficiency of the proposed algorithm, some benchmark problems of NOCPs are considered in some examples. The absolute error of the performance index, satisfaction of the problem's constraints, and the effect of particle size of control times and intervals are parameters for evaluating the algorithms.

The codes are developed using MATLAB software and the OpenMP programming interface. They were executed on a parallel computer at the Aerospace Research Institute with 48 (1.8 GHz) cores.
Let J be the performance index obtained by an algorithm, let φ_f define the error of final state constraints, and let J^* be the best solution obtained among all implementations, or the exact solution (when exists). Now, let N_t and m be defined as the division of time and control interval, and let the absolute error of J, gap_J, of the algorithm be defined by

$$gap_J = |J - J^*|. \tag{15}$$

The parameters of PHVND are $l_{\max} = 4$, u_{left}, u_{right}, t_0, and t_f. We consider $p_m = 0.4$ and $p_c = 0.6$. The other parameters of PHVND are given in what follows. To find the best value for the algorithm parameters, we ran the algorithm with different parameter values and then chose the best. However, the sensitivity of the parameters N_t, m are studied in the tables. Table 1 indicates the details of each case.

4.1 Comparison of results

The results on instances of partition time and control function are summarized in Table 12 where PHVND is compared with Modified Hybrid Genetic (see [24]), GA (see [11]), sequential unconstrained minimization technique (SUMT) (see [10]), sequential quadratic programming (SQP) (see [10]), hybrid algorithm by integrating an improved PSO with improved particle swarm optimization (IPSO-SQP)(see [22]), novel continuous genetic algorithm (CGA) (see [1]), numerical method proposed in [13] called Bézier, ho-

motopy perturbation method (HPM) (see [9]), parallel numerical method based on whale optimization algorithm (WOA) (see [20]), Linear interpolation (LI)(see [24]), and Spline interpolation (SI)(see [24]).

Comparison the best numerical results in 300 different runs of the PHVND method for different partitions of time and control function for each exam-

Table 1: The problem parameters for NOCPs examples

	$g = \frac{1}{2}(x_1^2 + x_2^2 + u^2)$	$x_0 = (1, 0)$	
Example 1	$f = \left[x_2, -x_2 + (1 - x_1^2)x_2 + u\right]^T$	$t = [0, 5]$	$M = 10^2$
	$\psi = x_1 - x_2 + 1$	$u = [-0.5, 2]$	
	$g = \frac{1}{2}(x_1^2 + x_2^2 + 0.1u^2)$	$x_0 = (0.05, 0)$	
Example 2	$f = [x_1 - 2(x_1 + 0.25) + (x_2 + 0.5)\exp\left(\frac{25x_1}{x_1+2}\right)$	$t = [0, 0.78]$	$M = 1$
	$-u(x_1 + 0.25), 0.5 - x_2 - (x_2 + 0.5)\exp\left(\frac{25x_1}{x_1+2}\right)]^T$		
	$\psi = [x_1, x_2]^T$	$u = [-1.5, 2]$	
	$g = u^2$	$x_0 = 0$	
Example 3	$f = \frac{1}{2}x^2\sin(x) + u$	$t = [0, 1]$	$M = 10^6$
	$\psi = x - 0.5$	$u = [0, 1]$	
	$\phi = -x_3$	$x_0 = (10, -2, 10)$	
Example 4	$f = [x_2, -2\frac{u}{x_3}, -0.01u]^T$	$t = [0, 5]$	$M = 1$
	$\psi = [x_1, x_2]^T$	$u = [-30, 30]$	
	$\phi = x_1$	$x_0 = (0, 1)$	
Example 5	$f = [-620000e^{\left(\frac{-5000}{u}\right)}x_1 + 4000e^{\left(\frac{-2500}{u}\right)}x_2^2,$	$t = [0, 1]$	
	$-4000e^{\left(\frac{-2500}{u}\right)}x_2^2]^T$	$u = [298, 398]$	
	$\phi = x_3$	$x_0 = (0, -1, 0)$	
Example 6	$f = [x_2, -x_2 + u, x_1^2 + x_2^2 + 0.005u^2]^T$	$t = [0, 1]$	$M = 1$
	$\psi = (x_1, x_2)$	$u = [-5, 5]$	
	$d(x, t) = x_2 + 0.5 - 8(t - 0.5)^2$		
	$g = 2x_1$	$x_0 = (2, 0)$	
Example 7	$f = [x_2, u]^T$	$t = [0, 3]$	$M = 1$
	$d(x, t) = -(6 + x_1)$	$u = [-2, 2]$	
	$g = x_1^2 + x_2^2 + 0.005u^2$	$x_0 = (0, -1)$	
Example 8	$f = [x_2, -x_2 + u]^T$	$t = [0, 1]$	$M = 10^2$
	$d(x, t) = -(8(t - 0.5)(t - 0.5) - 0.5 - x_2)$	$u = [-5, 5]$	
	$g = \frac{1}{2}(x_1^2 + x_2^2 + u^2)$	$x_0 = (1, 0)$	
Example 9	$f = [x_2, -x_1 + (1 - x_1^2)x_2 + u]^T$	$t = [0, 5]$	$M = 1$
	$d(x, t) = -(x_2 + 0.25)$	$u = [-1, 1]$	

ples are proposed in Table 2 and Tables 4–11. In the tables related to each example, such as Table 2 or Tables 4–11, the value of gap_J is the result of comparing the value of J for different partitions with the optimal value of PHVND. In Table 12, the value of gap_J is the result of comparing the best value of PHVND with other algorithms for each example.

4.2 Numerical Experiments

Example 1. This example is Van Der Pol Problem (VDP) [1], which has to be minimized. After scaling by taking the best-known value reached by PHVND, we obtain the following values: $J^* = 1.5404$ for $N_t = 100$ and $m = 100$. The numerical result for different N_t, m is given in Table 2. We have also compared our result with other algorithms in Table 12. The results objective function of implementing PHVND is less than the result of CGA, which equals $J^* = 1.7404$ and less than LI, $J^* = 1.5949$, and SI, $J^* = 1.5950$.

The error of proposed method is small around $\epsilon = 10^{-4}$. The value of speedup in Table 3 for parallel computing, shows the performance of parallelization and reduction time in terms of number of CPUs. The resulting control and states of Example 1 obtained from using the PHVND method for $(N_t, m) = (100, 100)$ are given in Figures 1 (a), 1(b). Figure 1 (c) shows the performance index for 300 iterations and indicates the convergence rate.

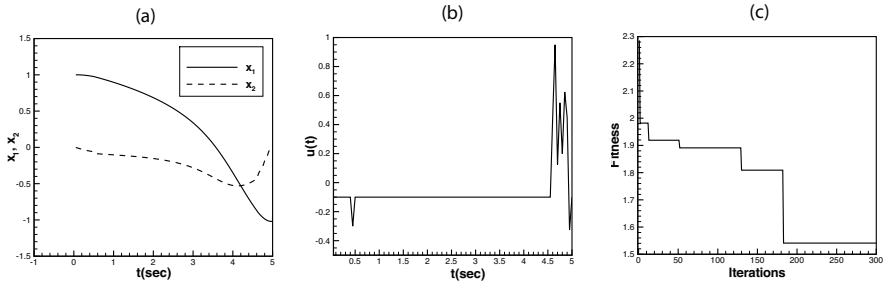


Figure 1: (a) The resulting optimal trajectories for $N_t = 100$, $m = 100$ for Example 1, (b) The resulting optimal control for $N_t = 100$, $m = 100$ for Example 1, (c) Performance index for iterations for Example 1 ($N_t = 100$, $m = 100$)

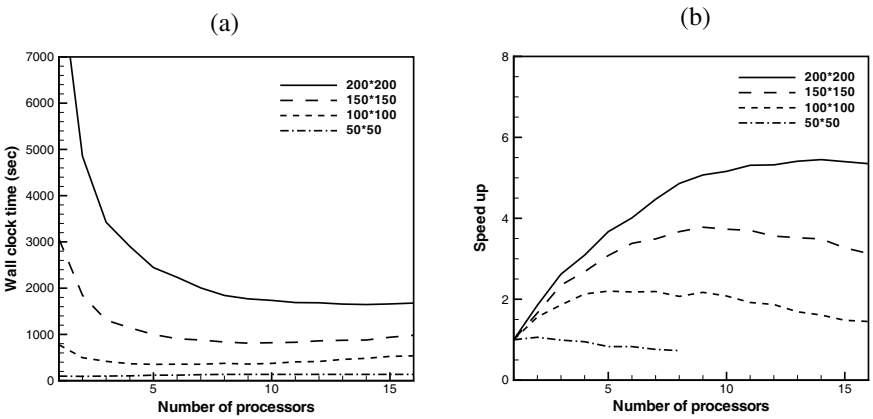


Figure 2: (a) The wall clock time of execution for Example 1, and (b) The speedup curve for Example 1.

In addition, Figure 2(a) shows the wall clock time of running, which shows that run time is decreased with increase the number of processors. The speed of the performance curve is also given in Figure 2(b) and compared with the ideal speedup. The resulting speedup curve is close to the ideal line up to 8 threads and its divergent behavior is normal after 8 threads because of the serial fraction of the algorithm and other overheads.

Based on the figures, the efficiency of the method in controlling the system to desired final states, rate of convergence and parallel performance can be deduced.

Example 2. The Chemical Reactor Problem (CRP) [1] is solved in this example. The proposed algorithm was applied to this problem where the results are given in Table 4.

The results have been compared with CGA, LI, and SI algorithm in Table 12. Although, in Table 12, the norm of final state constraints, φ_f, for CGA, equals 7.57×10^{-10} and for the LI and SI methods, which equal 1.15×10^{-9} and 5.99×10^{-9}, respectively, is less than φ_f of the PHVND methods, which equals 3.60×10^{-3}, but the performance index gap is very small, about 0.008 for CGA and 0.01 for LI and SI.
After solving with the proposed method, we obtain the following results: the best value $J^* = 0.033$ and the error function $\varphi_f = 3.60 \times 10^{-3}$ that obtained

Table 2: Comparison between the best numerical results of different partitions for Example 1

N_t	m	J	φ_f	gap_J
10	10	1.5659	2.19×10^{-3}	0.025
30	50	1.6299	2.85×10^{-3}	0.089
30	100	1.5811	3.12×10^{-3}	0.047
50	50	1.5966	9.06×10^{-4}	0.056
50	100	1.5689	4.14×10^{-4}	0.028
75	75	1.661	7.15×10^{-4}	0.120
100	100	1.5404	1.75×10^{-4}	0.000
150	150	1.7587	1.00×10^{-3}	0.218

Table 3: Comparison the results of PHVND for different partitions of time and control function for 16 number of threads for Example 1

Number of	$N_t = 100$, m=100		$N_t = 150$, m=150		$N_t = 200$, m=200	
threads	time (sec)	Speed up	time (sec)	Speed up	time (sec)	Speed up
1	778	1.00	3070	1.00	8967	1.00
2	497	1.56	1846	1.66	4849	1.85
3	418	1.86	1309	2.35	3425	2.62
4	366	2.13	1144	2.68	2904	3.09
5	353	2.20	998	3.08	2445	3.67
6	356	2.18	907	3.38	2235	4.01
7	355	2.19	879	3.49	2006	4.47
8	375	2.07	836	3.67	1843	4.86
9	359	2.17	811	3.78	1769	5.07
10	374	2.08	821	3.73	1736	5.16
11	406	1.92	830	3.70	1689	5.31
12	415	1.87	861	3.56	1684	5.32
13	460	1.69	873	3.52	1656	5.41
14	483	1.61	879	3.49	1646	5.45
15	525	1.48	939	3.27	1659	5.40
16	537	1.45	982	3.13	1678	5.35

Table 4: Comparison between the best numerical results of different partitions for Example 2

N_t	m	J	φ_f	gap_J
10	10	0.041	7.93×10^{-3}	0.007
30	50	0.033	3.60×10^{-3}	0.000
30	100	0.034	7.46×10^{-4}	0.000
50	50	0.045	5.09×10^{-3}	0.011
50	100	0.042	1.53×10^{-3}	0.008
75	75	0.059	1.02×10^{-2}	0.025
100	100	0.058	1.25×10^{-2}	0.024
150	150	0.060	1.11×10^{-2}	0.026

for $N_t = 30$ and $m = 50$. The resulting control and states of Example 2 obtained by the PHVND method are given in Figures 3(a) and 3(b). Furthermore, the speed-up curve is also shown in Figure 3(c) and compared with the ideal speedup. Figure 4(a) indicates the convergence rate of the performance index for 300 iterations.

The results shown in Figures 3 and 4(a) specify that the method has the ability to find nearly optimal control-trajectories with a reduction in time with the aid of parallel processing.

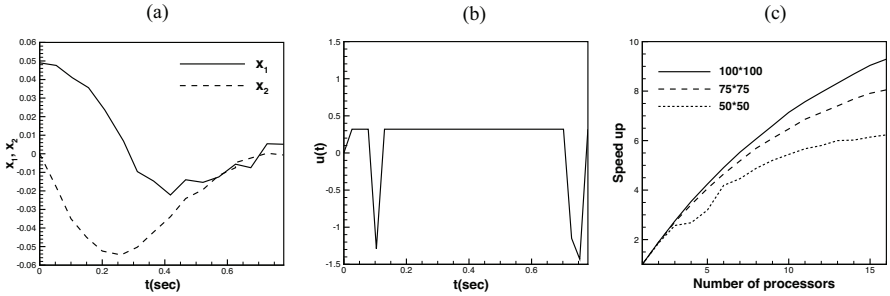


Figure 3: (a) The resulting optimal trajectories ($N_t = 16$, $m = 20$), (b) The resulting optimal control ($N_t = 30$, $m = 50$), and (c) The speedup curve of Example 2.

Example 3. This example [9], is a constraint nonlinear model. After solving this problem, the following results: the best value $J^* = 0.1825$ and the error function $\varphi_f = 6.81 \times 10^{-7}$ that obtained for $N_t = 100$ and $m = 100$. The

convergence rate of the performance index for 300 iterations is shown in Figure 4(b). The numerical results for different N_t, m are given in Table 5.

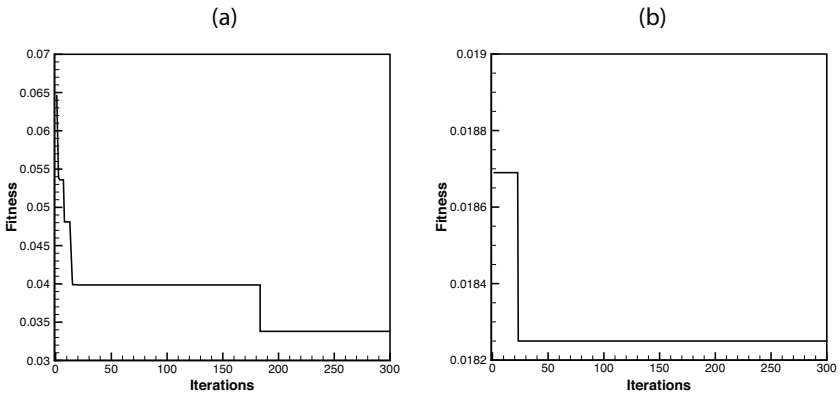


Figure 4: (a) Convergence rate of the performance index for Example 2 ($N_t = 30$, $m = 50$), and (b) Convergence rate of the performance index for Example 3 ($N_t = 100$, $m = 100$).

Table 5: Comparison between the best numerical results of different partitions for Example 3

N_t	m	J	φ_f	gap_J
10	10	0.2547	2.21×10^{-7}	0.072
30	50	0.1952	1.2×10^{-6}	0.012
30	100	0.2100	3.17×10^{-6}	0.027
50	50	0.1877	1.33×10^{-7}	0.005
50	100	0.1906	8.06×10^{-7}	0.013
75	75	0.1869	1.52×10^{-6}	0.004
100	100	0.1825	6.81×10^{-7}	0.000

From Table 5, it is obvious that the gap between the cost function and the best one for different values of m and N_t is less than 0.07, indicating that the objective functions in PHVND do not differ a lot. Although, from Table 5, the norm of final state constraints, φ_f, for our method, is equal to 10^{-7} or 10^{-6}. The numerical results of the proposed algorithm are compared with algorithms: GA in [11], LI, SI, and HPM. For different n, m, the performance

index, J,in our method is also less than $J^* = 0.2015$ which was proposed in [24] and $J^* = 0.3526$ which was taken by GA in [11].

Example 4. This problem [10] is to control the performance index with minimum effort. The problem is compared with several numerical methods, see Table 12. The best optimal objective was $J^* = -8.8157$ for $N_t = 30, m = 50$ (see Table 6) which is close to the result of LI and SI,-8.8692 or SQP and SUMT, -8.8690. The norm of final state constraint for LI and SI equals 1.45×10^{-10} and 2.79×10^{-10} and this criterion for the PHVND methods equals 4.8×10^{-4}. By increasing the repetition in the algorithm, better solutions can be obtained, as the convergence rate of the performance index for 300 iterations is shown in Figure 5(a).

Table 6: Comparison between the best numerical results of different partitions for Example 4

N_t	m	J	φ_f	gap_J
10	10	-8.3931	0.586	0.422
30	50	-8.8157	0.036	0.000
30	100	-8.7597	0.075	0.056
50	50	-8.7945	0.063	0.021
50	100	-8.7829	0.051	0.032
75	75	-8.8074	0.036	0.008
100	100	-8.8099	0.042	0.005

Example 5. The problem of dynamic optimization of consecutive reaction batch reactor ($A \to B \to C$) is a classical problem studied by several researchers [41, 8, 34]. The aim is to find the optimal temperature profile that maximizes the product performance of B temperature at the result of the process in the batch reactor. The state constraints and initial conditions are given in Table 1, where x_1 shows the concentration of B, x_2 represents the concentration of A, and u is the temperature.

To compare the performance of PHVND with existing methods, the comparison results are given in Table 7. The maximum value by applying PHVND is $J^* = 0.717$ was obtained very rapidly. The computation time with a discretization of $N_t = 9, m = 10$ subintervals, was 27.7 seconds.

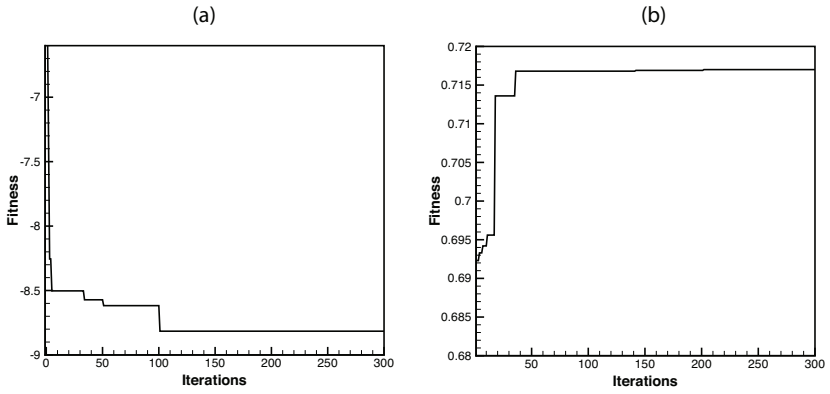


Figure 5: (a) Convergence rate of the performance index for Example 4 ($N_t = 30$, $m = 50$), and (b) Performance index versus iterations number in Example 5 ($N_t = 9$, $m = 10$).

Table 7: Comparison between the best numerical results of different partitions for Example 5

N_t	m	J	gap_J
9	10	0.717	0.000
10	10	0.677	0.04
30	50	0.612	0.105
30	100	0.612	0.105
50	50	0.609	0.108
50	100	0.609	0.108
75	75	0.608	0.109
100	100	0.607	0.11

In addition, we present our approach to the optimal control in Figure 5(b) and the results from the approach of [40] and [18] are 0.6107 and 0.6128 that exceeded by our method to 0.717 that shown in Table 12.

Example 6. This problem [22] contains an inequality constraint with minimum control effort. After solving with the proposed method, the best value $J^* = 0.2746$ was obtained for $N_t = 10$ and $m = 10$.

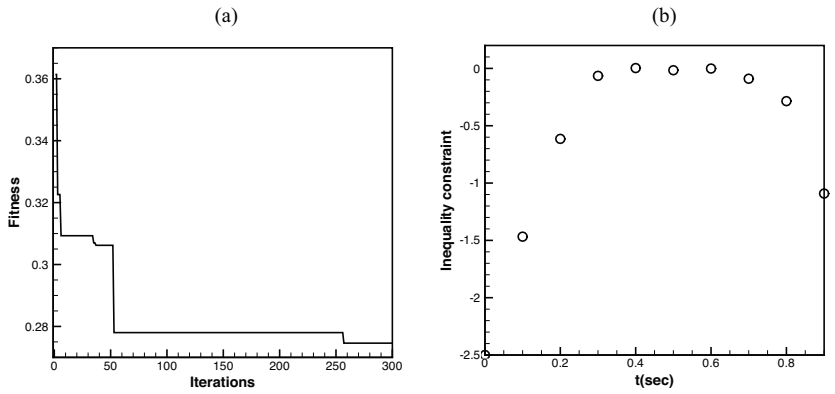


Figure 6: (a) Convergence rate of the performance index, and (b) the inequality constraint of Example 6 ($N_t = 10$, $m = 10$).

To see the convergence of the example in the 300 iterations, Figure 6(a) is presented, and Figure 6(b) shows an inequality constraint for 300 iterations. Table 8 shows the numerical results for different N_t, m.

Table 8: Comparison between the best numerical results of different partitions for Example 6

N_t	m	J	gap_J
10	10	0.2746	0.000
30	50	0.2797	0.005
30	100	0.2796	0.005
50	50	0.2839	0.009
50	100	0.2771	0.002
75	75	0.2871	0.012
100	100	0.2931	0.018

Example 7. This problem [13] contains an inequality constraint that should be minimized. After solving with the proposed method, the best value $J^* = -8.999$ was obtained for $N_t = 75$ and $m = 75$.

The convergence rate of the performance index for 300 iterations is given in Figure 7(a) and inequality constraint d is displayed in Figure 7(b). The numerical results for different N_t, m are compared in Table 9. The perfor-

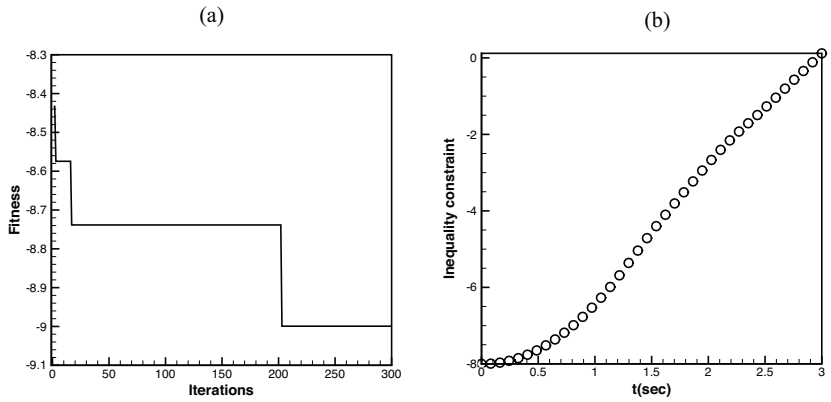


Figure 7: (a) The performance index versus iterations number, and (b) The inequality constraint of Example 7 ($N_t = 75$, $m = 75$).

Table 9: Comparison between the best numerical results of different partitions for Example 7

N_t	m	J	gap_J
10	10	-7.391	1.608
30	50	-8.885	0.114
30	100	-8.574	0.425
50	50	-8.917	0.082
50	100	-8.843	0.156
75	75	-8.999	0.000
100	100	-8.860	0.139

mance index in Table 12, J, for different N_t, m in our method is also low than $J = -5.3898$ by Bézier and $J = -5.4309$ by LI and SI.

Example 8. This problem [10] contains an inequality constraint,that J has to be minimized. After solving with the proposed method, the best value $J^* = 0.2646$ was obtained for $N_t = 30$ and $m = 50$.

The convergence rate of the performance index for 300 iterations is shown in Figure 8(a) and inequality constraint d for 300 iterations is shown in Figure 8(b). The numerical result for different N_t, m is given in Table 10. From Table 10, it is obvious that for different values of m and N_t, the gap of the

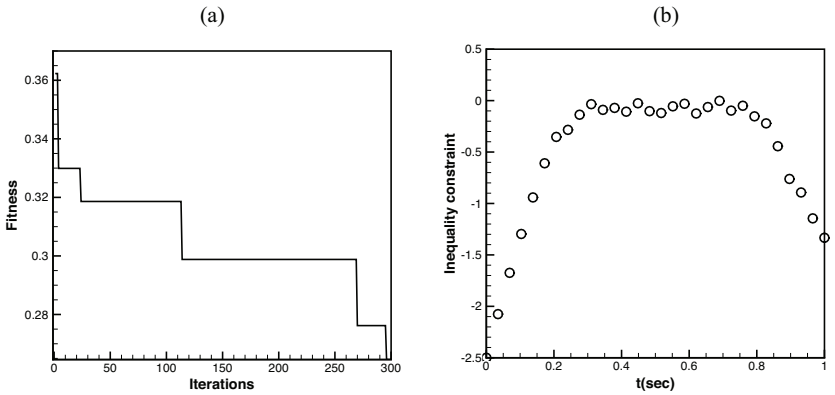


Figure 8: (a) The performance index versus iterations number, and (b) The inequality constraint of Example 8 ($N_t = 30$, $m = 50$).

Table 10: Comparison between the best numerical results of different partitions for Example 8

N_t	m	J	gap_J
10	10	0.2818	0.017
30	50	0.2646	0.000
30	100	0.3019	0.037
50	50	0.2846	0.02
50	100	0.2748	0.01
75	75	0.2755	0.01
100	100	0.2818	0.017

cost function from the best one, is less than 0.03, indicating that the objective functions in PHVND do not differ a lot. The performance index, J, for different N_t, m in our method is also do not differ a lot than $J = 0.2163$ by SQP and $J = 0.1703$ which was obtained by SUMT and $J = 0.1549$ by LI and SI. This difference will decrease with increasing repetition and will converge to the optimal solution.

Example 9. The aim of this problem [10], is minimizing the performance index. After solving with the proposed method, the best value $J^* = 1.476$ that obtained for $N_t = 30$ and $m = 50$.

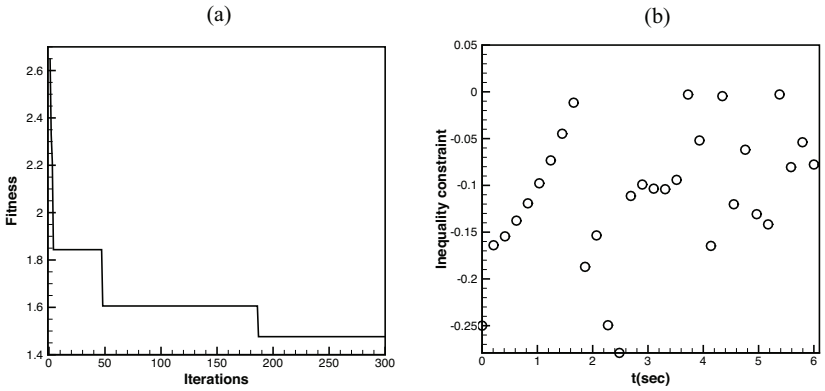


Figure 9: (a) The performance index versus iterations number, and (b) the inequality constraint of Example 9 ($N_t = 30$, $m = 50$).

The performance index convergence rate for 300 iterations is shown in Figure 9(a) and inequality constraint d for 300 iterations is shown in Figure 9(b). The numerical result for different N_t, m is presented in Table 11. According to Table 11, we find that increasing the number of partitions m and N_t, does not necessarily improve the solution. The performance index

Table 11: Comparison between the best numerical results of different partitions for Example 9

N_t	m	J	gap_J
10	10	1.822	0.346
30	50	1.476	0.000
30	100	1.546	0.07
50	50	1.795	0.319
50	100	1.596	0.12
75	75	1.993	0.517
100	100	2.118	0.642

in Table 12, J, for different N_t, m in our method is also less than $J = 1.7950$ by SQP and $J = 1.7980$ which was obtained by SUMT and $J = 1.6509$ by LI and SI. In general, its performance is influenced by the size of the problem.

Table 12: Comparison of the best value of the PHVND with some algorithms for nine examples

Problem	Algorithm	J	φ_f	gap_J
	CGA	1.7404	2.67×10^{-11}	0.2
	LI	1.5949	1.08×10^{-13}	0.054
Example 1	SI	1.5950	9.99×10^{-14}	0.054
	PHVND	1.5404	1.75×10^{-4}	0.000
	CGA	0.0163	7.57×10^{-10}	0.003
	LI	0.0127	1.15×10^{-9}	0.000
Example 2	SI	0.0127	5.99×10^{-9}	0.000
	PHVND	0.033	3.60×10^{-3}	0.021
	HPM	0.2353	4.2×10^{-6}	0.052
	LI	0.2015	2.35×10^{-9}	0.019
Example 3	SI	0.2015	2.82×10^{-10}	0.019
	GA	0.3526	4.09×10^{-5}	0.17
	PHVND	0.1825	6.81×10^{-7}	0.000
	SQP	-8.8690	0	0.000
	SUMT	-8.8690	0	0.000
Example 4	LI	-8.8692	1.45×10^{-10}	0.000
	SI	-8.8692	2.79×10^{-10}	0.000
	PHVND	-8.8157	3.6×10^{-2}	0.053
	Method of [18]	0.6107	-	0.106
Example 5	method of [40]	0.6128	-	0.104
	PHVND	0.717	-	0.000
	IPSO-SQP	0.1727	-	0.002
	LI	0.1699	-	0.000
Example 6	SI	0.1700	-	0.000
	PHVND	0.2746	-	0.104
	LI	-5.4309	-	3.56
Example 7	SI	-5.4309	-	3.56
	Bézier	-5.3898	-	3.6
	PHVND	-8.999	-	0.000
	SQP	0.2163	-	0.061
	LI	0.1549	-	0.000
Example 8	SI	0.1549	-	0.000
	SUMT	0.1703	-	0.015
	PHVND	0.2646	-	0.109
	SQP	1.7950	-	0.319
	LI	1.6509	-	0.174
Example 9	SI	1.6509	-	0.174
	SUMT	1.7980	-	0.322
	PHVND	1.476	-	0.000

4.2.1 Analysis of the numerical examples

By analyzing the number of times and controls that an approach used to reach its best solution, we can conclude that PHVND is more reliable than

other heuristics for problems. The results obtained by the algorithm are similar for small instances of the problem, but when the size increases, the PHVND approach improves the results in terms of quality of solutions and running times. Moreover, we improve the best-known solution for benchmark data sets presented in examples.

According to our computational results, PHVND is an efficient method in general since only a little computational time is added to the algorithm in each iteration. Besides, it does not require large RAM for relatively medium instances. Note that the PHVND can find optimal solutions (or feasible solutions in some cases), however, others cannot load the problem due to the shortage of memory. Whereas, the gap solution of others are larger than those of our algorithm for most cases (see Table 12).

Thus, the PHVND together with efficient partitioning for solving trajectory corresponding significantly improves our ability to tackle large instances of NOCP. On the other hand, by comparison of computational results of the number of partitions in Tables 2–12, our choice of the objective function has a significant impact on the number of iterations. In general, the proposed method can solve complex problems with accurate solutions as shown in Table 12. Note that the method outperforms control problems in average computational effort or the best solution gap for large instances. Tables 2–12 present results for network designs of NOCP for different partitions. Naturally, the selection of the number of nodes in the partitions resulting from the division of time and control space may depend on the dimensions of the problem, e.g. the number of state variables, but in any case, the numerical results show that the selection of finer partitions reduces the possibility of the solution getting worse. Basically, by dividing the time interval and control into more nodes, we can determine that in most cases, a worse criterion does not occur.

5 Conclusion

This paper proposes a PHVND algorithm to solve large-scale NOCPs. Comparing the results of the proposed method with existing methods for NOCP reveals that the proposed method leads to solutions closer to the optimal

solutions. The algorithm also increases the efficiency of the solution process, improves the scalability of performance by using the SVND process, and increases the diversification of solutions at the same time while reducing the execution time compared to the genetic algorithm. A set of large-scale NOCPs was used to test this algorithm. Examining the parallel implementation of the method shows that this method uses parallelism, which reduces the execution time. The proposed algorithm can significantly reduce execution time compared to other well-known algorithms in lower-scale problems. In most cases, our algorithm has achieved good results. In other cases, we have found a very close to the optimal solution, which, as seen in the convergence figures, will improve with the increasing number of repetitions. For future research, methods can be used to solve the minimum-optimal-time control problem, where the end of the time interval is not fixed. OCPs in dynamically distributed systems may also be genetically analyzed. From a computational point of view, this method can be used in computers with GPUs to take advantage of the acceleration of graphics processing and compare parallel performance on shared memory structures.

Abbreviations

ACO	Ant Colony Optimization
CGA	Continuous Genetic Algorithm
CRP	Chemical Reactor Problem
GA	Genetic Algorithm
HPM	Homotopy Perturbation Method
IM	Induction Motor
PSO	Particle Swarm Optimization
IPSO	Improved Particle Swarm Optimization
LI	Linear Interpolation
LQR	Linear Quadratic Regulator
NLP	Nonlinear Programming
NOCP	Nonlinear Optimal Control Problem
OPC	Optimal Control Problem

PHVND	Parallel Hybrid Variable Neighborhood Descent
PID	Proportional Integral Derivative
SI	Spline Interpolation
SQP	Sequential Quadratic Programming
SUMT	Sequential Unconstrained Minimization Technique
VND	Variable Neighborhood Descent
SVDN	Sequential Variable Neighborhood Descent
USApHMP	Uncapacitated Single Allocation p-hub Median problem
VDP	Van Der Pol
WOA	Whale Optimization Algorithm
C_j	jth feasible chromosome
f	A function representing the system dynamic
f_{opt}	Current optimal objective

Nomenclature

gap_J	Absolute error of the resulting objective value
J	Objective function
$\tilde{J}$	Discrete form of J
J^*	Optimal objective value
$l_{\max}$	Maximum number of iterations in Algorithm 1
m	Number of divisions of the control interval
$Maxiter$	Maximum number of iterations in Algorithm 2
M_1, M_2, M_{3k}	Penalty constants
N_p	Number of processors
N_{pop}	Number of population in GA
N_t	Number of time interval divisions
p_m	Mutation factor
p_c	Crossover constant
$pop(i)$	A randomly chosen individual
t	Time variable
t_j	jth time node

t_0	Initial time
t_f	Final time
u_l	Lower bound on the control values
u_b	Upper bound on the control values
$u(t), u$	Control variable
$u^*(t)$	Optimal control
$x(t), x$	State variable
$x^*(t)$	Optimal state
x_{opt}	Current optimal state
x_0	Initial state
x_f	Final state
φ_f	Total error in constraints
$\chi(t)$	Indicator function

References

[1] Abo-Hammour, Z.S., Asasfeh, A.G., Al-Smadi A.M. and Alsmadi, O.M., *A novel continuous genetic algorithm for the solution of optimal control problems*, Optim. Control Appl. Methods, 32(4) (2011), 414–432.

[2] Alzahrani, E.O. and Khan, M.A., *Modeling the dynamics of Hepatitis E with optimal control*, Chaos Solit. Fractals, 116 (2018), 287–301.

[3] Babaie-Kafaki, S., Ghanbari, R. and Mahdavi-Amiri, N., *Two effective hybrid metaheuristic algorithms for minimization of multimodal functions*, Int. J. Comput. Math. 88(11) (2011), 2415–2428.

[4] Betts, J.T., *Practical methods for optimal control and estimation using nonlinear programming*, Society for Industrial and Applied Mathematics, 2010.

[5] Borzabadi, A.H. and Mehne, H.H., *Ant colony optimization for optimal control problems*, Journal of Information and Computing Science, 4(4) (2009), 259–264.

[6] Chachuat, B., *Nonlinear and dynamic optimization: From theory to practice*, Automatic Control Laboratory, 2007.

[7] Costa, W.E., Goldbarg, M.C. and Goldbarg, E.G., *New VNS heuristic for total flowtime flowshop scheduling problem*, Expert Syst. Appl. 39(2) (2012), 8149–8161.

[8] Dadebo, S.A. and Mcauley, K.B., *Dynamic optimization of constrained chemical engineering problems using dynamic programming*, Comput. Chem. Eng. 19(5) (1995), 513–525.

[9] Effati, S. and Nik, H.S., *Solving a class of linear and non-linear optimal control problems by homotopy perturbation method*, IMA J. Math. Control Inf. 28(4) (2011), 539–553.

[10] Fabien, B.C., *Some tools for the direct solution of optimal control problems*, Adv. Eng. Softw. 29(1) (1998), 45–61.

[11] Fard, O.S. and Borzabadi, A.H., *Optimal Control Problem, Quasi-Assignment Problem and Genetic Algorithm*, World Academy of Science, Engineering and Technology, 33 (2007), 46–48.

[12] Gaing, Z.L., *A particle swarm optimization approach for optimum design of PID controller in AVR system*, IEEE Trans. Energy Convers. 19(2) (2004), 384–391.

[13] Ghomanjani, F., Farahi, M.H. and Gachpazan, M., *Bézier control points method to solve constrained quadratic optimal control of time varying linear systems*, Comput. Appl. Math. 31(3) (2012), 433–456.

[14] Hansen, P. and Mladenović, N., *Developments of Variable Neighborhood Search*, Essays and Surveys in Metaheuristics, Springer, 2002.

[15] Hansen, P., Mladenović, N., Moreno, P. and érez, J.A., *Variable neighbourhood search: methods and applications*, 4OR, 6(4) (2008), 319–360.

[16] Hansen, P., Mladenović, N. and Urošević, D., *Variable neighborhood search for the maximum clique*, Discret. Appl. Math. 145 (2004), 117–125.

[17] Ilić, A., Uroševic, D., Brimberg, J. and Mladenović, N., *A general variable neighborhood search for solving the uncapacitated single allocation p-hub median problem*, Eur. J. Oper. Res. 206(2) (2007), 289–300.

[18] Lazutkin, E., Geletu, A., Hopfgarten, S. and Li, P., *An Analytical Hessian and Parallel-Computing Approach for Efficient Dynamic Optimization Based on Control-Variable Correlation Analysis*, Ind. Eng. Chem. Res. 54(48) (2015), 12086–12095.

[19] Mehne, H.H., *Evaluation of parallelism in ant colony optimization method for numerical solution of optimal control problems*, J. Electr. Eng. Electron. Control Comput. Sci. Science, 1(2) (2015), 15–20.

[20] Mehne, H.H. and Mirjalili, S., *A parallel numerical method for solving optimal control problems based on whale optimization algorithm*, Knowl.-Based Syst. 151 (2018), 114–123.

[21] Michalewicz, Z., Krawczyk, J.B. and Kazemi, M., *Genetic algorithms and optimal control problems*, 29th IEEE Conference on Decision and Control, 3 (1990), 1664–1666.

[22] Modares, H. and Naghibi Sistani, M.B., *Solving nonlinear optimal control problems using a hybrid IPSO–SQP algorithm*, Eng. Appl. Artif. Intell. 24(3) (2011), 476–484.

[23] Nezhadhosein, S., Heydari, A. and Ghanbari, R., *A Modified Hybrid Genetic Algorithm for Solving Nonlinear Optimal Control Problems*, Math. Probl. Eng. 2015 (2015), 1–21.

[24] Nezhadhosein, S., Heydari, A. and Ghanbari, R., *Integrating Differential Evolution Algorithm with Modified Hybrid GA for Solving Nonlinear Optimal Control Problems*, Iran. J. Math. Sci. Inform. 12(1) (2017), 47–67.

[25] Nezhadhosein, S., Ghanbari, R. and Ghorbani-Moghadam, K., *Article Solving a Class of Nonlinear Optimal Control Problems Using Haar Wavelets and Hybrid GA*, Control and Optim. Appl. Math. 8 (2023), 1–17.

[26] Oliveira, F.A., de Sá, E.M., d. Souza, S. R. and Souza, M.J.F., *ILS-based algorithms for the profit maximizing uncapacitated hub network design problem with multiple allocation*, Comput. Oper. Res. 157 (2023), 0305–0548.

[27] Ozkan, Y., Aydin, Y.O., Saranli, A., Yazicioglu, Y., Saranli, U. and Leblebicioğlu, K., *Optimal control of a half-circular compliant legged monopod*, Control Eng. 33 (2014), 10–21.

[28] Rigatos, G., Abbaszadeh, M., Sari, B., Siano, P., Cuccurullo, G. and Zouari, F., *Nonlinear optimal control for a gas compressor driven by an induction motor*, Results Control. Optim. 11(2023) 100226.

[29] Rigatos, G., Siano, P., AL-Numay, M., Abbaszadeh, M. and Sari, B., *Nonlinear optimal and multi-loop flatness-based control of induction motor-driven desalination units*, Results Control. Optim. 14 (2024), 100360.

[30] Roy, T. and Chakraborty, D., *Optimal vibration control of smart fiber reinforced composite shell structures using improved genetic algorithm*, J. Sound Vib. 319 (2009), 15-40.

[31] Said, S.M. and Nakamura, M., *Asynchronous parallel algorithms for strategic hybrid on a mixture gaussin model, International Journal of Innovative Computing*, Inf. Control, 10(2) (2014), 459-479.

[32] Salimi, M., Borzabadi, A.H., Mehne, H.H. and Heydari, A., *The hub location's method for solving optimal control problems*, Evol. Intell. 14 (2021), 1671-1690.

[33] Shi, X.H., Wan, L.M., Lee, P.H., Yang, X.W., Wang, L.M. and Liang, Y.C., *An improved genetic algorithm with variable population-size and a PSO-GA based hybrid evolutionary algorithm*, Int. J. Syst. Sci. 3 (2003), 1735-1740.

[34] Sun, F., Du, W., Qi, R., Qian, F. and Zhong, W., *A Hybrid Improved Genetic Algorithm and Its Application in Dynamic Optimization Problems of Chemical Processes*, Chin. J. Chem. Eng. 21(2) (2013), 144-154.

[35] Tavakolpour, A.R., Mat Darus, I.Z., Tokhi, O. and Mailah, M., *Genetic algorithm-based identification of transfer function parameters for a rectangular flexible plate system*, Eng. Appl. Artif. Intell. 23(8) (2010), 1388-1397.

[36] Van Soest, A.J.K. and Casius, L.J.R.R., it The Merits of a Parallel Genetic Algorithm in Solving Hard Optimization Problems, J. Biomech. Eng.125(1) (2003), 141–146.

[37] Wang, Z. and Ju, G., *A parallel genetic algorithm in multi-objective optimization*, 2009 Chinese Control and Decision Conference, 3497–3501.

[38] Roberge, V., Tarbouchi, M. and Labonte, G., *Comparison of Parallel Genetic Algorithm and Particle Swarm Optimization for Real-Time UAV Path Planning*, IEEE Trans. Ind. Inform. 9(1) (2013), 132–141.

[39] Wolf, S. and Merz, P., *Evolutionary local search for the super-peer selection problem and the p-Hub median problem*, Hybrid Metaheuristics, (2007), 1–15.

[40] Wu, X., Lei, B., Zhang, K. and Cheng, M., *Hybrid stochastic optimization method for optimal control problems of chemical processes*, Chem. Eng. Res. Des. 126 (2017), 297–310.

[41] Zhang, B., Chen, D. and Zhao, W., *Iterative ant-colony algorithm and its application to dynamic optimization of chemical process*, Comput. Chem. Eng. 29(10) (2005), 2078–2086.

Iranian Journal of Numerical Analysis and Optimization
Vol. 15, No. 2, 2025, pp 457–474
https://doi.org/10.22067/ijnao.2024.89654.1503
https://ijnao.um.ac.ir/

Research Article

Solving mixed singular integro-differential equation (IDE) using the finite elements method

M. Erfanian*, H. Zeidabadi and M.S. Hussein

Abstract

In this work, we describe a new method for the calculation of the numerical solution of the mixed singular integro-differential equation. The method is mainly based on a finite element approximation. For this purpose, we obtain the variational form of the problem under consideration. The resulting system has been approximated numerically by linear B-spline function. The method of convergence and its order of convergence are established.

*Corresponding author

Received 31 August 2024; revised 12 December 2024; accepted 25 December 2024

Majid Erfanian
Department of Science, School of Mathematical Sciences,University of Zabol, Zabol, Iran. e-mail: erfaniyan@uoz.ac.ir

Hamed Zeidabadi
Department of Applied Mathematics, Faculty of Mathematics and Computer Sciences, Hakim Sabzevari University, Sabzevar, Iran. e-mail: h.zeidabadi@yahoo.com

M.S. Hussein
Department of Mathematics College of Science University of Baghdad. e-mail: mmmsh@sc.uobaghdad.edu.iq

How to cite this article
Erfanian, M., Zeidabadi, H. and Hussein, M.S., Solving mixed singular integro-differential equation (IDE) using the finite elements method. *Iran. J. Numer. Anal. Optim.*, 2025; 15(2): 457-474. https://doi.org/10.22067/ijnao.2024.89654.1503

Finally, to investigate the accuracy of the proposed method some test examples are presented. The numerically obtained results are compared with another method for the validity of our results and shown by some tables and figures.

AMS subject classifications (2020): 65L60,44A45, 45B05, 65R20.

Keywords: Singular integro-differential equation; finite element approximation; Convergence; Error analysis.

1 Introduction

The integro-differential equation is a sophisticated mathematical construct that blends the principles of integral and differential calculus, serving as a vital tool in various fields such as physics, engineering, and applied mathematics. Need to write an introduction here singular integro-differential equations (SIDEs) appear in a lot of science, for example, in applied sciences, electromagnetics, physics, many problems in acoustics, finance, biology, engineering, viscoelasticity, hydrology, analysis of dynamic systems, where the evolution of states is influenced by both local changes and historical interactions, and so on [5, 11, 6, 12, 9].

One of the key challenges in solving integro-differential equations lies in their inherent complexity, as traditional methods for solving ordinary differential equations may not be directly applicable. In the last decades, several techniques have been presented to successfully used to approximate different kinds of singular integral equations, such as Jacobi and B-spline collocation method, Crank–Nicolson finite difference method, Daubechies wavelet, Hermite wavelet method, wavelet methods, trigonometric Hermit wavelet, discrete collocation method, Legendre wavelet and Legendre multi-wavelet methods and Petrov-Galerkin method with a wide range of applications, Block boundary value methods, operational matrix-based schemes with relaxation processes [4, 10, 13, 7, 2, 8]. The significance of integro-differential equations is further underscored by their ability to model phenomena that cannot be adequately described by either differentials or integrals alone, thus

offering a more comprehensive framework for understanding complex systems wherein time delays or memory effects play a critical role.

2 Mathematical formulation

The integral equations frequently appear in many fields of mathematics such as game theory, control theory, and numerical analysis. Also, they appear in real physical world problems in other fields of science like as fluid dynamics, queuing theory, filtration theory, electromagnetic, chemical kinetics, laminar boundary-layer theory, solid state physics, plasma physics, the diffusion of neutrons in nuclear reactor dynamics, and so on [5, 11, 6, 12, 9, 1].

In this work, we use the finite element method (FEM) to solve mixed singular integro-differential equation (MSIDE), as follows:

$$\begin{aligned} -u''(x) + f_1(x)u'(x) + f_2(x)u(x) =& f_3(x) + \int_a^x \frac{K_1(x,t)u(t)}{(x-t)^\alpha}dt \\ &+ \int_a^b \frac{K_2(x,t)u(t)}{(x-t)^\beta}dt, \end{aligned} \tag{1}$$

with

$$u(a) = 0, \quad u(b) = 0, \quad \Omega = [a,b] \subset \mathbb{R}, \tag{2}$$

where $K_1(x,t)$, $K_2(x,t)$, and $f_i(x)$, $\quad i = 1,2,3$ are known continuous functions belong to $C^1(\Omega)$, $u(x)$ is the unknown function to be determined, and $\alpha = \beta = \frac{1}{2}$.

The major goal of this research is to apply Lagrange polynomials combined with FEM for solving MSIDE of (1)–(2) and obtain an approximate solution. First, we obtain the weak and variational form of (1). In section 3, we analyze the application and prove the bilinear form is a $V-$ellipticity and continuous. Also, we proved that (1) has a unique solution. By using some theorem and lemma we calculated the order of convergence is $O(h^2)$ in section 4. We present some numerical test examples in section 5. Finally, the conclusions are highlighted in section 6.

3 Finite element method

One of the best techniques for numerically resolving differential equations in structural analysis, electromagnetic potential, fluid flow, ferroelectric, and electromagnetics in engineering and boundary value problems, algebraic equations, complex mathematical problems, and heat transfer, in mathematical modeling, began in the mid-1950s, used to compute such approximations nominated as FEM [1].

First, we obtain the virtual work (or weak form) and variational form of (1). For this purpose, we assume a Sobolev space of $V = H_0^1(\Omega)$, with the following norm:

$$||u||_V^2 = ||u||_{L^2(\Omega)}^2 + ||u'||_{L^2(\Omega)}^2.$$

Also, the bilinear form is $B : V \times V \to \mathbb{R}$; linear functional has a $L : V \to \mathbb{R}$; and

$$B(u, v) = L(v) = \int_\Omega f_3(x)v(x)dx, \tag{3}$$

for all $v(x) \in V$ is an arbitrary function. In addition,

$$\begin{aligned} B(u, v) = &\int_\Omega f_2(x)u(x)v(x)dx + \int_\Omega u'(x)\left(f_1(x)\,v(x) + v'(x)\right)dx \\ &- \int_\Omega v(x)\left(\int_a^x \frac{K_1(x,t)\,u(t)}{(x-t)^\alpha}dt\right)dx \\ &- \int_\Omega v(x)\left(\int_a^b \frac{K_2(x,t)\,u(t)}{(x-t)^\beta}dt\right)dx. \end{aligned} \tag{4}$$

Since we have

$$\begin{aligned} B(\eta_1 u + \eta_2 w, v) = &\int_\Omega \big(\eta_1 u'(x) + \eta_2 w'(x)\big)v'(x)dx \\ &+ \int_\Omega f_1(x)\left(\eta_1 u'(x) + \eta_2 w'(x)\right)v(x)dx \\ &+ \int_\Omega f_2(x)\left(\eta_1 u(x) + \eta_2 w(x)\right)v(x)dx \\ &- \int_\Omega v(x)\int_a^x \frac{K_1(x,t)\left(\eta_1 u(x) + \eta_2 w(x)\right)}{(x-t)^\alpha}dtdx \\ &- \int_\Omega v(x)\int_a^b \frac{K_2(x,t)\left(\eta_1 u(x) + \eta_2 w(x)\right)}{(x-t)^\alpha}dtdx \end{aligned}$$

$$=\eta_1 B(u,v)+\eta_2 B(w,v),$$

then, B is a bilinear form. In this work, we use the linear B-spline (LBS) function as the basis functions of subspace $V_h = \text{span}\{\phi_1, \phi_2, \ldots, \phi_n\}$. If $a = x_0 < x_1 < \cdots < x_n = b$ is a grid with $n+1$ points in the interval $[a,b]$, where $x_i = ih$ for $i = 1, \ldots, n$, and $\Delta x = h = \frac{1}{n}$, then, for $i = 1, \ldots, n$, the LBS is defined as

$$\phi_i(x) = \frac{1}{h}\begin{cases} (x - x_{i-1}), & x \in [x_{i-1}, x_i], \\ (x_{i+1} - x), & x \in [x_i, x_{i+1}], \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Thus $\phi_i(x_i) = 1, \phi_i(x_j) = \delta_{ij}$, for $i, j = 1, 2, \ldots, n$, and the value of ϕ in the other nodes is equal to zero. Therefore, we approximate $u_h(x)$ and $v_h(x)$ as

$$u_h(x) = \sum_{i=1}^{n} a_i \phi_i(x), \qquad v_h(x) = \phi_j(x). \tag{6}$$

By the combination of (6) and variational formulation of the problem, we have

$$\begin{aligned} \int_\Omega f(x)\phi_j(x)dx = \sum_{i=1}^{n} a_i \Bigg\{ & \int_\Omega f_2(x)\phi_i(x)\phi_j(x)dx \\ & + \int_\Omega \phi_i'(x)\left(\phi_j'(x) + f_1(x)\phi_j(x)\right)dx \\ & - \int_\Omega \phi_j(x)\left(\int_a^x \frac{K_1(x,t)}{(x-t)^\alpha}\,\phi_i(t)\,dt\right)dx \\ & - \int_\Omega \phi_j(x)\left(\int_a^b \frac{K_2(x,t)}{(x-t)^\beta}\,\phi_i(t)\,dt\right)dx \Bigg\}. \end{aligned} \tag{7}$$

Now, we define

$$\begin{aligned} C_{i,j} = & \int_\Omega \phi_i'(x)\left(\phi_j'(x) + f_1(x)\phi_j(x)\right)dx + \int_\Omega f_2(x)\phi_i(x)\phi_j(x)dx \\ & - \int_\Omega \phi_j(x)\left(\int_a^x \frac{K_1(x,t)}{(x-t)^\alpha}\,\phi_i(t)\,dt\right)dx \\ & - \int_\Omega \phi_j(x)\left(\int_a^b \frac{K_2(x,t)}{(x-t)^\beta}\,\phi_i(t)\,dt\right)dx, \end{aligned} \tag{8}$$

and

$$F_j = \int_\Omega f(x)\phi_j(x)dx. \tag{9}$$

Thus the above system can be expressed as a system of an algebraic equation of the form

$$\sum_{i=1}^{n} C_{ij}a_i = F_j, \quad j = 1, 2, \ldots, n. \tag{10}$$

4 Error analysis

In this section, we aim to establish a lower and upper bound for approximation error. To prove that we need to define the following definition.

Definition 1. The inner product energy can be described as

$$(\cdot,\cdot)_B : V \times V \to \mathbb{R},$$

such that for the Hilbert space V and the $V-$elliptic bilinear form B, we have $(u,v)_B = B(u,v)$. Also, we can define the energy norm as

$$||u||_E^2 = (u,u)_B.$$

Lemma 1. If $f_1(x) \in [M_1, M_2]$, $f_2(x) \in [P_1, P_2]$, and B is a bilinear form defined by (4), then B is continuous.

Proof. First, by (4), we have

$$\begin{aligned} |B(u,v)| = \Bigg| & \int_\Omega f_2(x)u(x)v(x)dx + \int_\Omega u'(x)\left(v'(x) + f_1(x)v(x)\right)dx \\ & - \int_\Omega v(x)\left(\int_a^x \frac{K_1(x,t)\,u(t)}{(x-t)^\alpha}dt\right)dx \\ & - \int_\Omega v(x)\left(\int_a^b \frac{K_2(x,t)\,u(t)}{(x-t)^\beta}dt\right)dx\Bigg|. \end{aligned} \tag{11}$$

In (11) by applying the Sobolev norm and the Cauchy–Schwarz inequality, we have

$$\begin{aligned} |B(u,v)| \leq & \, ||u||_{H^1}||v||_{H^1} + P_2\,||u||_{H^1}||v||_{H^1} + M_2\,||u||_{H^1}||v||_{H^1} \\ & + K_1\left|\int_a^b v(x)\int_a^x \frac{u(t)}{(x-t)^\alpha}dtdx\right| \end{aligned}$$

$$
\begin{aligned}
&+ K_2 \left| \int_a^b v(x) \int_a^b \frac{u(t)}{(x-t)^{\beta}} dt dx \right| \\
&= \left(1 + P_2 + M_2\right) ||u||_{H^1} ||v||_{H^1} \\
&+ K_1 \left| \int_a^b v(x) u(\eta_x) \int_a^x \frac{1}{(x-t)^{\alpha}} dt dx \right| \\
&+ K_2 \left| \int_a^b v(x) u(\zeta_x) \int_a^b \frac{1}{(x-t)^{\beta}} dt dx \right| \\
&\leq \left(1 + P_2 + M_2\right) ||u||_{H^1} ||v||_{H^1} \\
&+ K_1 \left| \int_a^b v(x) u(\eta_x) \frac{1}{1-\alpha} (x-t)^{1-\alpha} \, |_{t=a}^{t=x} \, dx \right| \\
&+ K_2 \left| \int_a^b v(x) u(\zeta_x) \frac{1}{1-\beta} (x-t)^{1-\beta} \, |_{t=a}^{t=b} \, dx \right| \\
&\leq \left(1 + P_2 + M_2\right) ||u||_{H^1} ||v||_{H^1} \\
&+ \frac{K_1 (b-a)^{1-\alpha}}{1-\alpha} \left| \int_a^b v(x) u(\eta_x) dx \right| \\
&+ \frac{K_2 (b-a)^{1-\beta}}{1-\beta} \left| \int_a^b v(x) u(\eta_x) dx \right| \\
&\leq \left(1 + P_2 + M_2 + \frac{K_1 (b-a)^{1-\alpha}}{1-\alpha} + \frac{K_2 (b-a)^{1-\beta}}{1-\beta}\right) ||u||_{H^1} \, ||v||_{H^1},
\end{aligned}
$$

where

$$K_1 = \max |K_1(x,t)|, \quad a \leq x \leq b, \quad a \leq t \leq x$$

and

$$K_2 = \max |K_2(x,t)|, \quad a \leq x \leq b, \quad a \leq t \leq x.$$

Thus B is continuous. □

In addition to the hypothesis of Lemma 1, suppose $T_2 = \max_{a \leq x \leq b} |f_1^{'}(x)|$. Now, we consider the $V-$ ellipticity of B. By using (4) and Lemma 1, we write

$$
\begin{aligned}
\int_{\Omega} f_1(x) v'(x) v(x) dx &= \frac{-1}{2} \int_a^b f_1'(x) (v(x))^2 dx \\
&\geq \frac{-T_2}{2} \int_a^b (v(x))^2 dx
\end{aligned}
$$

$$\geq \frac{-T_2}{2}||v||^2_{H^1}, \tag{12}$$

and

$$\int_\Omega v'(x)v'(x)dx + \int_\Omega f_2(x)v(x)v(x)dx \geq \int_\Omega v'^2(x)dx \geq \frac{1}{1+c}||v||^2_{H^1}. \tag{13}$$

Thus

$$\begin{aligned} B(v,v) &= \int_\Omega v'(x)v'(x)dx + \int_\Omega f_1(x)v'(x)v(x)dx + \int_\Omega f_2(x)v(x)v(x)dx \\ &\quad - \int_\Omega v(x)\int_a^x \frac{K_1(x,t)v(t)}{(x-t)^\alpha}dtdx - \int_\Omega v(x)\int_a^b \frac{K_2(x,t)v(t)}{(x-t)^\beta}dtdx \\ &\geq \left(\frac{1}{1+c} - \frac{T_2}{2}\right)||v||^2_{H^1} - K_1\left(\frac{(b-a)^{1-\alpha}}{1-\alpha}\right)||v||^2_{H^1} \\ &\quad - K_2\left(\frac{(b-a)^{1-\beta}}{1-\beta}\right)||v||^2_{H^1} \\ &= \eta\ ||v||^2_{H^1(\Omega)}, \end{aligned} \tag{14}$$

so

$$B(v,v) \geq \eta\ ||v||^2_{L_2(\Omega)}, \tag{15}$$

where

$$\eta = \frac{1}{1+c} - K_2\left(\frac{(b-a)^{1-\beta}}{1-\beta}\right) - \frac{T_2}{2} - K_1\left(\frac{(b-a)^{1-\alpha}}{1-\alpha}\right),$$

where c is Poincaré's constant. Then, the following lemma can be expressed.

Lemma 2. If $\eta > 0$, then B is $V-$ellipticity.

By applying the Lax–Milgram theorem and Lemmas 1 and 2, the equation (1) has a unique solution.

Since for each particular $\tilde{v}_h$ in V_h, $\inf||u-v_h||_V \leq ||u-\tilde{v}_h||_V$, thus for seeking of an upper bound $u-u_h$, we can take $\tilde{v}_h$ equal to $\tilde{u}_h$. Then

$$||u-u_h||_V \leq c||u-\tilde{u}_h||_V.$$

By using Cea's lemma [3] and

$$||u-u_h||_E = \min_{v_h \in V_h} ||u-v_h||_E$$

for each particular $\tilde{v}_h$ in V_h, we have

$$\alpha||u-u_h||_V^2 \leq B(u-u_h, u-u_h) = ||u-u_h||_E^2. \tag{16}$$

Thus $||u-u_h||_V \leq \frac{1}{\sqrt{\alpha}}||u-u_h||_E$. Also,

$$||u-v_h||_E^2 = B(u-v_h, u-v_h) \leq C||u-v_h||_V||u-v_h||_V. \tag{17}$$

Therefore $||u-v_h||_E \leq \sqrt{C}||u-v_h||_V$. By (16) and (17),

$$||u-u_h||_V \leq \sqrt{\frac{C}{\alpha}} \min_{v_h \in V_h} ||u-v_h||_V. \tag{18}$$

If we define interpolation error $E(x) = u(x) - \tilde{u}_h(x)$, such that error on Ω_e, then

$$E(x_1^{(e)}) = E(x_2^{(e)}) = E(x_3^{(e)}) = 0.$$

Also, by using Rolle's theorem, we have

$$E'(\xi_1) = 0 \qquad \xi_1 \in \left(x_1^{(e)}, x_2^{(e)}\right),$$

$$E'(\xi_2) = 0 \qquad \xi_2 \in \left(x_2^{(e)}, x_3^{(e)}\right),$$

$$E''(\eta) = 0 \qquad \eta \in \left(x_1^{(e)}, x_3^{(e)}\right).$$

We know the polynomial interpolation that we used in our equation, is a piecewise quadratic polynomial. Thus

$$|E''(x)| = u''(x) - u''_h(x) = \int_\eta^x E'''(t)dt = \left|\int_\eta^x u'''(t)dt\right|$$
$$\leq \int_\eta^x |u'''(t)|dt \leq \int_{x_1^{(e)}}^{x_3^{(e)}} |u'''(t)|dt \leq ||1||_{L^2}||u'''||_{L^2},$$

or

$$|E''(x)| \leq h^{\frac{1}{2}}|u|_{H^3(\Omega_e)}.$$

Then

$$\int_{\Omega_e} |E''(x)|^2 dx \leq h|u|^2_{H^3(\Omega_e)} \int_{\Omega_e} dx.$$

Thus

$$||E''||^2_{L^2(\Omega_e)} = |E|^2_{H^2(\Omega_e)} \leq h^2 |u|^2_{H^3(\Omega_e)}.$$

Therefore,

$$\begin{aligned} ||E''||^2_{L^2(\Omega)} = |E|^2_{H^2(\Omega)} &= \int_\Omega |E''(x)|^2 dx = \sum_{e=1}^{M} \int_{\Omega_e} |E''(x)|^2 dx = \sum_{e=1}^{M} |E|^2_{H^2(\Omega_e)} \\ &\leq h^2 \sum_{e=1}^{M} |u|^2_{H^3(\Omega_e)} = h^2 \sum_{e=1}^{M} \int_{\Omega_e} (u'''(x))^2 dx \\ &= h^2 \int_\Omega |u'''(x)|^2 dx = h^2 |u|^2_{H^3(\Omega)}, \end{aligned}$$

that is,

$$|E(x)|_{H^2(\Omega)} \leq h |u|_{H^3(\Omega)}. \tag{19}$$

Similarly, we have

$$|E|^2_{H^1(\Omega)} \leq h^2 |E|^2_{H^2(\Omega)}, \tag{20}$$

that is,

$$||E||_{L^2(\Omega)} \leq h \, |E|_{H^1(\Omega)}. \tag{21}$$

By (19), (20) and (21), we have

$$||E||^2_{L^2(\Omega)} \leq h^2 h^2 h^2 |u|^2_{H^3(\Omega)} = h^6 |u|^2_{H^3(\Omega)}. \tag{22}$$

By using the Sobolev norm, we have

$$\begin{aligned} ||E||^2_{H^1(\Omega)} = ||E||^2_{L^2(\Omega)} + ||E'||^2_{L^2(\Omega)} &\leq h^6 |u|^2_{H^3(\Omega)} + |E|^2_{H^1(\Omega)} \\ &\leq h^6 |u|^2_{H^3(\Omega)} + h^4 |u|^2_{H^3(\Omega)} \leq 2h^4 |u|^2_{H^3(\Omega)}. \end{aligned} \tag{23}$$

Thus the upper bound for the interpolation error as

$$||E||_{H^1(\Omega)} \leq \sqrt{2} h^2 |u|_{H^3(\Omega)}.$$

The unique solution to the variational form is known. Therefore $|u|_{H^2(\Omega)}$ is a constant number. Also, according to (18), if α is the V-ellipticity constant

and C is the continuity constant, then

$$||u - u_h||_V \leq \frac{C}{\alpha}||u - \tilde{u}_h||_V.$$

Thus

$$||u - u_h||_{H^1(\Omega)} \leq \frac{C}{\alpha}\sqrt{2}h^2|u|_{H^3(\Omega)}.$$

Since $|u|_{H^3(\Omega)}$ is constant, and $h \to 0$, thus the method is convergence, and the rate of method is $O(h^2)$.

5 Numerical examples

In this section, we present two examples: one of MSIDE and another singular Volterra integro-differential equation (SVIDE). We compare numerical and accurate solutions of our method with FEM and radial basis function (RBF) methods in tables and figures. Also, we compute the error of examples shown in the Figures, and the accuracy of the method is shown for all examples. Also, we define the root mean square (RMS) or total error, as

$$RMS = \sqrt{\frac{1}{M}\sum_{i=1}^{M}\left(p(t_i)^{exact} - p(t_i)^{numerical}\right)^2},$$

where the total error is RMS and the total number of estimated values is $M = 5$.

In addition, to define the sequence of approximating functions $\{u_i\}_{i\in\mathbb{N}}$, the initial function $u_0 \in C([0,1])$ can be chosen arbitrarily. The following algorithm, based on the method presented in section 3, has been used to solve Examples 1 and 2.

Algorithm 4.1 Step 1. Make the boundary conditions homogeneous
Step 2. Input $n, [a, b]$
Step 3. Set $h = \frac{b-a}{n}$
Step 4. Obtain the virtual work (or weak form) and variational form of (1).
Step 5. For $i = 1$ to n, $x_i = a + ih$, and $\phi_i(x)$ is computed form (5).
Step 6. Compute $u_h(x), v_h(x)$ from (6) for the assumed point.
Step 7. Compute $C_{i,j}$ from (8).

Step 8. Product matrices F_j from (9).
Step 9. Solving (10), obtain an approximate solution of (1).

All results are computed by using Maple and a Lenovo Laptop (Legion Y545) to run the programs. Finally, the RMS of Examples 1 and 2 is given in the last line of the tables.

Example 1. Consider the MSIDE and $0 \leq x \leq 1$ as follows:

$$\begin{aligned} &- u''(x) + e^x u'(x) + x\, u(x) - \int_0^x \frac{1}{\sqrt{x-t}} u(t)dt \\ &- \int_0^1 \frac{1}{\sqrt{x-t}} u(t)dt - f(x) = 0, \end{aligned}$$

where the exact solution $u(x) = 3x^4 - x$, and

$$\begin{aligned} f(x) = &- 37\, x^2 + 12\, \mathrm{e}^x x^3 - \mathrm{e}^x + 3\, x^5 - \frac{512\, x^{\frac{9}{2}}}{105} + \frac{8}{3}\, x^{\frac{3}{2}} - \frac{4}{7}\sqrt{x-1}\,(x) \\ &+ \frac{32\sqrt{x-1}x^2}{35} + \frac{128\sqrt{x-1}x^3}{105} + \frac{256\sqrt{x-1}x^4}{105}. \end{aligned}$$

The inhomogeneous conditions must be transformed into homogeneous boundary conditions via transformation formulas. Comparison of numerical and accurate solutions by our method with FEM and RBF methods are shown in Table 1 and Figure 1. Also, we compute the error of Example 1 in Figure 2.

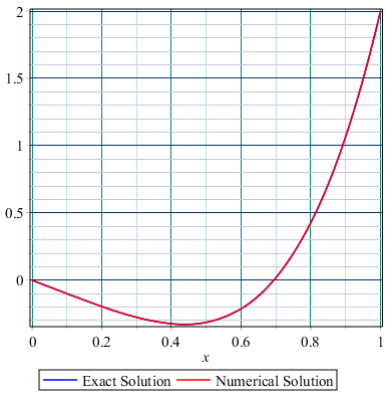


Figure 1: Graph of exact and numerical solutions for Example 1.

Table 1: The result obtain for Example 1.

x	$Exact$	$FEMMethod$	$RBFMethod$	$Cheb-FEMMethod$
0.1	-0.0997000	-0.0998079	-0.0982779	-0.1003080
0.2	-0.1952000	-0.1952376	-0.1936725	$-.19516122$
0.3	-0.2757000	-0.2759063	-0.2740468	$-.27504775$
0.4	-0.3232000	-0.3232568	-0.3214751	-0.3224741
0.5	-0.3125000	-0.3128260	-0.310652	-0.31219453
0.6	-0.2112000	-0.2112657	-0.2093563	-0.2114496
0.7	0.02030000	0.0198278	0.0222087	0.01978290
0.8	0.42880000	0.4287426	0.4306302	0.42852739
0.9	1.06830000	1.0676660	1.0701542	1.0685356
RMS		2.82×10^{-4}	1.66×10^{-3}	4.56×10^{-4}

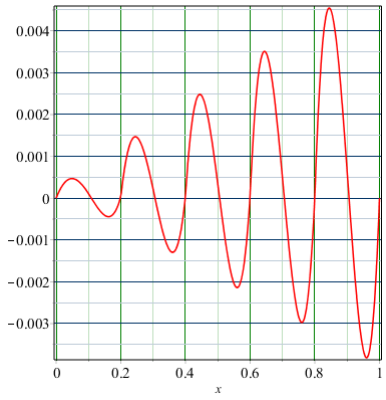


Figure 2: Diagrams of error for Example 1.

Example 2. We assume the SVIDE and $0 \le x \le 1$ as follows:

$$-u''(x)+\sqrt{x}u'(x)+e^x\,u(x)-\int_0^x \frac{1}{\sqrt{x-t}}u(t)dt$$
$$-\int_0^1 \frac{1}{\sqrt{x-t}}u(t)dt-f(x)=0,$$

where the exact solution is a $u(x)=x^3+2\,e^x$, and

$$f(x)=2\,\mathrm{erf}\left(\sqrt{x-1}\right)\sqrt{\pi}\mathrm{e}^x+\frac{1}{35}\left(32\,x^3+16\,x^2+12\,x+10\right)\sqrt{x-1}+3\,x^{5/2}$$

$$-\frac{64x^{7/2}}{35}-4\mathrm{erf}\left(\sqrt{x}\right)\sqrt{\pi}\mathrm{e}^{x}+2\mathrm{e}^{2x}+2\sqrt{x}\mathrm{e}^{x}+\frac{1}{35}\left(35x^{3}-70\right)\mathrm{e}^{x}-6x.$$

First, we convert the boundary conditions at $x = 0$ and $x = 1$ until the example is homogeneous. A comparison between exact and numerical solutions by our method with FEM and RBF methods is shown in Table 2 and Figure 3. Also, we compute the error of Example 2 in Figure 4.

Table 2: The result obtain for Example 2.

x	$Exact$	$FEMMethod$	$RBFMethod$	$Cheb-FEMMethod$
0.1	2.2113418	2.2113258	2.2111969	2.2115616
0.2	2.4508055	2.4507929	2.4507205	2.4507892
0.3	2.7267176	2.7266735	2.7266832	2.7264834
0.4	3.0476493	3.0476138	3.0476679	3.0473934
0.5	3.4224425	3.4223740	3.4225271	3.4223379
0.6	3.8602376	3.8601901	3.8603829	3.8603269
0.7	4.3705054	4.3704358	4.3707312	4.3706852
0.8	4.9630818	4.9630550	4.9633822	4.9631763
0.9	5.6482062	5.6481528	5.6486308	5.6481290
RMS		4.32×10^{-5}	2.06×10^{-4}	1.61×10^{-4}

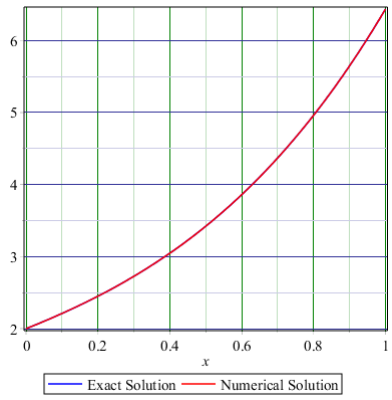


Figure 3: Graph of exact and FEM method for Example 2.

Example 3. We assume the SVIDE and $0 \le x \le 1$ as follows:

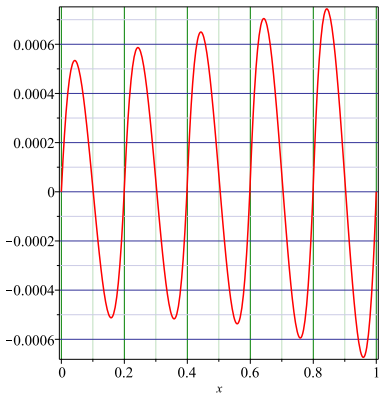


Figure 4: Diagrams of error for Example 2.

$$-u''(x)+x^2u'(x)+2x\,u(x)-\int_0^x\frac{2xt^2}{(x-t)^{\frac{1}{3}}}u(t)dt \\ -\int_0^1\frac{3x^2+t+2}{(x-t)^{\frac{2}{5}}}u(t)dt-f(x)=0,$$

where the exact solution is $u(x)=x^3-x+1$, and

$$\begin{aligned}f(x)=&-4\,x+5\,x^4-3\,x^2-\frac{2187}{2618}x^{\frac{20}{3}}+\frac{243\,x^{\frac{14}{3}}}{220}-\frac{27\,x^{\frac{11}{3}}}{20}-\frac{625}{312}x^{\frac{28}{5}}\\&-\frac{3125}{5382}x^{\frac{23}{5}}+\frac{1675}{936}x^{\frac{18}{5}}-\frac{655}{156}x^{\frac{13}{5}}+\frac{25\,x^{\frac{8}{5}}}{24}-\frac{10}{3}\,x^{\frac{3}{5}}\\&+\frac{10135\,(x-1)^{\frac{3}{5}}\,x^2}{2392}-\frac{10325\,(x-1)^{\frac{3}{5}}\,x^3}{21528}\\&+\frac{38375\,(x-1)^{\frac{3}{5}}\,x^4}{21528}+\frac{625\,(x-1)^{\frac{3}{5}}\,x^5}{312}+\frac{66665\,(x-1)^{\frac{3}{5}}}{21528}\\&-\frac{13775\,(x-1)^{\frac{3}{5}}\,(x)}{21528}.\end{aligned}$$

First, we convert the boundary conditions at $x=0$ and $x=1$ until the example is homogeneous. Comparison between exact and numerical solutions by our method with FEM and RBF method is shown in Table 3 and Figure 5. Also, we compute the error of Example 3 in Figure 6.

Table 3: The result obtain for Example 3.

x	$Exact$	$FEMMethod$	$RBFMethod$	$Cheb-FEMMethod$
0.1	0.9010000	0.9009946	0.8847265	0.9010000
0.2	0.8080000	0.8079910	0.7780601	0.8080000
0.3	0.7270000	0.7269861	0.6861300	0.7270000
0.4	0.6640000	0.6639860	0.6150528	0.6640000
0.5	0.6250000	0.6249798	0.5711002	0.6250000
0.6	0.6160000	0.6159850	0.5607239	0.6160000
0.7	0.6430000	0.6429763	0.5907674	0.6430000
0.8	0.7120000	0.7119892	0.6685135	0.7120000
0.9	0.8290000	0.8289770	0.8020527	0.8290000
RMS		1.52×10^{-5}	3.96×10^{-2}	5.29×10^{-11}

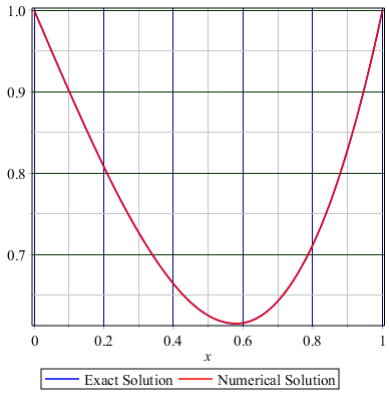


Figure 5: Graph of exact and FEM method for Example 3.

6 Conclusions

In this work, we used the FEM and Lagrange polynomials to solve one of the most important kinds of singular integral equations. The main purpose of this work is to apply Lagrange polynomials and FEM for solving MSIDE (1) and obtain an approximate solution. First, we obtain the weak and variational form of (1), and with the use of the system (10), we can obtain the approximate solution. Then, we proved the convergence of the method. Also, the order of the method is computed. Finally, For efficiency, we presented

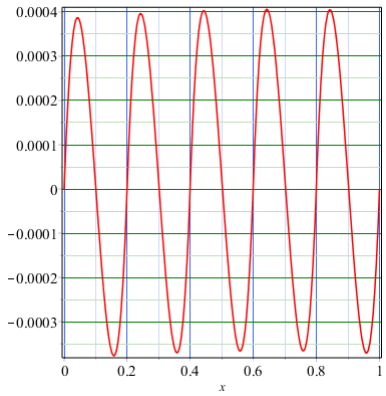


Figure 6: Diagrams of error for Example 3.

some numerical examples and used polynomials of degree 2. Also, the results of comparing between exact FEM method with other methods such as RBF are shown in tables and figures. We conclude the comparison of the results shows that this method has superior accuracy. Also, we computed the error of examples.

References

[1] Ames, W.F. *Nonlinear partial differential equations in engineering*, Academic Press, New York, 1965.

[2] Behera, S. and Saha Ray, S.*An operational matrix based scheme for numerical solutions of nonlinear weakly singular partial integro-differential equations*, Appl. Math. Comput. 367 (2020) 124771.

[3] Brenner, S. and Ridgway Scott. L. *The mathematical theory of finite element methods*, Texts in Applied Mathematics, Springer, 2007.

[4] El-Wakili, S. , Elhanbaly, A. and Abdou, M.A. *Adomian decomposition method for solving fractional nonlinear differential equations*. Appl. Math. Comput, 182 (2006) 313–324.

[5] Liu, H., Ma, Y., Li, H. and Zhang, W. *Combination of discrete technique on graded meshes with barycentric rational interpolation for solving a*

class of time-dependent partial integro-differential equations with weakly singular kernels, Comput. Math. Appl. 141, (2023) 159–169.

[6] Rezazadeh, T. and Najafi, E. *Jacobi collocation method and smoothing transformation for numerical solution of neutral nonlinear weakly singular Fredholm integro-differential equations*, Appl. Numer. Math. 181 (2022) 135–150.

[7] Rosa, M.A., Cuminato, J.A. and McKee, S. *A polynomial collocation method for singular integro-differential equations in weighted spaces*, J. Comput. Appl. Math. 368 (2020) 112526.

[8] Singh, V.K. and Postnikov, E.B. *Operational matrix approach for solution of integro-differential equations arising in theory of anomalous relaxation processes in vicinity of singular point*, Appl. Math. Model. 37(10) (2013) 6609–6616.

[9] Wang, Y.M. and Zhang, Y.J. *A Crank-Nicolson-type compact difference method with the uniform time step for a class of weakly singular parabolic integro-differential equations*, Appl. Numer. Math. 172 (2022) 566–590.

[10] Xu, D. *Numerical solution of partial integro-differential equation with a weakly singular kernel based on Sinc methods,* Math. Comput. Simul. 190 (2021) 140–158.

[11] Zemlyanova, A.Y. and Machina, A. *A new B-spline collocation method for singular integro-differential equations of higher orders*, J. Comput. Appl. Math. 380 (2020) 112949.

[12] Zheng, Z.Y. and Wang, Y.M. *A second-order accurate Crank–Nicolson finite difference method on uniform meshes for nonlinear partial integro-differential equations with weakly singular kernels*, Math. Comput. Simul. 205 (2023) 390–413,

[13] Zhou, Y. and Stynes, M. *Block boundary value methods for solving linear neutral Volterra integro-differential equations with weakly singular kernels*, J. Comput. Appl. Math. 401 (2022) 113747.

Iranian Journal of Numerical Analysis and Optimization
Vol. 15, No. 2, 2025, pp 475–507
https://doi.org/10.22067/ijnao.2025.89754.1508
https://ijnao.um.ac.ir/

Research Article

Unite and conquer approach for data clustering based on particle swarm optimization and moth flame optimization

E. Mosavi, S.A. Shahzadeh Fazeli*, E. Abbasi and F. Kaveh-Yazdy

Abstract

Data clustering is a widely used technique in various domains to group data objects according to their similarity. Clustering molecules is a useful

*Corresponding author

Received 10 September 2024; revised 28 December 2025; accepted 15 January 2025

Esra Mosavi
Department of Computer Science, Yazd University, Yazd, Iran.
Parallel Processing Laboratory, Yazd University, Yazd, Iran. e-mail: esramosavi@stu.yazd.ac.ir

Seyed Abolfazl Shahzadeh Fazeli
Department of Computer Science, Yazd University, Yazd, Iran.
Parallel Processing Laboratory, Yazd University, Yazd, Iran.e-mail: fazeli@yazd.ac.ir

Elham Abbasi
Department of Computer Science, Yazd University, Yazd, Iran. e-mail: e.abbasi@yazd.ac.ir

Fatemeh Kaveh-Yazdy
Researcher at Oncober, Basel, Switzerland. e-mail: f.kaveh@oncober.ch

How to cite this article
Mosavi, E., Shahzadeh Fazeli, S.A., Abbasi, E. and Kaveh-Yazdy, F., Unite and conquer approach for data clustering based on particle swarm optimization and moth flame optimization. *Iran. J. Numer. Anal. Optim.*, 2025; 15(2): 475-507. https://doi.org/10.22067/ijnao.2025.89754.1508

process where you can easily subdivide and manipulate and large datasets to group compounds into smaller clusters with similar properties. To discover new molecules with optimal properties and desired biological activity, can be used by comparing molecules and their similarities. A prominent clustering technique is the k-means algorithm, which assigns data objects to the nearest cluster center. However, this algorithm relies on the initial selection of the cluster centers, which can affect its convergence and quality.

To address this issue, metaheuristic algorithms have been proposed as a type of approximate optimization algorithm capable of identifying almost optimal solutions. In this paper, a new meta-heuristic approach is proposed by combining two algorithms of particle swarm optimization (PSO) and moth flame optimization (MFO), following that, it is used to improve data clustering. The efficiency of the proposed approach is evaluated utilizing benchmark functions F1-F23. Its efficiency is evaluated with PSO and MFO algorithms on different datasets. Our experiential results show that the suggested approach exceeds the PSO and MFO algorithms with respect to speed of convergence and clustering quality.

AMS subject classifications (2020): 68T10; 62H30.

Keywords: Data clustering; k-means clustering; Metaheuristic optimization; Particle swarm optimization; Moth flame optimization.

1 Introduction

Clustering methods have become increasingly popular in various fields of science and engineering, making it an important field for data mining study. Using this method, objects are grouped into clusters where they have the least amount of distance between each object and the cluster's center. Ideally, objects within a cluster should be as similar as possible and distinct from those in other clusters. As a powerful tool in chemistry, clustering helps to analyze complex data and make decisions in various processes. Some of these applications such as analysis of chemical compounds, medicinal chemistry, analysis of spectra, and design of new materials can be mentioned. Clustering can be used to group chemical compounds with similar properties, to

identify chemical compounds with similar biological properties, or to analyze spectral data by identifying compounds. It can also design new materials with desirable properties through the grouping of compounds [44]. Several algorithms exist for clustering. One of the most popular clustering techniques with a wide range of applications is the k-means algorithm [20, 27]. Extensive research has been conducted on k-means clustering, leading to various extensions, and it has been applied in various fundamental domains [4, 26, 29, 47]. However, this algorithm's performance depends heavily on the selection of the centers, making it susceptible to getting stuck in a local optimum. One way to overcome this problem is using optimization algorithms for clustering. Optimization involves finding the best possible solutions to a specific problem. Mathematical optimization techniques are deterministic but unable to avoid local optima. However, random optimization algorithms (metaheuristics) are much more efficient at avoiding local optimization by randomly exploring the search domain. While these algorithms can capture a certain pattern, their aim is to find an approximate solution that is close to the exact solution. Several approximate approaches are known as global search approaches referred to as metaheuristics. These methods have been designed to overcome the limitations of traditional techniques. Metaheuristic algorithms utilize the intelligence observed in natural phenomena to find solutions that are either near-global or global. Over the past few decades, numerous metaheuristic algorithms have been presented, drawing inspiration from natural processes, collective behavior, art, physics, and mathematical principles. Such as whale optimization algorithm (WOA) [31], particle swarm optimization (PSO) [11], grey wolf optimizer (GWO) [32], moth flame optimization (MFO) [30], cuckoo search (CS) [15], ant colony optimization [10], social spider optimization [25], mountain gazelle optimizer [1], dwarf mongoose optimization [2]. These studies showed that meta-heuristic algorithms can solve complex optimization problems with high accuracy and reasonable execution time. These algorithms have several advantages, such as simplicity of implementation, suitable accuracy and reasonable execution time. In clustering, which is considered as a problem of optimization with the aim of minimizing the sum of intra-cluster distances, choosing the appropriate meta-

heuristic algorithm is very important because the search domain is large and the goal of distance calculation is nonlinear [5, 6, 23, 28, 33, 40, 41].

The efficacy of MFO in producing a quick convergence rate and population diversity in clustering problems has been proven [42]. The PSO includes a built-in guidance strategy that helps solutions learn from better ones, improving their own solutions [21]. This article proposes a data clustering technique based on a combination of MFO and PSO called MFO_PSO_CLUST. The potential to avoid local optima and quickly converge to the global optimal solution is one of the main algorithm's features. Our main purpose is to cluster data using MFO_PSO_CLUST, aiming for more thorough search space coverage and improved accuracy compared to previously used methods. In brief, we will describe the new findings and contributions presented in this paper.

- The proposed approach MFO_PSO is a combination of PSO and MFO algorithms.
- Six datasets are utilized for evaluating the proposed approach: Iris, Wine, Banknote_authentication, Vowel, Glass, and Zoo.
- The proposed method's solution quality is evaluated against the MFO and PSO algorithms.
- The 23 benchmark functions are used to undertake tests and evaluations.
- MFO_PSO_CLUST method is presented for data clustering and evaluated in terms of FMI, NMI, and *Silhouette_score* criteria.
- The performance of the proposed approach is justified based on convergence curves and experimental results.

The sections of the article are organized as follows: Section 2 examines previous literature, while Section 3 introduces the main terms and definitions used in the paper. Section 4 presents the proposed approach based on MFO and PSO algorithms and utilizes it for the k-means clustering method. Section 5 includes experimental analysis using benchmark functions and the different datasets, demonstrating that the suggested approach is effective. The conclusion is discussed in Section 6.

2 Literature review

Several studies have been conducted in the field of data clustering; some of which are mentioned below. Saida, Nadjet, and Omar [36] proposed a data clustering with the CS optimization algorithm while Esmin, Coelho, and Matwin [13] proposed using the PSO as a globalized search algorithm for faster clustering convergence. However, complexity remains a significant challenge with this approach. Nasiri and Khiyabani [34] introduced a clustering method with the WOA, inspired by humpback whales' foraging behavior, and a technique for improving PSO using Renyi entropy clustering was presented by Emre Comak [9]. It sorts particles based on the entropy parameter to achieve the best outcomes. But when used to bigger datasets, this method becomes too complex and is not flexible enough to adjust to changing weights for the Renyi entropy parameter. Therefore, it is only appropriate for smaller datasets.

Qin-Hu Zhang et al. [46] presented the multivariant optimization algorithm, designed for global search to find optimal solutions. This method maintains the best cluster center values, offering stability and accuracy in addressing clustering challenges. However, its performance becomes more complex with increased dimensional size. Kumar and Sahoo [22] introduced a hybrid metaheuristic algorithm, MCSS-PSO, which combines MCSS and PSO algorithms. This approach improves exploration capabilities and ensures convergence at the global minimum instead of local minima. Indeed, its efficiency and robustness are primarily applicable to partition-based clustering methods. Jadhav and Gomathi [17] presented the KEGWO method, which efficiently calculates cluster centroids, while being excellent at identifying ideal centroids. However, this method may not be suitable for complex clustering. Jadhav and Gomathi [18] proposed the WGC algorithm to determine the appropriate center for the clustering algorithm. The WGC algorithm employs the new fitness function and hybrids the WOA and WEGWO. The WEGWO algorithm is a combination of the EGWO and the WOA algorithms, where the weight-based position update added to the GWO creates the EGWO.

Alswaitti, Albughdadi, and Isa [7] proposed a variance-based differential evolution algorithm for data clustering that improves both quality and convergence speed and includes an optional crossover, and an approach to improve clustering problems using the bio-inspired cuttlefish algorithm, aiming to find the best cluster centers to minimize clustering metrics was presented in [12]. Singh [39] presented WOA approach for data clustering because of its capacity to maintain search space diversity and its positive rate of convergence, and Singh et al. [42] proposed a novel method for solving data clustering problems that is based on the MFO. In addition, Kumar and Kumar [21] presented a fuzzy clustering technique aimed at enhancing the convergence performance of clustering methods. It enhances fuzzy c-means by creating a new measure that can withstand noisy surroundings. The PSO method is applied to solve the fuzzy c-means initialization problem. Preliminaries will be introduced, in the next section.

3 Preliminaries

Clustering is an unsupervised learning technique that helps organize and comprehend big data sets by grouping data into homogeneous clusters based on patterns and similarities. The main purpose of this classification is to put similar data in a cluster, while the data in different clusters have less similarity so that more appropriate information can be obtained from them [3, 8, 19]. There are different types of clustering. One of the most famous algorithms in clustering and classification is the k-means. The first step of this technique is to choose k points at random to serve as the initial cluster centers. Then, in an iterative loop, it finds the closest center for each object and assigns the object to that group. After, all objects are assigned to an appropriate group, the center of each group is recomputed. Until the centers converge, this process is repeated [14]. The k-means algorithm is indicated in Algorithm 1.

The primary challenge faced here is the identification of the best centroid. Many optimization algorithms are utilized for this purpose. Next, two types of these algorithms will be described.

Algorithm 1 k-means algorithm

Initialize k centroids randomly
while cluster assignments change **do**
  for each data point x_i **do**
    Assign x_i to the closest centroid c_j
  end for
  for each centroid c_j **do**
    Update the position of c_j to the mean position of all data points assigned to it
  end for
end while

3.1 Particle swarm optimization

The PSO is an approach for meta-heuristic optimization that is utilized for solving optimization problems. The algorithm starts by generating a population of particles, each particle representing a potential solution to the optimization problem. Each particle p_i with a specified position and a velocity, moves in the search space for the optimal solution. For each particle, the best position is the location with the best fitness value that the particle has yet encountered, and the global best position is the position with the best fitness value that any particle might find in the population.

The following equations are utilized to update the particle's position and velocity:

$$q_i(t+1) = wq_i(t) + c_1r_1(pbest_i - P_i(t)) + c_2r_2(gbest - P_i(t)), \tag{1}$$

$$P_i(t+1) = P_i(t) + q_i(t+1), \tag{2}$$

where $q_i(t+1)$ is the velocity of particle p_i at iteration $t+1$, $P_i(t)$ is the position of particle p_i at iteration t, $pbest_i$ is the personal best position of particle p_i, $gbest$ is the global best position of the population, w is a constant referred to as the inertia weight, c_1 and c_2 are acceleration constants, and r_1 and r_2 are random numbers in the range $[0, 1]$ [11, 38]. The PSO algorithm is shown in Algorithm 2.

Algorithm 2 PSO algorithm

Initialize the swarm of particles S with n particles
Initialize the best position of each particle $pbest_i$ to its initial position
Initialize the best position of the swarm $gbest$ to the best position of any particle in S
Initialize the velocity of each particle p_i, q_i to 0
while stopping criterion not met **do**
  for each particle p_i in S **do**
    Update the particle p_i's velocity with (1)
    Update the particle p_i's position with (2)
    if fitness of the new position is an improvement over the fitness of $pbest_i$ **then**
      Update $pbest_i$ to the new position
    end if
    if fitness of the new position is an improvement over the fitness of $gbest$ **then**
      Update $gbest$ to the new position
    end if
  end for
end while

The simplicity and efficiency of PSO have made it a popular choice for optimization problems in many fields of research and industry.

3.2 Moth flame optimization

The MFO algorithm is a metaheuristic optimization algorithm inspired by the behavior of moths around a flame. The algorithm was presented by Mirjalili [30] in 2015. The foundational idea of MFO is to simulate the behavior of moths around a flame to optimize a given function. The flame represents the optimal solution, and the moths move towards the flame while also keeping exploration and exploitation in balance. The MFO algorithm consists of three phases: initialization, moth movement, and flame updating. In the

first phase, the population of moths is randomly generated. It is shown in a matrix size of $n \times d$,

$$m_i = \vec{l} + \beta.(\vec{u} - \vec{l}). \tag{3}$$

The variable m_i represents the ith solution, $(m_{i,1}m_{i,2}\dots m_{i,d})$, where i ranges from 1 to n, the population's size. The value of β is a randomly generated number between 0 and 1. The variable $\vec{l}$ represent the lower bound and the variable $\vec{u}$ represent upper bound for the problem being considered. The operator "." represents a point-by-point multiplication of the vectors. Each moth is evaluated using fitness value and included in a matrix of size $n \times 1$ called OM. Flame is another component in the MFO algorithm. They are also included in a matrix of size $n \times d$ called F. Every flame's fitness is assessed and added to an $n \times 1$ matrix called OF. In the moth movement phase, each moth searches for the flame by following a certain set of rules, which are based on the distance between the flame and the moth.

$$L(M_i, F_j) = D_i e^{bt} \cos(2\pi t) + F_j, \tag{4}$$

where D_i represents the distance from the ith moth to the jth flame, t is a random number in the range $[-1, 1]$, and for defining the shape of the logarithmic spiral of a moth, has been used b constant. Here, the calculation of D_i is as follows:

$$D_i = |F_j - M_i|, \tag{5}$$

where M_i denotes the ith moth and F_j denotes the jth flame. By updating their positions at n different locations in the search space, the moths can reduce the exploitation of the most promising solutions. Thus, to mitigate this effect, using the following formula, the number of flames reduces throughout the course of iterations:

$$\text{flameno} = \text{round}\left(F_j - It \times \frac{(F_j - 1)}{Maxit}\right), \tag{6}$$

where It represents the current iteration, $Maxit$ denotes the most number of iterations, and F_j corresponds to the maximum number of flames [37]. Until the termination condition is satisfied, the process continues. Finally, in the flame updating phase, the position of the flame is updated based on the

position of the moths [30, 42]. The MFO algorithm is shown in Algorithm 3.

Algorithm 3 MFO algorithm

Initialize population $M = \{M_{n,1}, M_{n,2}, \ldots, M_{n,d}\}$ randomly.
Update flameno using (6).
(OM), Evaluate the fitness of each moth $M_{n,i}$ using the objective function.
while stopping criterion not met **do**
  $F = sorted(M)$;
  $OF = sorted(OM)$;
  for each moth $M_{n,i}$ in M **do**
    Calculate D_i using (5) for the corresponding moth;
    Update Moth(i) using (4) for the corresponding moth;
  end for
  Increment the iteration counter t;
end while
$F = sorted(M)$;
Return the best moth $F(0)$;

While the original PSO and MFO algorithms have demonstrated commendable optimization performance, they still encounter certain limitations, such as becoming trapped in local optima and exhibiting slow convergence. Therefore, we have presented an algorithm, which is the result of hybridizing the PSO and MFO algorithms, as a solution to this problem. The next section provides an explanation for this.

4 Proposed method

Today, numerous metaheuristic algorithms exist. The no free lunch theorem (NFL) [45] explicates the reason for the existence of numerous methods for optimization. NFL is a basic concept in machine learning and optimization, stating that there is not a single method for solving all optimization or machine learning problems. This principle emphasizes that not a single algorithm suits every optimization or machine learning task. In simpler terms,

one algorithm may excel at solving one type of problem but not be as effective for another type. In consideration of the theorem above as well as the MFO and PSO algorithms, we made the decision to present a novel approach to incorporate these two techniques. When compared to the PSO and MFO methods, this method expressed better convergence and increased efficiency. The operation of this proposed approach will be explained in the next section. The proposed approach begins with the MFO algorithm. After this process, the optimum result determined by the objective function value is compared with the best solution. If it is improved, then the best solution is updated accordingly. Following that, the MFO algorithm's top s results are chosen and incorporated as input to the PSO algorithm to some extent, ensuring that the best solutions are further improved. The value of s is selected at random within the range of 3 particles and $\frac{1}{4}$ of the initial population. The chosen results replace the random initial population in PSO. The best s results from the MFO are assigned at the start of the initial population list in PSO, instead of creating a completely random starting population. Then, randomly generated solutions are inserted into the remaining spaces. After that, PSO is carried out with this improved initial population. Now, it is the turn of the PSO algorithm to run. The output produced by PSO is evaluated, and once it completes, the best result from PSO is compared with the best solution found overall. The best solution is updated once again if an improvement is found. Choose the top s outcomes from the optimal solutions determined by PSO. Following that, these values are added back into the MFO algorithm. The MFO algorithm's initial population is established through the utilization of a similar approach to PSO. The cluster centers are the end result of this iterative process, which continues until the required number of repeats is obtained. In summary, each iteration of the algorithm starts with the previous one and uses the output of the previous step as the input for the subsequent one. By investigating the results and taking consideration of the objective function value, the algorithm identifies the best cluster centers for the data. Algorithm 4 presents the suggested algorithm, and Figure 1 illustrates the method used to improve k-means clustering.

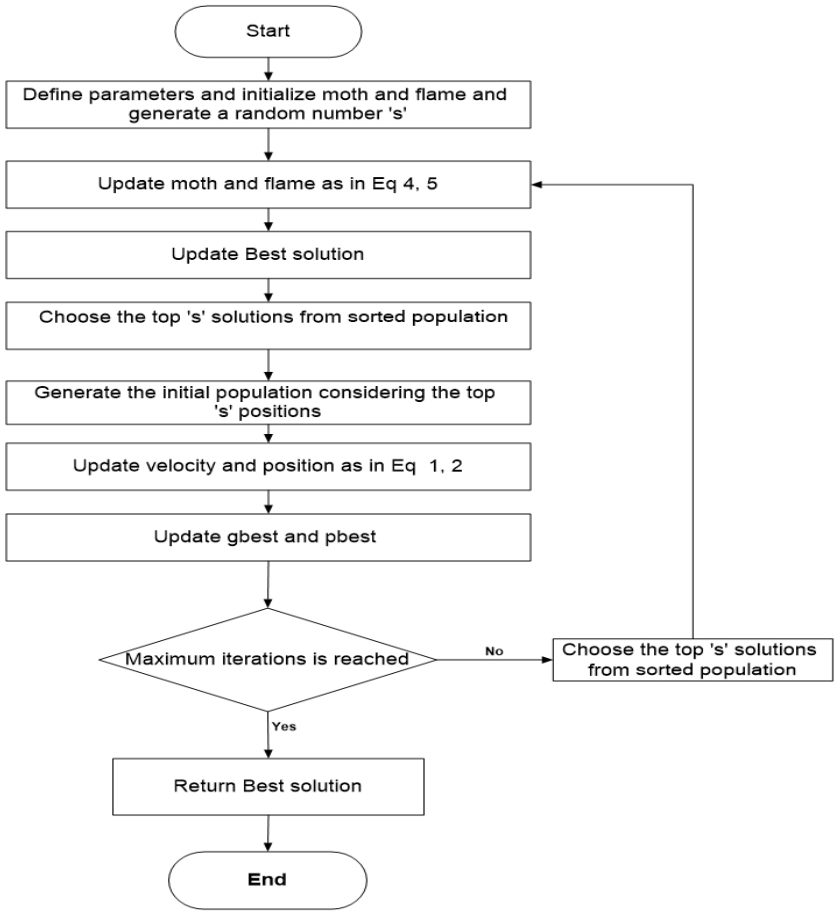


Figure 1: MFO_PSO flowchart

4.1 Clustering and the proposed method

To improve clustering, it is essential to find the most appropriate centroid of clusters that minimize intra-cluster distance and maximize the distance between clusters. It is possible to use optimization methods to identify the optimal cluster centers. As mentioned earlier, an optimization algorithm is a collection of algorithms designed to determine the best feasible solution or values for certain parameters. The aim of optimization while meeting certain constraints is to either reduce or increase an objective function. When employed for this purpose, rather than the global optimum, many optimization

Algorithm 4 MFO_PSO algorithm

Create a random number s and the population X.
while Stopping criteria not achieved **do**
  if iter = 1 **then**
    Use the MFO algorithm.
    if new solution's fitness is an improvement over the *best_solution*'s fitness **then**
      Update best_solution
    end if
    Best_position is the top s solutions from sorted population (F);
  else
    $\{M_{n,1}, M_{n,2}, \ldots, M_{n,s}\} = Best\_position$;
    Initialize population $\{M_{n,s+1}, M_{n,s+2}, \ldots, M_{n,d}\}$ randomly;
    $M = \{M_{n,1}, M_{n,2}, \ldots, M_{n,s}, M_{n,s+1}, \ldots, M_{n,d}\}$;
    calculate the fitness of each moth $M_{n,i}$
    for each moth $M_{n,i}$ in M **do**
      Calculate D_i using (5)
      Update Moth(i) using (4)
    end for
    if the new solution's fitness is superior to the *best_solution*'s fitness **then**
      Update *best_solution*;
    end if
    $\{p_1, p_2, \ldots, p_s\} = Best\_position$;
    Initialize population $\{p_{s+1}, p_{s+2}, \ldots, p_n\}$ randomly;
    $S = \{p_1, p_2, \ldots, p_s, p_{s+1}, \ldots, p_n\}$;
    for each particle p_i in S **do**
      Update the particle p_i's velocity with (1)
      Update the particle p_i's position with (2)
      if the new position's fitness is an improvement over the $pbest_i$'s fitness **then**
        Update $pbest_i$;
      end if
      if the new position's fitness is an improvement over the *gbest*'s fitness **then**
        Update *gbest*;
      end if
    end for
    $Best\_position = \{pbest_1, pbest_2, \ldots, pbest_s\}$;
    if new solution's fitness is an improvement over the *best_solution*'s fitness **then**
      Update *best_solution*;
    end if
    iter=iter+1;
  end if
end while

approaches usually converge to a local optimum. Consequently, the necessity arises for an efficient optimization algorithm. Here, the focus is on utilizing a unite and conquer approach of optimization algorithms to identify centroids that minimize the distance from the remaining data points and achieve optimal performance. In this article, the proposed approach was utilized, which

hybridizes the PSO and MFO algorithms, to improve clustering. The following explains the operation of the MFO_PSO_CLUST algorithm.

The goal of clustering is to minimize the intra-cluster distance while maximizing the inter-cluster distance. The objective function is expressed as:

$$Fit = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - c_i\|^2 \tag{7}$$

where x presents data points, c_i is cluster centroid, and C_i is data points in cluster i. The MFO_PSO algorithm aims to minimize Fit by optimizing the placement of centroids. MFO_PSO generates an initial population of candidate solutions, where each solution represents a set of centroids. The objective function Fit is computed for each solution to assess its clustering performance. Next, the MFO_PSO algorithm is executed, leading to the optimization of the centroids. The algorithm iterates until the maximum number of iterations is reached or minimal improvement in Fit is observed. The optimized centroids obtained from MFO_PSO are used to initialize k-means, ensuring high-quality clustering. Algorithm 5 displays the MFO_PSO_CLUST algorithm.

5 Results and discussion

Several widely used benchmark functions, each of which may display a distinct set of MFO_PSO_CLUST abilities, have been used to assess the efficiency of the MFO_PSO_CLUST algorithm. With only one global optimum point, the benchmark functions (F1-F7) show the optimization algorithms' exploitative powers. The definition of optimization techniques for exploration is done using standard functions (F8-F23). It is worth mentioning that the evaluation functions (F1-F23) are sourced from CEC 2005 special session [24, 43]. Tables 1, 2, and 3 display the functions properties and mathematical formulas. The MFO_PSO_CLUST algorithm was tested on a system comprising MATLAB R2022a and Windows 10 Enterprise 64-bit operating system. The hardware used included 4.00 GB of RAM and an Intel Core(TM) i5-3210M 2.50GHz CPU. Similar evaluations were conducted for MFO and PSO optimization algorithms. In these assessments, the

Algorithm 5 MFO_PSO_CLUST algorithm

Input: Data set $\{x_1, x_2, \ldots, x_n\}$, number of clusters k, maximum iterations T

Output: Optimized cluster centroids

Initialize population with candidate solutions (set of centroids) randomly

for each iteration $t = 1 \ldots T$ **do**

for each solution in population **do**

Evaluate the objective function Fit for the current centroid configuration

end for

Optimization Phase:

for each moth in population **do**

Update the position of the particle using the MFO_PSO algorithm towards optimal centroids

end for

if stopping criteria met (e.g., max iterations or minimal improvement in Fit) **then**

Exit the loop

end if

end for

Finalizing Clusters:

Use the optimized centroids obtained from MFO_PSO to initialize k-means

Run k-means with the optimized centroids to finalize clustering

MFO_PSO_CLUST algorithm was subjected to 30 populations, each with a maximum of 600 iterations. The results gained from 10 autonomous results were applied to compare the findings. The parameters of corresponding optimization methods were set using the settings available in the MFO [30] and PSO [11]. Details of the parameters that these optimization techniques used are presented in Table 4. In Table 2, the values of y_i and $Q(x_i, a, k, m)$ are computed as follows:

$$y_i = 1 + \frac{x_i + 1}{4},$$

$$Q(x_i, a, k, m) = \begin{cases} k(x_i - a)^m & \text{if } x_i > a \\ 0 & \text{if } -a < x_i < a \\ k(-x_i - a)^m & \text{if } x_i < -a \end{cases}.$$

Table 1: Information on unimodal benchmark functions [24, 43].

No	Function	Dimensions	Range	Fmin
F1	$f(x) = \sum_{i=1}^{d} x_i^2$	30	$[-100, 100]^d$	0
F2	$f(x) = \sum_{i=1}^{d} \lvert x_i \rvert + \prod_{i=1}^{d} \lvert x_i \rvert$	30	$[-10, 10]^d$	0
F3	$f(x) = \sum_{i=1}^{d} (\sum_{j=1}^{i} x_j)^2$	30	$[-100, 100]^d$	0
F4	$f(x) = \max_i \{\lvert x_i \rvert\}, 1 \leq i \leq d$	30	$[-100, 100]^d$	0
F5	$f(x) = \sum_{i=1}^{d-1} [100(x_{i+1} - x_i^2)^2 + (x_i - 1)^2]$	30	$[-30, 30]^d$	0
F6	$f(x) = \sum_{i=1}^{d} (\lvert x_i + 0.5 \rvert)^2$	30	$[-100, 100]^d$	0
F7	$f(x) = \sum_{i=1}^{d} i x_{4i} + \text{random}[0, 1)$	30	$[-128, 128]^d$	0

5.1 Qualitative and quantitative MFO_PSO results

In this section, MFO_PSO quality was assessed using six common benchmark functions (F1, F3, F7, F8, F12, and F13). Three main factors were examined: convergence behavior, average population fitness, and changes in the first dimension of the first particle. During optimization, the diagram of convergence behavior tracked the cost of the best solution, while the average population fitness curve displayed how the cost for each search parameters changed during the optimization procedure. Moreover, the modifications made to the first particle were shown in the plots for its first dimension. The convergence diagrams for the best solution during optimization are shown in Figure 2. They demonstrate that MFO_PSO exhibits strong convergence capabilities with a fast decrease in cost.

Table 2: Information on multimodal benchmark functions [24, 43].

No	Function	Dimensions	Range	Fmin
F8	$f(x) = \sum_{i=1}^{d} \left(x_i \sin\left(\sqrt{\lvert x_i \rvert} \right) \right)$	30	[-500, 500]d	$-418.9829 \times n$
F9	$f(x) = 10d + \sum_{i=1}^{d} \left[x_i^d - 10\cos(2\pi x_i) \right]$	30	[-5.12, 5.12]d	0
F10	$f(x) = -20\exp\left(-0.2\sqrt{\frac{1}{d}\sum_{i=1}^{d} x_i^2} \right) - \exp\left(\frac{1}{d}\sum_{i=1}^{d} \cos(2\pi x_i) \right) + 20 + e$	30	[-32, 32]d	0
F11	$f(x) = \frac{1}{4000}\sum_{i=1}^{d} x_i^2 - \prod_{i=1}^{d} \cos\left(\frac{x_i}{\sqrt{i}} \right) + 1$	30	[-600, 600]d	0
F12	$f(x) = \frac{\pi}{d}\left\{ 10\sin(\pi y_1) + \sum_{i=1}^{d-1} (y_i - 1)^2 \left[1 + 10\sin^2(\pi y_{i+1}) \right] + (y_d - 1)^2 \sum_{i=1}^{d} Q(x_i, 10, 100, 4) \right\}$	30	[-50, 50]d	0
F13	$f(x) = 0.1\left\{ \sin^2(3\pi x_1) + \sum_{i=1}^{d} (x_i - 1)^2 \left[1 + \sin^2(3\pi x_i + 1) \right] + (x_d - 1)^2 \right\} + \sum_{i=1}^{d} Q(x_i, 5, 100, 4)$	30	[-50, 50]d	0

Table 3: Information on benchmark functions [24, 43].

No	Function	Dimensions	Range	Fmin
F14	$f(x) = \left[\frac{1}{500} + \sum_{i=1}^{25} \frac{1}{i+\sum_{j=1}^{2}(x_j - a_{j,i})^6}\right]^{-1}$	30	$[-65, 65]^d$	-1
F15	$f(x) = \sum_{i=1}^{d} [a_i - \frac{x_1(b_{2i}+b_i x_2)}{b_i^2 + b_i x_3 + x_4}]^2$	30	$[-5, 5]^d$	0.00030
F16	$f(x) = 4x_1^2 - 2.1x_1^4 + \frac{1}{3}x_1^6 + x_1x_2 - 4x_2^2 + 4x_2^4$	30	$[-5, 5]^d$	-1.0316
F17	$f(x) = \left(x_2 - \frac{5.1}{4\pi^2}x_1^2 + \frac{5}{\pi}x_1 - 6\right)^2 + 10\left(1 - \frac{1}{8\pi}\right)\cos x_1 + 10$	30	$[-5, 5]^d$	0.398
F18	$f(x) = \left[1 + (x_1 + x_2 + 1)^2(19 - 14x_1 + 3x_1^2 - 14x_2 + 6x_1x_2 + 3x_2^2)\right]$ $\times \quad \left[30 + (2x_1 - 3x_2)^2 \times (18 - 32x_1 + 12x_1^2 + 48x_2 - 36x_1x_2 + 27x_2^2)\right]$	30	$[-2, 2]^d$	3
F19	$f(x) = -\sum_{i=1}^{4} a_i \exp\left(-\sum_{j=1}^{3} b_{ij}(x_j - p_{ij})^2\right)$	30	$[1, 3]^d$	-3.86
F20	$f(x) = -\sum_{i=1}^{4} a_i \exp\left(-\sum_{j=1}^{6} b_{ij}(x_j - p_{ij})^2\right)$	30	$[0, 1]^d$	-3.32
F21	$f(x) = -\sum_{i=1}^{5} \left[(X - a_i)(X - a_i)^T + c_i\right]^{-1}$	30	$[0, 10]^d$	-10.1532
F22	$f(x) = -\sum_{i=1}^{7} \left[(X - a_i)(X - a_i)^T + c_i\right]^{-1}$	30	$[0, 10]^d$	-10.4028
F23	$f(x) = -\sum_{i=1}^{10} \left[(X - a_i)(X - a_i)^T + c_i\right]^{-1}$	30	$[0, 10]^d$	-10.5363

Table 4: Parameters for comparing and evaluating optimization algorithms.

Algorithm	Parameter	Value
PSO	v_{Max}	0.1
	c_1	2
	c_2	2
MFO	Convergence constant a	[-2, -1]
	Logarithmic spiral constant b	1
MFO_PSO_CLUST	v_{Max}	0.1
	c_1	1.496
	c_2	1.496
	Convergence constant a	[-2, -1]
	Logarithmic spiral constant b	1

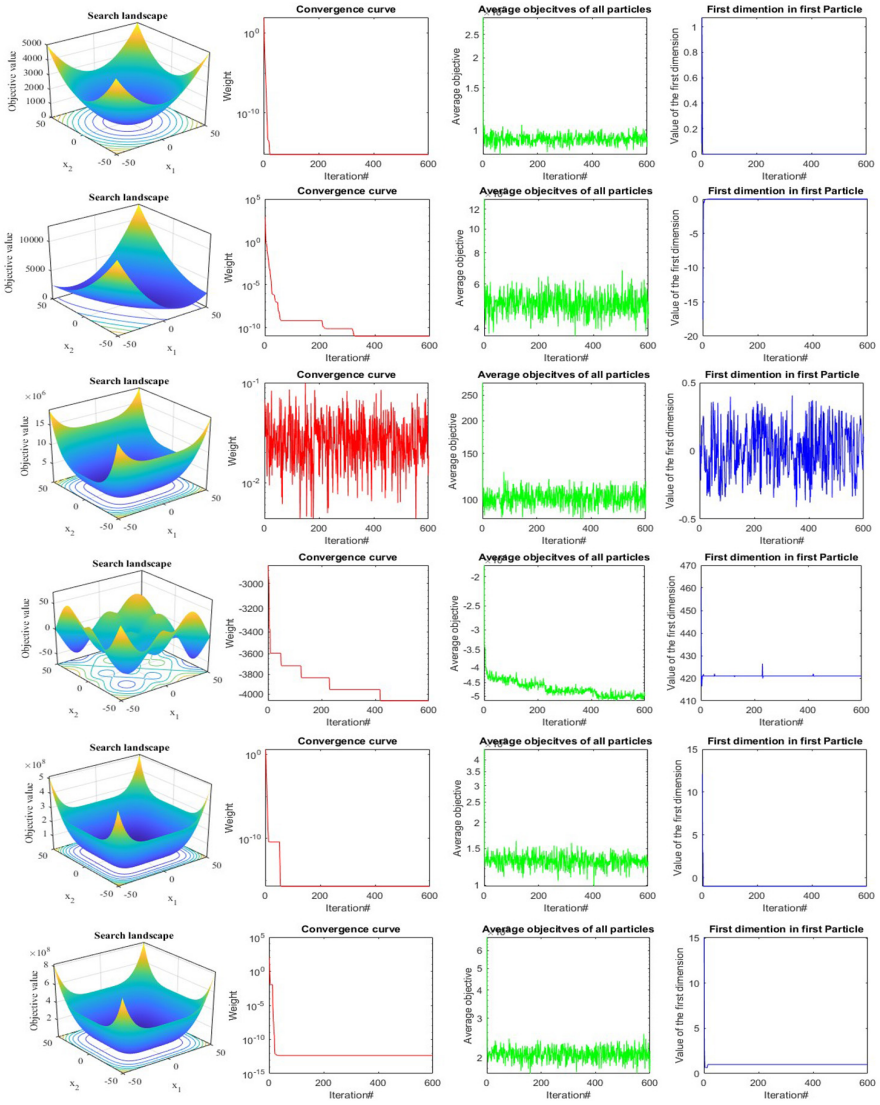


Figure 2: The qualitative results of F1, F3, F7, F8, F12, and F13 function using MFO_PSO algorithm

The quantitative efficiency of MFO_PSO was tested and assessed. The results achieved by MFO_PSO were evaluated with the results of two other optimization methods. The testing utilized 23 benchmark functions, including scalable F1-F23 functions. The tests included results from 10 independent

iterations and included the AVG, Worst, and Best mean error criterion. The outcomes of scalability evaluations for the F1-F23 functions, the MFO_PSO algorithm, and associated algorithms are indicated in Table 5. MFO_PSO performed better than two other optimization algorithms (MFO, PSO) across various sizes and the majority of benchmark functions. This is highlighted in Table 5, where the results of MFO_PSO are in comparison to those of MFO and PSO algorithms for the benchmark functions F1-F23 with a dimension of 30.

5.2 Experiential results of the MFO_PSO_CLUST algorithm

This section includes multiple examples to show the efficiency of our proposed approach for improving k-means clustering by comparing it with other methods. We utilized this hybrid optimization method to Iris, Wine, Banknote_authentication, Vowel, Glass and Zoo datasets and evaluated its efficiency in finding optimal solutions.

Table 6 evaluates the fitness values for different datasets, where the MFO_PSO_CLUST, MFO_CLUST, and PSO_CLUST algorithms were utilized. The fitness values, including the best, mean, and worst outcomes, were analyzed for each dataset with the corresponding number of clusters. Each algorithm was examined using 30 populations, with each run having a maximum of 500 iterations and 10 independent iterations for each dataset. The results reported in the table represent the average values obtained from 10 iterations.

For the Iris dataset, consistent fitness values 96.6556 were achieved by the MFO_PSO_CLUST algorithm across the best, mean, and worst cases, indicating stable performance. In contrast, a significantly higher mean fitness value 205.9479 was observed for MFO_CLUST, however, its variability (best: 195.5371, worst: 212.1155) suggested less stability. Average performance with slightly higher variability (mean: 135.4312, worst: 146.0318) was demonstrated by PSO_CLUST. In the Banknote_authentication dataset, the best fitness value (7200.6576) was achieved by the MFO_PSO_CLUST

Table 5: Benchmark function results (F1-F23) using 30 dimensions.

No.Function	Type	MFO_PSO	PSO	MFO
F1	Best	4.7922E-21	7.1071E-09	5.3715E-01
	Worst	6.6727E-11	7.3614E-05	2.0002E+04
	Mean	1.1744E-11	4.4590E-06	1.6722E+03
F2	Best	2.9181E-12	5.6941E-06	1.7467E-01
	Worst	2.3649E-06	2.0504E-02	8.0223E+01
	Mean	5.8226E-07	3.3272E-03	3.5329E+01
F3	Best	3.4763E-12	1.8171E+01	2.3463E+03
	Worst	8.0948E-09	3.4851E+03	4.7370E+04
	Mean	1.5474E-09	5.8934E+02	2.0003E+04
F4	Best	5.1595E-06	2.6406E-01	4.7078E+01
	Worst	2.9516E-05	2.1015E+00	8.5407E+01
	Mean	1.4508E-05	7.9578E+00	6.9113E+01
F5	Best	6.4368E-04	1.9874E+01	1.3794E+02
	Worst	7.6082E+00	1.0846E+02	8.0123E+07
	Mean	8.607E-01	4.6616E+01	2.6939E+06
F6	Best	1.0626E-20	6.7447E-09	7.4584E-01
	Worst	2.2514E-10	4.1088E-05	1.9801E+04
	Mean	4.1271E-11	2.9114E-06	2.6696E+03
F7	Best	1.6307E-02	4.0726E-02	4.0524E-02
	Worst	6.9200E-01	1.5602E-01	5.9234E+01
	Mean	3.5400E-02	9.5056E-02	2.7669E+02
F8	Best	-4.1898E+03	-3.315E+03	-1.0353E+04
	Worst	-3.8345E+03	-1.949E+03	-6.3081E+03
	Mean	-4.0359E+03	-2.590E+03	-8.5561E+03
F9	Best	0	1.9899E+01	1.0158E+02
	Worst	4.0956e-11	7.2632E+01	2.6097E+02
	Mean	7.0031e-12	3.8671E+01	1.5809E+02
F10	Best	2.1231e-08	3.7741E-06	1.4912E+00
	Worst	4.6272e-06	2.4083E+00	1.9964E+01
	Mean	1.4335e-06	3.1302E-01	1.5907E+01
F11	Best	2.4605E-02	6.3190E+01	6.3238E-01
	Worst	9.8345E-02	1.0377E+02	9.1184E+01
	Mean	4.9904E-02	8.3190E+01	3.9966E+01
F12	Best	2.2284e-22	9.3714E-11	6.5723E-01
	Worst	7.4927e-13	1.5674E+00	3.4177E+02
	Mean	1.8579e-13	2.4928E-01	2.1472E+01
F13	Best	1.5237e-17	2.9246E-11	2.9276E+00
	Worst	3.9970e-12	1.1006E-02	4.1582E+03
	Mean	8.6499e-13	2.5662E-03	1.7993E+02
F14	Best	9.9800E-01	9.9800E-01	9.9800E-01
	Worst	9.9800E-01	1.9926E+00	1.0763E+01
	Mean	9.9800E-01	1.3294E+00	2.3131E+00
F15	Best	3.0749e-04	3.0749E-04	0.0011604
	Worst	3.0749e-04	2.0363E-02	4.1582E+03
	Mean	3.0749e-04	1.2877E-03	1.7993E+02
F16	Best	-1.0316E+00	-1.0316E+00	-1.0316E+00
	Worst	-1.0316E+00	-1.0316E+00	-1.0316E+00
	Mean	-1.0316E+00	-1.0316E+00	-1.0316E+00
F17	Best	3.9789E-01	3.9789E-01	3.9789E-01
	Worst	3.9791E-01	3.9789E-01	3.9789E-01
	Mean	3.9789E-01	3.9789E-01	3.9789E-01
F18	Best	3.0000E+00	3.0000E+00	3.0000E+00
	Worst	3.0000E+00	3.0000E+00	3.0000E+00
	Mean	3.0000E+00	3.0000E+00	3.0000E+00
F19	Best	-3.8628E+00	-3.8628E+00	-3.8628E+00
	Worst	-3.8628E+00	-3.8549E+00	-3.8628E+00
	Mean	-3.8628E+00	-3.8620E+00	-3.8628E+00
F20	Best	-3.3220E+00	-3.3220E+00	-3.3220E+00
	Worst	-3.2031E+00	-2.9564E+00	-3.1376E+00
	Mean	-3.2150E+00	-3.2389E+00	-3.2324E+00
F21	Best	-1.0153E+01	-1.0153E+01	-1.0153E+01
	Worst	-5.1008E+00	-2.6305E+00	-2.6305E+00
	Mean	-8.6375E+00	-6.7321E+00	-5.8820E+00
F22	Best	-1.0403E+01	-1.0403E-01	-1.0403E-01
	Worst	-1.0403E+01	-2.7519E+00	-2.7519E+00
	Mean	-1.0403E+01	-6.7370E+00	-7.4010E+00
F23	Best	-1.0536E+01	-1.0536E+01	-1.0536E+01
	Worst	-1.0536E+01	-2.4217E+00	-2.4217E+00
	Mean	-1.0536E+01	-7.2984E+00	-8.0934E+00

algorithm, which outperformed MFO_CLUST 7674.9282 and PSO_CLUST (7348.2371). The mean fitness value for MFO_PSO_CLUST (7200.6908) indicated high consistency, whereas greater variation was observed for the other

algorithms. For the Wine dataset, the lowest best fitness value (16292.1846) was achieved by MFO_PSO_CLUST, outperforming both MFO_CLUST (8379
4.3558) and PSO_CLUST (16409.438). Robustness in MFO_PSO_CLUST was evidenced by its nearly identical mean and worst fitness values. In the Vowel dataset, MFO_PSO_CLUST achieved superior performance (148967.2408) compared to MFO_CLUST (1036524.9394) and PSO_CLUST (177395.8541). Its lower mean fitness value (150320.9081) further highlighted its efficiency. For the Glass dataset, the best fitness value (210.4287) was achieved by MFO_PSO_CLUST, surpassing MFO_CLUST (939.6499) and PSO_CLUST (388.7199). The lower variability in its fitness values indicated superior stability compared to the other algorithms. In the Zoo dataset, MFO_PSO_CLUST demonstrated the best performance (best fitness value: 101.1195) in comparison to MFO_CLUST (222.2408) and PSO_CLUST (181.4975). Its consistent mean fitness value 107.1153 further reinforced its reliability. Across all datasets, MFO_PSO_CLUST consistently achieved better fitness values with minimal variation between the best, mean, and worst outcomes. This high consistency in its optimization process demonstrated the robustness and reliability of MFO_PSO_CLUST. In contrast, MFO_CLUST exhibited significant variability, suggesting susceptibility to suboptimal solutions in some iterations. Although PSO_CLUST displayed moderate performance, it failed to match the accuracy and stability of MFO_PSO_CLUST consistently. Altogether, MFO_PSO_CLUST demonstrated superior performance across all datasets by achieving better fitness values with a high degree of consistency and robustness. These results underscore its effectiveness in addressing diverse clustering challenges when compared to MFO_CLUST and PSO_CLUST. Table 6 presents additional detailed results.

Figure 3 illustrates the convergence behavior of the MFO_PSO_CLUST, MFO_CLUST, and PSO_CLUST algorithms across six datasets: Iris, Banknote_authentication, Wine, Zoo, Vowel, and Glass. These figures display the algorithm's progress towards the optimal solution across a defined number of iterations. For the Iris dataset, the MFO_PSO_CLUST algorithm converges rapidly to a stable solution with a significantly lower score

Table 6: Evaluation of the fitness values for different datasets.

Algorithm	Dataset	MFO_PSO_CLUST	MFO_CLUST	PSO_CLUST	No. cluster	No. iteration
best fit. val	Iris	96.6556	195.5371	119.7239	3	10
mean fit. val		96.6555	205.9479	126.0913		
worst fit. val		96.6555	212.1155	132.3762		
best fit. val	Banknote_	7200.6576	7674.9282	7348.2371	2	10
mean fit. val	authentication	7200.6908	7896.7872	7488.0422		
worst fit. val		7200.9767	8186.5658	7610.7892		
best fit. val	Wine	16292.1846	83794.3558	16409.438	3	10
mean fit. val		16292.2329	83794.6123	16469.1824		
worst fit. val		16292.6672	83794.9193	16520.0933		
best fit. val	Vowel	148967.2408	1036524.9394	177395.8541	6	10
mean fit. val		150320.9081	1036524.9394	185652.7939		
worst fit. val		160668.3641	1036524.9394	193752.0194		
best fit. val	Glass	210.4287	939.6499	388.7199	6	10
mean fit. val		226.0724	948.786	405.1631		
worst fit. val		246.6828	957.2347	423.1936		
best fit. val	Zoo	101.1195	202.2408	181.4975	7	10
mean fit. val		107.1153	212.7521	185.0522		
worst fit. val		112.0928	222.7578	189.0228		

compared to MFO_CLUST and PSO_CLUST, demonstrating superior efficiency and faster convergence. In contrast, MFO_CLUST exhibits a slower and less stable convergence process, while PSO_CLUST reaches a higher score, indicating suboptimal performance. In the Banknote_authentication dataset, MFO_PSO_CLUST achieves the best score earlier in the iterations, outperforming both MFO_CLUST and PSO_CLUST. The curves for MFO_CLUST and PSO_CLUST display higher variability, reflecting less consistent optimization performance. For the Wine dataset, the MFO_PSO_-CLUST algorithm demonstrates robust convergence, achieving a stable and optimal solution earlier than the other methods. MFO_CLUST and PSO_CLUST converge at higher scores, with MFO_CLUST showing slower convergence and higher variability throughout the iterations. The Zoo dataset showcases the significant advantage of MFO_PSO_CLUST, which converges rapidly to a stable and optimal solution. Both MFO_CLUST and PSO_CLUST exhibit prolonged and less effective convergence, with higher final scores, indicating poorer optimization. In the Vowel dataset, the performance trend remains consistent, with MFO_PSO_CLUST converging to the lowest score, followed by MFO_CLUST and PSO_CLUST. The gap between the algorithms final scores further emphasizes the superior performance of MFO_PSO_CLUST in minimizing the fitness function. Finally, for the Glass dataset, MFO_PSO_CLUST achieves the best score with minimal variability, outperforming the other algorithms. Both MFO_CLUST and PSO_CLUST converge at significantly higher scores, showcasing their limitations in handling this dataset.

The convergence curves across all datasets indicate the consistent superiority of MFO_PSO_CLUST in terms of faster convergence, lower fitness values, and greater stability, making it a highly robust and efficient algorithm for clustering tasks.

The efficiency evaluation of the proposed approach, utilizing the Iris, Wine, and Banknote_authentication datasets, is shown in Table 7. It is assessed in terms of the Normalized mutual information (NMI), Fowlkes-Mallows index (FMI), and $Silhouette\_score$ criteria [16, 35]. The table mentioned shows optimal efficiency ranges. For all criteria, A value that is larger and closer to 1 indicates higher efficiency.

Table 7: Assessing MFO_PSO_CLUST and k-means using evaluation criteria

	Dataset	MFO_PSO_CLUST	k-means	number of cluster	Note
NMI	Wine	0.42876	0.42411	3	The ranges [0,1].
	Iris	0.75821	0.74193	3	A higher index indicates
	Banknote_authentication	**0.030324**	**0.030324**	2	greater similarity.
	vowel	0.4948	0.48029	6	
	Glass	0.42693	0.4126	6	
	Zoo	0.7616	0.73226	7	
FMI	Wine	0.58387	0.5314	3	The ranges [0,1].
	Iris	0.80518	0.79821	3	A higher index indicates
	Banknote_authentication	**0.52769**	**0.52769**	2	greater similarity.
	vowel	0.47726	0.4635	6	
	Glass	**0.40502**	**0.40813**	6	
	Zoo	0.81281	0.74278	7	
Silhouette_score	Wine	0.7323	0.71578	3	The ranges [0,1].
	Iris	0.73546	0.73421	3	A higher index indicates
	Banknote_authentication	**0.63519**	**0.63519**	2	greater similarity.
	vowel	0.53741	0.53001	6	
	Glass	0.65475	0.60565	6	
	Zoo	0.6264	0.53085	7	

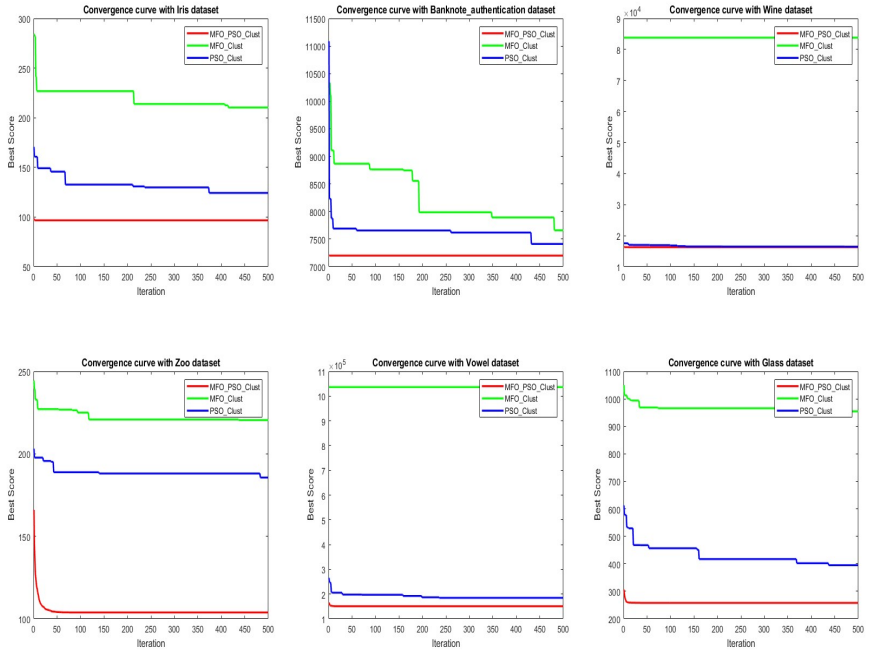


Figure 3: Convergence curve MFO_PSO_CLUST, MFO_CLUST and PSO_CLUST with different datasets

The NMI is a technique used for the analysis of clustering's performance against the true labeling. The results indicate that MFO_PSO_CLUST performs better than k-means in most datasets, such as Iris (0.75821 vs. 0.74193) and Zoo (0.7616 vs. 0.73226). However, for the Banknote_authentication dataset, both algorithms performed poorly, achieving an identical NMI score of 0.030324, which highlights a potential challenge in clustering this specific dataset. FMI metric assesses the balance between precision and recall in clustering. The MFO_PSO_CLUST algorithm demonstrated superior performance across most datasets, except for Glass, where k-means performed slightly better (0.40813 vs. 0.40502). Notably, for the Iris dataset, MFO_PSO_CLUST achieved a significantly higher FMI score (0.80518 vs. 0.79821), indicating a more balanced and accurate clustering. *Silhouette _score* measure evaluates the cohesion within clusters and the separation between clusters. MFO_PSO_CLUST consistently outperformed k-means in most datasets, including Wine (0.7323 vs. 0.71578) and Zoo (0.6264 vs.

0.53085). Based on these results, the new algorithm can generate tighter and better-separated clusters of data points as opposed to k-means clustering. The results indicate that MFO_PSO_CLUST generally performs better than k-means in terms of clustering quality, particularly in the NMI and $Silhouette\_score$ metrics. This demonstrates the effectiveness of the proposed algorithm in producing more coherent and meaningful clusters across various datasets. However, there are specific cases, such as the Glass dataset with the FMI metric, where k-means shows a slight advantage. The uniqueness of this dataset may be an explanation for this results in the clustering performance.

6 Conclusion

In this article, using the metaheuristic algorithm and their combination, we achieved better results for clustering. The advantages of several algorithms can be utilized to get better results by merging them. MFO and PSO methods were combined in the MFO_PSO algorithm. In the first part, MFO_PSO was tested with 23 benchmark functions and proved to be a robust algorithm for solving various problems. Its excellent performance in various tests suggests that it consistently finds superior solutions. Then, the MFO_PSO algorithm was applied to cluster data and determine the centroid with the lowest fitness function as the optimal center. Once the best centroid is identified by utilizing the MFO_PSO_CLUST algorithm, the data are clustered using it. Based on three datasets and three NMI, FMI, and $Silhouette\_score$ criteria, performance analysis demonstrated that the MFO_PSO_CLUST algorithm outperforms the MFO_CLUST and PSO_CLUST methods. The outcomes of this method can be further improved by adjusting the parameters.

Data Availability Statement

The datasets used in this paper are as follows:

- The Banknote_authentication dataset from https://archive.ics.uci.edu/ml/datasets/banknote+authentication.

- The Iris dataset from https://archive.ics.uci.edu/dataset/53/iris.

- The Vowel dataset from https://archive.ics.uci.edu/ml/datasets/Vowel.

- The Zoo dataset from https://archive.ics.uci.edu/dataset/111/zoo.

- The Glass dataset from https://archive.ics.uci.edu/dataset/42/glass+identification.

- The Wine dataset from https://archive.ics.uci.edu/dataset/109/wine.

Acknowledgements

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

References

[1] Abdollahzadeh, B., Gharehchopogh, F.S., Khodadadi, N. and Mirjalili, S. *Mountain gazelle optimizer: a new nature-inspired metaheuristic algorithm for global optimization problems*, Advances in Engineering Software, 174 (2022), 103282.

[2] Agushaka, J.O., Ezugwu, A.E. and Abualigah, L. *Dwarf mongoose optimization algorithm*, Comput. Methods Appl. Mech. Engrg, 391 (2022), 114570.

[3] Ahmad, A. and Dey, L. *A k-mean clustering algorithm for mixed numeric and categorical data*, Data & Knowledge Engineering, 63 (2007), 503–27.

[4] Alhawarat, M. and Hegazi, M. *Revisiting K-Means and topic modeling, a comparison study to cluster arabic documents*, IEEE Access, 6 (2018), 42740–42749.

[5] Aljarah, I., Mafarja, M., Heidari, A.A., Faris, H. and Mirjalili, S. *Clustering analysis using a novel locality-informed grey wolf-inspired clustering approach*, Knowledge and Information Systems, 62 (2020), 507–539.

[6] Almasri, A., Alkhawaldeh, R.S. and Çelebi, E. *Clustering-Based EMT Model for Predicting Student Performance*, Arab. J. Sci. Eng, 45 (2020), 10067–10078.

[7] Alswaitti, M., Albughdadi, M. and Isa, N.A.M. *Variance-based differential evolution algorithm with an optional crossover for data clustering*, Appl. Soft Comput, 80 (2019), 1–17.

[8] Barbakh, W.A. Wu, Y. and Fyfe, C. *Review of clustering algorithms, In Non-Standard Parameter Adaptation for Exploratory Data Analysis*, Springer: Berlin/Heidelberg, Germany, (2009), 7–28.

[9] Çomak, E. *A modified particle swarm optimization algorithm using Renyi entropy-based clustering*, Neural Computing and Applications, 27 (5) (2016), 1381–1390.

[10] Dorigo, M., Bonabeau, E. and Theraulaz, G. *Ant algorithms and stigmergy*, Future Generation Computer System, 16 (2000), 851–871.

[11] Eberhart, R. and Kennedy, J. *Particle swarm optimization*, InProceedings of the IEEE international conference on neural networks, 4 (1995), 1942-1948.

[12] Eesa, A.S. and Orman, Z. *A new clustering method based on the bio-inspired cuttlefish optimization algorithm*, Expert Syst, 37 (2020), e12478.

[13] Esmin, A.A., Coelho, R.A. and Matwin, S. *A review on particle swarm optimization algorithm and its variants to clustering high-dimensional data*, Artif. Intell. Rev, 44 (2015), 23–45.

[14] Gan, G., Ma, C. and Wu, J. *Data clustering: theory, algorithms, and applications*, SIAM, Society for Industrial and Applied Mathematics, (2007).

[15] Gandomi, A.H., Yang, X.S. and Alavi, A.H. *Cuckoo search algorithm, a metaheuristic approach to solve structural optimization problems*, Eng. Comput, 29 (2013), 17–35.

[16] Hu, F., Liu, J., Li, L. and Liang, J. *Community detection in complex networks using Node2vec with spectral clustering*, Physica A: Statistical Mechanics and its Applications, 545 (2020), 123633.

[17] Jadhav, A.N. and Gomathi, N. *Kernel-based exponential grey wolf optimizer for rapid centroid estimation in data clustering*, Jurnal Teknologi, 78 (11) (2016), 65–74.

[18] Jadhav, A.N. and Gomathi, N. *WGC: Hybridization of exponential grey wolf optimizer with whale optimization for data clustering*, Alex. Eng. J, 57 (2018), 1569–1584.

[19] Jain, A.K. *Data clustering: 50 years beyond K-means*, Pattern Recognit. Lett, 31 (2010), 651–666.

[20] Jain, A.K. and Dubes, R.C. *Algorithms for Clustering Data*, Englewood Cliffs, NJ, USA: Prentice-Hall, (1988).

[21] Kumar, N. and Kumar, H. *A fuzzy clustering technique for enhancing the convergence performance by using improved Fuzzy c-means and Particle Swarm Optimization algorithms*, Data & Knowledge Engineering, 140 (2022), 102050.

[22] Kumar, Y. and Sahoo, G. *Hybridization of magnetic charge system search and particle swarm optimization for efficient data clustering using neighborhood search strategy*, Soft Computing, 19 (12) (2015), 3621–3645.

[23] Kushwaha, N., Pant, M., Kant, S. and Jain, V.K. *Magnetic optimization algorithm for data clustering*, Pattern Recognition Letters, 115 (2018), 59–65.

[24] Liang, J., Suganthan, P. and Deb, K. *Novel composition test functions for numerical global optimization, in Swarm Intelligence Symposium*, 2005. SIS 2005, Proceedings 2005 IEEE, (2005), 68–75.

[25] Luque-Chang, A., Cuevas, E., Fausto, F., Zaldívar, D. and Pérez, M. *Social spider optimization algorithm: modifications, applications*, and perspectives, Mathematical Problems in Engineering, 1 (2018), 6843923.

[26] Lv, Z., Liu, T., Shi, C., Benediktsson, J.A. and Du, H. *Novel land cover change detection method based on k-Means clustering and adaptive majority voting using bitemporal remote sensing images*, IEEE Access, 7 (2019), 34425–34437.

[27] MacQueen, J. *Some methods for classification and analysis of multivariate observations*, in Proc. 5th Berkeley Symp. Math. Statist. Probab, 1 (1967), 281–297.

[28] Mansalis, S., Ntoutsi, E., Pelekis, N. and Theodoridis, Y. *An evaluation of data stream clustering algorithms*, Statistical Analysis and Data Mining: The ASA Data Science Journal, 11 (2018), 167–187.

[29] Meng, Y., Liang, J., Cao, F. and He, Y. *A new distance with derivative information for functional k-means clustering algorithm*, Inf. Sci. 463 (2018), 166–185.

[30] Mirjalili, S. *Moth-flame optimization algorithm: A novel nature-inspired heuristic paradigm*, Knowledge-based systems. 89 (2015), 228–249.

[31] Mirjalili, S. and Lewis, A. *The whale optimization algorithm*, Adv. Eng. Softw. 95 (2016), 51–67.

[32] Mirjalili, S., Mirjalili, S.M. and Lewis, A. *Grey wolf optimizer*, Advances in engineering software. 69 (2014), 46–61.

[33] Nanda, S.J. and Panda, G. *A survey on nature inspired metaheuristic algorithms for partitional clustering*, Swarm Evol. Comput. 16 (2014), 1–18.

[34] Nasiri, J. and Khiyabani, F.M. *A whale optimization algorithm (WOA) approach for clustering*, Cogent Math. Stat. 5 (2018), 1483565.

[35] Rodriguez, M.Z., Comin, C.H.v Casanova, D., Bruno, O.M., Amancio, D.R., Costa, L.D.F. and Rodrigues, F.A. *Clustering algorithms: A comparative approach*, PLoS ONE 14(1), e0210236 (2019). https://doi.org/10.1371/journal.pone.0210236

[36] Saida, I.B., Nadjet, K. and Omar, B. *A new algorithm for data clustering based on cuckoo search optimization, In Genetic and Evolutionary Computing*, Springer: Berlin/Heidelberg, Germany. (2014), 55–64.

[37] Sayed, G.I. and Hassanien, A.E. *A hybrid SA-MFO algorithm for function optimization and engineering design problems*, Complex & Intelligent Systems. 4 (2018), 195–212.

[38] Shojaee, Z., Shahzadeh Fazeli, S.A., Abbasi, E. and Adibnia, F. *Feature Selection based on Particle Swarm Optimization and Mutual Information*, AI and Data Mining. 9(1) (2021), 39–44.

[39] Singh, T. *A novel data clustering approach based on whale optimization algorithm*, Expert Syst. 38 (2020), e12657.

[40] Singh, T. and Mishra, K.K. *Data clustering using environmental adaptation method In International Conference on Hybrid Intelligent Systems*, Springer. (2019), 156–164.

[41] Singh, T., Mishra, K.K. and Ranvijay. *A variant of EAM to uncover community structure in complex networks*, Int. J. Bio-Inspired Comput. 16 (2020), 102–110.

[42] Singh, T., Saxena, N., Khurana, M., Singh, D., Abdalla, M. and Alshazly, H. *Data clustering using moth-flame optimization algorithm*, Sensors. 21(12) (2021), 4086.

[43] Suganthan, P.N., Hansen, N., Liang, J. J., Deb, K., Chen, Y.P., Auger, A. and Tiwari, S. *Problem definitions and evaluation criteria for the CEC 2005 special session on real-parameter optimization*, KanGAL report. 2005005 (2005).

[44] Talevi, A. and Bellera, CL. *Clustering of small molecules: new perspectives and their impact on natural product lead discovery*, Frontiers in Natural Products. 3 (2024), 1367537.

[45] Wolpert, D.H. and Macready, W.G. *No free lunch theorems for optimization*, IEEE Trans Evol Comput. 1 (1997), 67–82.

[46] Zhang, Q.H., Li, B.L., Liu, Y.J., Gao, L., Liu, L.J. and Shi, X.L. *Data clustering using multivariant optimization algorithm*, Int. J. Mach. Learn. Cybern. 7 (5) (2016), 773–782.

[47] Zhu, J., Jiang, Z., Evangelidis, G.D., Zhang, C., Pang, S. and Li, Z. *Efficient registration of multi-view point sets by K-means clustering*, Inf. Sci. 488 (2019), 205–218.

Iranian Journal of Numerical Analysis and Optimization
Vol. 15, No. 2, 2025, pp 508–530
https://doi.org/10.22067/ijnao.2025.89779.1511
https://ijnao.um.ac.ir/

Research Article

Exploring hyperchaotic synchronization of a fractional-order system without equilibrium points: A sliding mode control approach

R.A. Meskine and S. Kaouache*

Abstract

Recently, constructing hidden attractors of chaotic systems without equilibrium point has become a key discussion point in the application fields of chaos and hyperchaos science. This paper introduces a novel hyperchaotic system without equilibrium points, distinct from existing systems that rely on the Shilnikov criterion for demonstrating hyperchaos. This study investigates the qualitative properties of the system, including its hyperchaotic

*Corresponding author

Received 13 September 2024; revised 11 December 2024; accepted 17 January 2025

Rayene Abir Meskine
Laboratory of mathematics and their interactions, Department of Mathematics, Institute of Mathematics and Computer Science, Abdelhafid Boussouf University Center, Mila. Algeria e-mail: r.meskine@centre-univ-mila.dz

Smail Kaouache
Laboratory of mathematics and their interactions, Department of Mathematics, Institute of Mathematics and Computer Science, Abdelhafid Boussouf University Center, Mila, Algeria. e-mail: smailkaouache@gmail.com

How to cite this article
Meskine, R.A. and Kaouache, S., Exploring hyperchaotic synchronization of a fractional-order system without equilibrium points: A sliding mode control approach. *Iran. J. Numer. Anal. Optim.*, 2025; 15(2): 508-530. https://doi.org/10.22067/ijnao.2025.89779.1511

attractors, Poincare map, Lyapunov exponents, and Kaplan-Yorke dimension. To enhance the practical applicability of this system, an integral sliding mode control method for synchronization is proposed. Lyapunov theory ensures the stability and effectiveness of the synchronization scheme. The efficiency of the approach is demonstrated by numerical simulations, which validate the potential of the system for various applications.

AMS subject classifications (2020): 34C28, 93C10, 93D05, 93D20

Keywords: Fractional derivative; Synchronization; Lyapunov stability theory; Lyapunov exponents; Sliding mode control.

1 Introduction

Fractional calculus, a mathematical discipline with over 300 years of history, extends traditional calculus to noninteger order derivatives and integrals. The study of fractional calculus has attracted significant attention due to its potential applications in various fields. Fractional-order systems provide an explicit description and further insight into physical processes. They also serve as valuable tools in modeling many phenomena. The use of fractional derivatives has been investigated in various research fields, including electromagnetism [41], secure communication [19], cryptography [34], viscolelasticty [15], diffusion [18], and chemical processing [16], due to their elegant description of interdisciplinary applications.

Chaos theory is the study of systems that obey deterministic rules but exhibit seemingly random behavior. Small differences in initial conditions can lead to very different outcomes, creating a fascinating interplay between order and randomness. The study of chaos allows us to understand and potentially predict complex behavior in a variety of scientific disciplines. Hyperchaotic systems, a subset of chaotic systems, exhibit even more complex dynamics with multiple positive Lyapunov exponents [42]. Many chaotic systems, including the Chua circuit [17], Duffing system [32, 46], Chen system [30], Lu system [35], and Rossler system [31], can be elegantly modeled using fractional-order equations.

Synchronization refers to a fascinating phenomenon in which two systems adjust their dynamics to become identical, essentially when one of two systems adapts its trajectory to follow that of the other [27, 45]. In 1990, Pecora and Carroll [38] established a pioneering scheme for achieving synchronization, known as the drive-response method. Nowadays, various synchronization techniques, such as complete synchronization [5, 8], phase synchronization [40], projective synchronization [7, 19, 47], inverse matrix projective synchronization [23], modified projective synchronization [21], modified hybrid synchronization [25], generalized synchronization [22, 37, 53], $Q - S$ synchronization [9], fixed-time synchronization [58], and combination synchronization [3, 2, 20, 24, 26] have been described. Recently, due to the wide range of applications, many different control schemes have been applied to synchronize fractional-order chaotic systems, such as active control [4, 14], adaptive control [56], scalar transmitted signal methods [36], impulsive control [52], event-triggered controller [43, 55], and sliding mode control (SMC) [1, 29, 54], among others.

The SMC is a powerful control strategy renowned for its robustness to parameter variations [6]. Additionally, SMC offers several advantages, including excellent transient performance, rapid dynamic responses, and disturbance rejection capability.

The analysis of chaotic systems has traditionally been conducted through the lens of equilibrium points, due to the sensitivity of these systems to initial conditions and their unpredictable long-term behavior. However, recent research has revealed a fascinating class of chaotic systems that challenge this convention: nonequilibrium chaotic systems [13, 44, 51, 57]. Despite the absence of stable equilibrium points, nonequilibrium chaotic systems are capable of generating intricate, confined trajectories, designated as hidden attractors. This counterintuitive behavior offers promising opportunities for potential applications in a range of fields. As research progresses, the challenge of synchronizing these systems remains a captivating challenge [39, 48, 49, 50].

This work makes a contribution to the field of hyperchaotic systems with hidden attractors by introducing a novel fractional-order system that lacks equilibrium points. A comprehensive study of the system's dynamical behavior is conducted with the objective of achieving synchronization using integral

sliding mode control (ISMC). The synchronization controller and parameter identification technique are designed based on Lyapunov stability theory, and the theoretical findings are validated through computer simulations. The primary contributions of this work include:

1. A groundbreaking four-dimensional fractional-order hyperchaotic system is introduced from the $3-D$ Rikitake dynamo system.

2. Rigorous analysis confirms the hyperchaotic nature of the proposed system, highlighting its complex and unpredictable behavior.

3. Developing of an ISMC strategy for synchronizing two identical instances of the novel system.

4. Lyapunov theory and numerical simulations to validate the effectiveness of the synchronization approach.

The rest of the manuscript is arranged as follows: Section 2 presents some preliminary concepts for fractional calculus. Section 3 introduces the construction of four dimensional fractional-order system, it delves into a detailed analysis of its fundamental dynamical property. Section 4 develops an integral sliding mode controller for synchronizing two identical hyperchaotic systems. The feasibility and effectiveness of both control schemes are verified by numerical simulations.

2 Basic definitions for fractional-order systems

In this section, some basic definitions related to fractional differentiation are presented.

Definition 1. In the field of fractional calculus, a vital function is the Euler gamma function Γ, which is defined as follows

$$\Gamma(a) = \int_0^{+\infty} t^{a-1} e^{-t} dt. \tag{1}$$

Definition 2. Fractional calculus extends the concepts of integration and differentiation to non-integer orders. It uses a fundamental operator, denoted

${}_aD_t^\alpha$, which defines fractional integration and differentiation. The integro-differential operator ${}_aD_t^\alpha$ is defined by

$$ {}_aD_t^\alpha = \begin{cases} \dfrac{d^\alpha}{(dt)^\alpha}, & \alpha > 0, \\ 1, & \alpha = 0, \\ \int_a^t (d\tau)^\alpha, & \alpha < 0, \end{cases} \tag{2} $$

where a and t are the bounds of the operator and $t \in \mathbb{R}$.

Several definitions exist regarding fractional derivative, but the Caputo definition in (3) is used the most in engineering applications, since this definition incorporates initial values for a function f and its standard derivatives, that is, initial values that are physically appealing in the traditional way. Also, a modified version of Adams–Bashforth–Moulton method [12] will be employed for numerical simulation of the Caputo fractional-order equations.

Definition 3 (Caputo fractional derivative [28])**.** The Caputo fractional derivative of order α is defined as follows:

$$ D^\alpha f(t) = I^{n-\alpha} f^n(t), \tag{3} $$

where, $n-1 < \alpha \leq n$, $n \in \mathbb{N}$, I^α is the αth-order Riemann–Liouville fractional integration given by

$$ I^\alpha f(t) = \frac{1}{\Gamma(\alpha)} \int_a^t (t-\xi)^{\alpha-1} f(\xi) d\xi. \tag{4} $$

Theorem 1. Consider the following commensurate fractional-order dynamics system:

$$ D^\alpha x(t) = f(x(t)), $$

where D^α is a Caputo differential operator, $0 < \alpha < 1$, and $x \in \mathbb{R}^n$ is the state vector of system and f is continuous vector function of the system. The equilibrium points of this system are calculated by solving the equation $f(x) = 0$. These points are locally asymptotically stable if all eigenvalues λ_i of the Jacobian matrix $J = \partial f / \partial x$ satisfy

$$ |arg(\lambda_i)| > \alpha \frac{\pi}{2}. $$

3 Model description of the novel fractional-order hyperchaotic system

As reported in the literature, the $3-D$ Rikitake dynamo system [41] is described as follows

$$\begin{cases} \dot{x_1} = -ax_1 + x_2x_3, \\ \dot{x_2} = -ax_2 + x_1(x_3 - b), \\ \dot{x_3} = 1 - x_1x_2, \end{cases} \tag{5}$$

where x_1, x_2, and x_3 are the state variables, a and b are positive parameters and represent the resistive dissipation and the difference in the angular velocities of the two disks, respectively.

An extensive study established chaos in this system has been described by Cook and Roberts [10].

By adding a state feedback control to the previous system and replacing standard derivatives by fractional-order derivatives, we get the following system

$$\begin{cases} D^\alpha x_1 = -ax_1 + x_2x_3 - cx_4, \\ D^\alpha x_2 = -ax_2 + x_1(x_3 - b), \\ D^\alpha x_3 = d - x_1x_2, \\ D^\alpha x_4 = x_1, \end{cases} \tag{6}$$

where x_1, x_2, x_3, and x_4 are the state variables of the system, a, b, c, and $d \neq 0$ are real constants, $\alpha \in (0, 1]$ is the fractional-order derivative, and D^α is the Caputo differential operator.

In the following, some basic properties of the novel system (6) are numerically analyzed, including: Dissipativity, equilibrium points, Poincaré map, hyperchaotic attractors, Lyapunov exponents, and Kaplan-Yorke dimension.

3.1 Dissipativity

We use the divergence of the vector field. The system is dissipative if the sum of the coefficients of the fractional derivatives is negative. The divergence of the system (6) is calculated as

$$\nabla f = \frac{\partial D^{\alpha} x_1}{\partial x_1} + \frac{\partial D^{\alpha} x_2}{\partial x_2} + \frac{\partial D^{\alpha} x_3}{\partial x_3} + \frac{\partial D^{\alpha} x_4}{\partial x_4}.$$

Substituting the partial derivatives

$$\nabla f = -2a,$$

since $a > 0$, the sum of the coefficients of the fractional derivatives is negative, that is, $\nabla f < 0$. Therefore, the system is dissipative. In this case, all the trajectories of the system tend to an attractor, when $t \to +\infty$.

3.2 Equilibrium points

In order to find the equilibrium points of the system (6), we solve the four following equations

$$\begin{cases} -ax_1 + x_2 x_3 - cx_4 = 0, \\ -ax_2 + x_1(x_3 - b) = 0, \\ d - x_1 x_2 = 0, \\ x_1 = 0. \end{cases} \tag{7}$$

Since $d \neq 0$, the last two equations in the system (7) contradict each other. This means that the proposed system (6) has no equilibrium points. Thus, the system lacks the characteristic bifurcation phenomena (e.g., pitchfork, Hopf) typically observed in hyperchaotic systems. Additionally, the absence of equilibrium points precludes the existence of a stable sink state.

3.3 Poincaré map

To elucidate the complexity and characteristic features of the novel system, we employ the Poincaré map. This technique provides valuable insights into the underlying behavior.

Figures 1 and 2 show the Poincaré map of the system (6) in $2 - D$ plan and $3 - D$ space, respectively.

The system's Poincaré map (Figure 2) reveals intricate structures, suggesting hyperchaotic behavior.

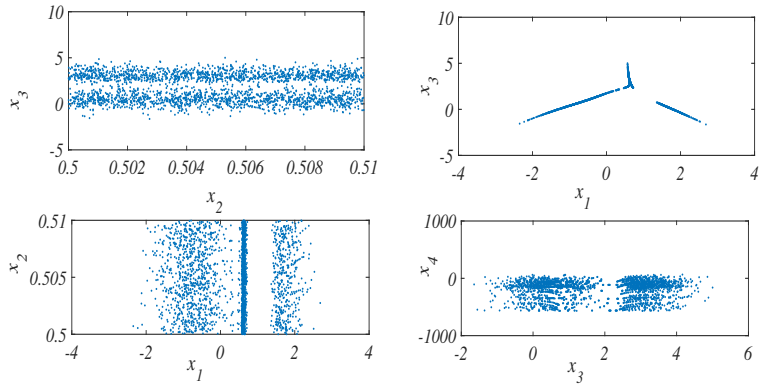


Figure 1: Poincaré map in $2-D$ plan.

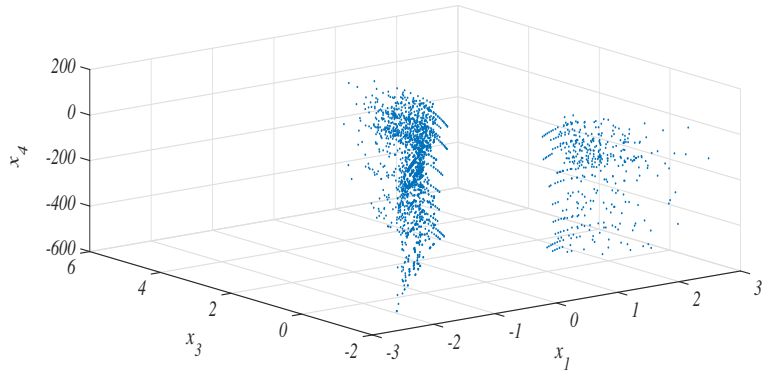


Figure 2: Poincaré map in $3-D$ space.

3.4 Hyperchaotic attractors

It is interesting that our system can generate hyperchaotic signals although there is no equilibrium points. The system (6) exhibits hyperchaotic behavior, as shown in Figure 3, where $\alpha = 0.98$, the system's parameters are as follows

$$a = b = d = 1 \quad and \quad c = 0,$$

and the initial values of the hyperchaotic system (6) are selected as

$$x_1(0) = x_2(0) = x_3(0) = -0.1 \quad and \quad x_4(0) = 0.2. \tag{8}$$

We used the MATLAB/Simulink environment to develop a comprehensive

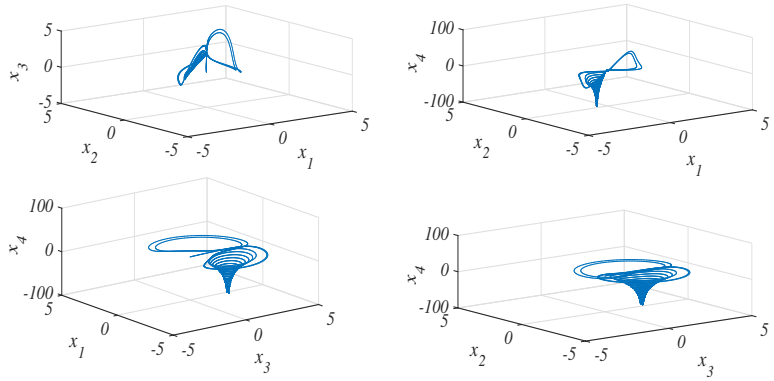


Figure 3: Hyperchaotic attractors of system (6), when $\alpha = 0.98$.

simulation framework for the system described by (6). To achieve accurate modeling of fractional-order dynamics, we utilized the capabilities of the FOMCON toolkit in Simulink. This toolbox provides robust blocks for implementing fractional integration of various orders. The detailed block diagram of the simulation setup is presented in Figure 4. The temporal evolution

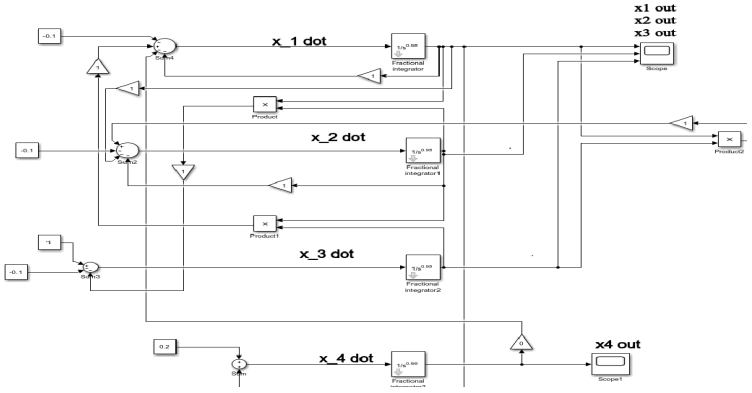


Figure 4: The block diagram of the new hyperchaotic system (6) in MATLAB Simulink.

of the state variables x_1, x_2, x_3, *and* x_4 are illustrated in Figures 5 and 6, respectively.

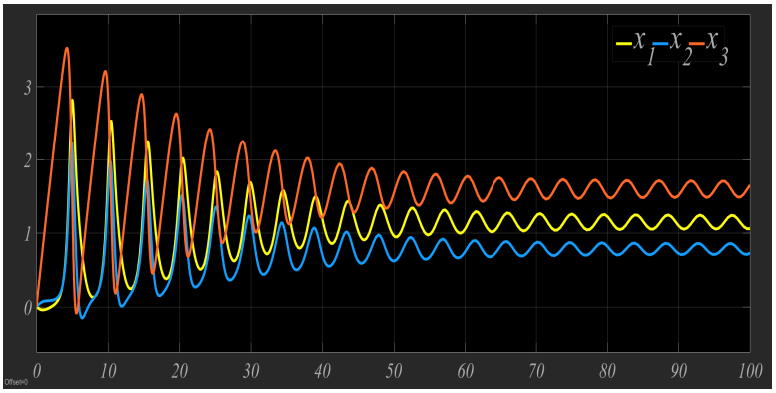


Figure 5: Temporal evolution of x_1, x_2, x_3.

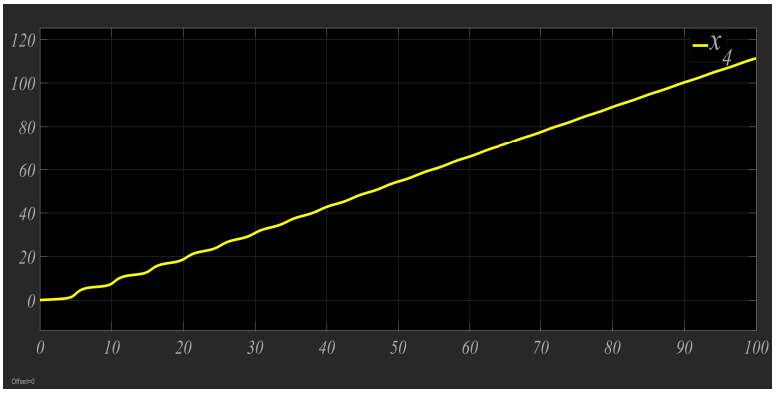


Figure 6: Temporal evolution of x_4.

In order to provide definitive confirmation of the hyperchaotic behavior exhibited by the proposed system, a computation of the Lyapunov exponents will be undertaken.

3.5 Lyapunov exponents

In this section, we assume that the parameters a, b, and d remain constant and c is varied in $[0, 5]$. By using MATLAB code for Lyapunov exponents of fractional-order systems [11], the Lyapunov exponents spectrum of system (6) with $a = 1, b = 1$, and $d = 1$ is represented in Figure 7.

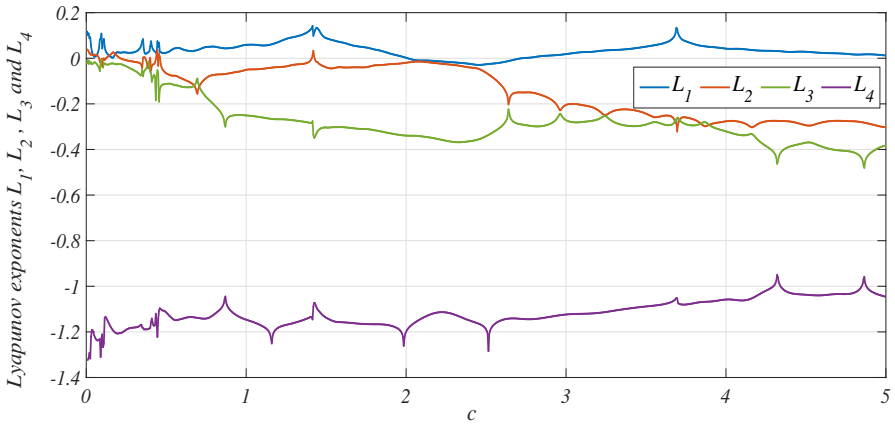


Figure 7: The dynamics of the Lyapunov exponents of the system (6).

Particularly, when the parameter systems are taken as

$$a = b = d = 1 \quad and \quad c = 0.03,$$

the Lyapunov exponents of the proposed system are obtained as

$$L_1 = 0.1208,\ L_2 = 0.0202,\ L_3 = 0,\ \text{and}\ L_4 = -1.2740.$$

The Kaplan–Yorke dimension of the system is calculated as

$$D_{KY} = 3 + \frac{L_1 + L_2 + L_3}{|L_4|} = 3.1107.$$

Thus, the system is hyperchaotic.

4 Sliding mode controller design and synchronization analysis

To assess the synchronization capabilities of the proposed hyperchaotic system, SMC is employed in order to achieve complete synchronization between the states of the master-slave systems.

4.1 General method of synchronization

Consider the fractional-order dynamic system defined by

$$D^{\alpha}x(t) = f(x(t)), \tag{9}$$

where D^{α} is Caputo differential operator, $0 < \alpha < 1$, $x \in \mathbb{R}^n$ is the state vector of system, and f is continuous vector function. We consider the system (9) as a drive system. The controlled response system is given by

$$D^{\alpha}y(t) = f(y(t)) + u. \tag{10}$$

Define the synchronization error as $e(t) = y(t) - x(t)$. The main purpose is to choose a suitable controller u, so that the response system accurately tracks the drive system's trajectory, such that

$$\lim_{t\to\infty} ||e(t)|| = 0. \tag{11}$$

4.2 Sliding mode controller design

The drive system is described by

$$\begin{cases} D^{\alpha}x_1 = -ax_1 + x_2x_3 - cx_4, \\ D^{\alpha}x_2 = -ax_2 + x_1(x_3 - b), \\ D^{\alpha}x_3 = d - x_1x_2, \\ D^{\alpha}x_4 = x_1. \end{cases} \tag{12}$$

The controlled response system is given by

$$\begin{cases} D^{\alpha}y_1 = -ay_1 + y_2y_3 - cy_4 + u_1, \\ D^{\alpha}y_2 = -ay_2 + y_1(y_3 - b) + u_2, \\ D^{\alpha}y_3 = d - y_1y_2 + u_3, \\ D^{\alpha}y_4 = y_1 + u_4, \end{cases} \tag{13}$$

where u_1, u_2, u_3, u_4 are the integral sliding mode controllers to be designed.

Define the synchronization errors as $e_i(t) = y_i(t) - x_i(t),\ i = 1, 2, 3, 4$. It is easy to see that the error dynamics system can be obtained as follows

$$\begin{cases} D^{\alpha}e_1 = -ae_1 + y_2y_3 - x_2x_3 - ce_4 + u_1, \\ D^{\alpha}e_2 = -ae_2 + y_1y_3 - x_1x_3 - be_1 + u_2, \\ D^{\alpha}e_3 = -y_1y_2 + x_1x_2 + u_3, \\ D^{\alpha}e_4 = e_1 + u_4. \end{cases} \tag{14}$$

In order to achieve the stability of system (14), four sliding surfaces s_1, s_2, s_3, and s_4 are introduced as

$$s_i = (D^{\alpha} + \lambda_i) \int_0^t e_i(\tau)d\tau, \qquad i = 1, 2, 3, 4. \tag{15}$$

A simple calculation yields

$$\begin{cases} \dot{s_1} = D^{\alpha}e_1(t) + \lambda_1 e_1(t), \\ \dot{s_2} = D^{\alpha}e_2(t) + \lambda_2 e_2(t), \\ \dot{s_3} = D^{\alpha}e_3(t) + \lambda_3 e_3(t), \\ \dot{s_4} = D^{\alpha}e_4(t) + \lambda_4 e_4(t), \end{cases} \tag{16}$$

where $\lambda_i > 0$ for $i = 1, 2, 3, 4$.
To design SMC, the reaching law [33] is chosen as

$$\begin{cases} \dot{s_1} = -\beta_1 sgn(s_1) - K_1 s_1, \\ \dot{s_2} = -\beta_2 sgn(s_2) - K_2 s_2, \\ \dot{s_3} = -\beta_3 sgn(s_3) - K_3 s_3, \\ \dot{s_4} = -\beta_4 sgn(s_4) - K_4 s_4, \end{cases} \tag{17}$$

where $\beta_i > 0$, $K_i > 0$, for $i = 1, 2, 3, 4$, and $sgn(\cdot)$ represents sign function, that is

$$sgn(s_i) = \begin{cases} 1, & s_i > 0, \\ 0, & s_i = 0, \\ -1, & s_i < 0. \end{cases}$$

By comparing (16) and (17), we deduce

$$\begin{cases} D^{\alpha}e_1(t) + \lambda_1 e_1(t) = -\beta_1 sgn(s_1) - K_1 s_1, \\ D^{\alpha}e_2(t) + \lambda_1 e_2(t) = -\beta_2 sgn(s_2) - K_2 s_2, \\ D^{\alpha}e_3(t) + \lambda_1 e_3(t) = -\beta_3 sgn(s_3) - K_3 s_3, \\ D^{\alpha}e_4(t) + \lambda_1 e_4(t) = -\beta_4 sgn(s_4) - K_4 s_4. \end{cases} \tag{18}$$

Combining (14) and (18), we get

$$\begin{cases} -ae_1 + y_2y_3 - x_2x_3 - ce_4 + u_1 + \lambda_1 e_1(t) = -\beta_1 sgn(s_1) - K_1 s_1, \\ -ae_2 + y_1y_3 - x_1x_3 - be_1 + u_2 + \lambda_2 e_2(t) = -\beta_2 sgn(s_2) - K_2 s_2, \\ -y_1y_2 + x_1x_2 + u_3 + \lambda_3 e_3(t) = -\beta_3 sgn(s_3) - K_3 s_3, \\ e_1 + u_4 + \lambda_4 e_4(t) = -\beta_4 sgn(s_4) - K_4 s_4. \end{cases} \tag{19}$$

In order to synchronize the drive system (12) with the response system (13), the control input can be obtained as

$$\begin{cases} u_1 = (a - \lambda_1)e_1 + ce_4 - y_2y_3 + x_2x_3 - \beta_1 sgn(s_1) - K_1 s_1, \\ u_2 = -b^* e_1 + (a - \lambda_2)e_2 - y_1y_3 + x_1x_3 - \beta_2 sgn(s_2) - K_2 s_2, \\ u_3 = -\lambda_3 e_3 + y_1y_2 - x_1x_2 - \beta_3 sgn(s_3) - K_3 s_3, \\ u_4 = -e_1 - \lambda_4 e_4 - \beta_4 sgn(s_4) - K_4 s_4. \end{cases} \tag{20}$$

4.3 Synchronization analysis

To enhance the practical applicability of the proposed system, an ISMC method is designed to achieve complete synchronization of the system. In addition, numerical simulations are performed to demonstrate and verify the theoretical analysis.

Theorem 2. The systems (12) and (13) are globally and asymptotically synchronized under the integral sliding mode controller (20), where α_i, β_i, K_i, $(i = 1, 2, 3, 4)$ are constant positives.

Proof. We establish this theorem using Lyapunov stability theory. First, we consider the quadratic and positive definite Lyapunov function defined by

$$V(s_1, s_2, s_3, s_4) = \frac{1}{2}(s_1^2 + s_2^2 + s_3^2 + s_4^2).$$

Differentiating V with respect to the time t, we obtain

$$\dot{V} = s_1\dot{s_1} + s_2\dot{s_2} + s_3\dot{s_3} + s_4\dot{s_4}. \tag{21}$$

Substituting from (17) to (21), we get

$$\dot{V} = \sum_{i=1}^{4}[-\beta|s_i| - K_i s_i^2)], \tag{22}$$

which is a negative definite function on $\mathbb{R}^4$.

Thus, by the Lyapunov stability theory, we conclude that $s_i \to 0$ as $t \to +\infty$, for $i = 1, 2, 3, 4$. Hence, it follows that $e_i \to 0$, as $t \to +\infty$, for $i = 1, 2, 3, 4$. □

4.4 Numerical simulations

For MATLAB simulations, we take the parameter values for the drive and response systems as in the case of hyperchaotic attractors

$$a = 1, b = 1, c = 0, \ and \ d = 1.$$

We take the sliding constants as

$$\lambda_i = 0.2, \ \beta_i = 0.2, \ and \ K_i = 3, \qquad for \ all \ i = 1, 2, 3, 4.$$

We set the initial conditions of the drive and the response systems, respec-

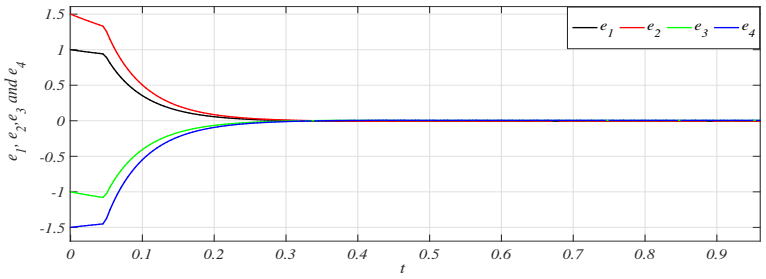


Figure 8: Time history of the synchronization errors.

tively, as

$$(x_1(0), x_2(0), x_3(0), x_4(0)) = (0.5, -0.5, 0.5, 1)$$

and

$$(y_1(0), y_2(0), y_3(0), y_4(0)) = (1.5, 1, -0.5, -0.5).$$

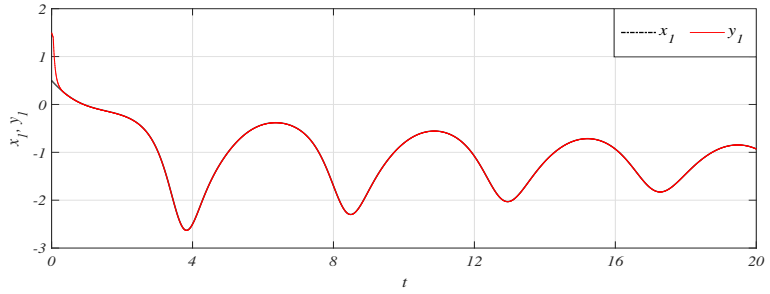


Figure 9: Complete synchronization between signals x_1 and y_1.

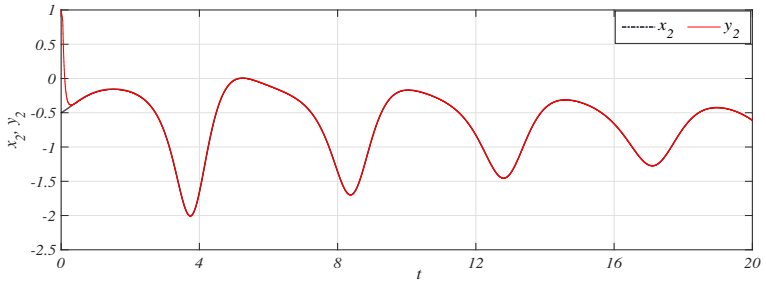


Figure 10: Complete synchronization between signals x_2 and y_2.

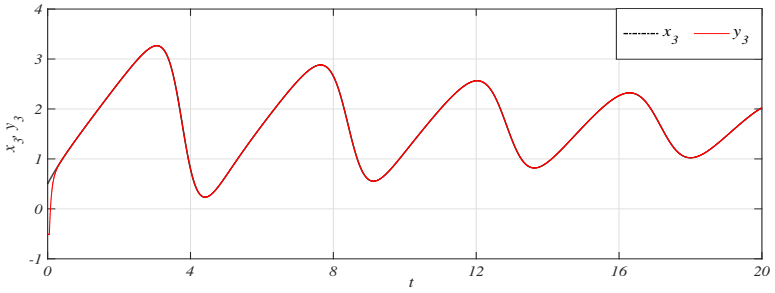


Figure 11: Complete synchronization between signals x_3 and y_3.

Figure 8 shows time history of the synchronization errors. Furthermore, Figures 9, 10, 11, and 12 show the complete synchronization between the states x_1 and y_1, x_2 and y_2, x_3 and y_3, x_4 and y_4, respectively .

From these figures, it is straightforward to see the complete synchronization between the drive and response systems (12) and (13) are achieved.

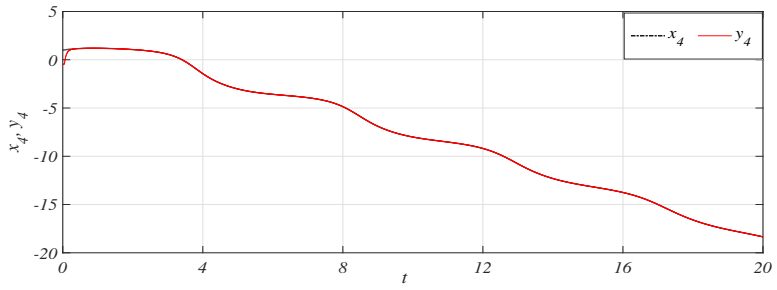


Figure 12: Complete synchronization between signals x_4 and y_4.

5 Conclusion

There is great interest in the modeling of hyperchaotic systems devoid of equilibrium points within the field of chaos literature. This research introduced a novel fractional-order system exhibiting complex dynamics, thereby expanding the understanding of chaotic behavior beyond traditional equilibrium-based frameworks. The comprehensive analysis conducted in this study, including the identification of hyperchaotic attractors and the computation of Lyapunov exponents, has yielded valuable insights into the intricate dynamics of the system.

Furthermore, the creation and application of an ISMC strategy for synchronization have illustrated the potential for controlling and coordinating these complex systems. The rigorous validation through MATLAB simulations has confirmed the effectiveness of the proposed approach, even in the face of the inherent challenges posed by the system's sensitivity to initial conditions.

The findings of this research have significant implications for a number of applications in fields such as secure communications, signal processing, and control systems. The capacity to generate and regulate complex, erratic dynamics can be harnessed for operations that necessitate randomness, noise, or chaotic behavior. Moreover, the demonstrated synchronization capabilities provide a foundation for the construction of complex networks and the coordination of multiple chaotic systems.

Acknowledgements

Authors are grateful to the anonymous referees and editor for their valuable comments and suggestions to improve the quality of the manuscript.

References

[1] Al-sawalha, M.M. *Synchronization of different fractional-order chaotic systems using modify adaptive sliding mode control*, Adv. Diff. Equ. 1 (2020), 417.

[2] Bensimessaoud, S. and Kaouache, S. *A general fractional control scheme for compound combination synchronization between different fractional-order identical chaotic systems*, Aust. J. Math. Anal. Appl. 20 (2) (2023), Art. 12, 9 pp.

[3] Bensimessaoud, S., Kaouache, S. and Abdelouahab, M.-S. *Chaos combination anti-synchronization (CCAS) of some fractional-order uncertain chaotic systems by some random noise*, Dyn. Contin. Discrete Impuls. Syst. Ser. A Math. Anal. 31 (2024), 199–211.

[4] Bhalekar, S. and Daftardar-Gejji, V. *Synchronization of different fractional-order chaotic systems using active control*, Commun. Nonlinear Sci. Numer. Simul. 15 (2010), 3536–3546.

[5] Bhat, M. A. and Shikha. *Complete synchronization of non-identical fractional order hyperchaotic systems using active control*, Int. J. Autom. Control, 13(2) (2019), 140–157.

[6] Boiko, I., Fridman, I., Iriarte, R., Pisano, A. and Usai, E. *Parameter tuning of second-order sliding mode controllers for linear plants with dynamic actuators*, Automatica, 42 (2006), 833–839.

[7] Bouzeriba, A., Boulkroune, A. and Bouden, T. *Projective synchronization of two different fractional-order chaotic systems via adaptive fuzzy control*, Neural. Comput. Appl. 27 (2016), 1349–1360.

[8] Bowonga S., Kakmenib, M. and Koinac, R. *Chaos synchronization and duration time of a class of uncertain systems*, Math. Comput. Simul. 71(3) (2006), 212–228.

[9] Chai, Y., Chen, L., Wu, R. and Dai, J. *$Q-S$ synchronization of the fractional-order unified system*, Pramana, 80 (2013), 449–461.

[10] Cook, A.E. and Roberts, P.H. *The Rikitake two disk dynamo system*, Math. Proc. Cambridge, 68(2), (1970), 547–569.

[11] Danca, M.F. and Kuznetsov, N. *MATLAB code for Lyapunov exponents of fractional-order systems*, Int. J. Bifurcat. Chaos, 28 (5) (2018) 1850067.

[12] Diethelm K. and Ford, N.J. *Analysis of fractional differential equations*, J. Math Anal. Appl. 265(2) (2002), 229–248.

[13] Dong, C. and Wang, J. *Hidden and coexisting attractors in a novel $4-D$ hyperchaotic system with no equilibrium point*, Fractal Fract. 6(6) (2022), 306.

[14] Du, C., Liu, L. and Shi, S. *Synchronization of fractional-order complex chaotic system using active control method*, IEEE UEMCON. (2019), 0817–0823.

[15] El-Shahed, M. *MHD of a fractional viscoelastic fluid in a circular tube*, Mech. Res. Commun. 33(2) (2006), 261–268.

[16] Flores-Tlacuahuac, F. and Biegler, L.T. *Optimization of fractional-order dynamic chemical processing systems*, Ind. Eng. Chem. Res. 53(13) (2014), 5110–5127.

[17] Hartley, T.T., Lorenzo, C.F. and Qammar, H.K. *Chaos in a fractional order Chua system*, IEEE Trans. Circuits Syst. I, 42(8) (1996), 485–490.

[18] Jesus, I.S. and Machado, J.T. *Fractional control of heat diffusion systems*, Nonlinear Dyn. 54(3) (2008), 263–282.

[19] Kaouache, S. *Projective synchronization of the modified fractional-order hyperchaotic Rossler system and its application in secure communication*, Univ. J. Math. Appl. (4) (2021), 50–58.

[20] Kaouache, S. *General method for hybrid projective combination synchronization of a class of nonlinear fractional-order chaotic systems*, Int. J. Appl. Math. 36(4) (2023), 509–520.

[21] Kaouache, S. and Abdelouahab, M.S. *Modified projective synchronization between integer order and fractional order hyperchaotic systems*, JARDCS 10(5) (2018), 96–104.

[22] Kaouache, S. and Abdelouahab, M.S. *Generalized synchronization between two chaotic fractional non-commensurate order systems with different dimensions*, Nonlinear Dyn. Syst. Theory. 18 (2018), 273–284.

[23] Kaouache, S. and Abdelouahab, M.S. *Inverse matrix projective synchronization of novel hyperchaotic system with hyperbolic sine function non-linearity*, Dyn. Contin. Discrete Impuls. Syst. Ser. B: Appl. Alg. 27 (2020), 145–154.

[24] Kaouache, S., Abdelouahab, M.S. and Bououden, R. *Reduced generalized combination synchronization between two n-dimentional integer-order hyperchaotic systems and one m-dimentional fractional-order chaotic system*, Aust. J. Math. Anal. Appl. 17(2) (2020), Art. 19, 8 pp.

[25] Kaouache, S. and Bouden, T. *Modified hybrid synchronization of identical fractional hyperchaotic systems with incommensurate order*, Dyn. Contin. Discrete Impuls. Syst. Ser. A: Math. Anal. 28 (2021), 25–36.

[26] Kaouache, S., Hamri, N.D., Hacinliyan, A.S., Kandiran, E., Deruni, B. and Keles, A.C. *Increaszd order generalized combination synchronization of non identical dimensional fractional-order systems by introducing different observable variable functions*, Nonlinear Dyn. Syst. Theory, 20 (2020), 307–315.

[27] Kiani-B, A., Fallahi, L., Pariz, K. and Leung, H. *A chaotic secure communication scheme using fractional chaotic systems based on an extended fractional Kalman filter*, Commun. Nonlinear Sci. Numer. Simul. 14 (2009), 863–879.

[28] Kilbas, A., Srivastava H. and Trujillo J. *Theory and applications of fractional differential equations*, Elsevier, 2006.

[29] Labed, B., Kaouache, S. and Abdelouahab, M.S. *Control of a novel class of uncertain fractional-order hyperchaotic systems with external disturbances via sliding mode controller*, Nonlinear Dyn. Syst. Theory, 20(2) (2020), 203–213.

[30] Li, C. and Chen, G. *Chaos in the fractional order Chen system and its control*, Chaos Solitons Fract. 22(3) (2004), 549–554.

[31] Li, C. and Chen, G. *Chaos and hyperchaos in the fractional-order Rossler equations*, Phys. A, 341(1) (2004), 55–61.

[32] Li, C. and Deng, W.H. *Chaos synchronization of fractional-order differential systems*, Int. J. Mod. Phys. B, 20(7) (2006), 791–803.

[33] Lin, J., Yan, J. and Liao, T. *Chaotic synchronization via adaptive sliding mode observers subject to input nonlinearity*, Chaos Solitons Fract. 24 (2005), 371–381.

[34] Lin, L., Zhuang, Y., Xu, Z., Yang, D. and Wu, D. *Encryption algorithm based on fractional order chaotic system combined with adaptive predefined time synchronization*, Front. Phys. 11 (2023), 1202871.

[35] Lu, J. *Chaotic dynamics of the fractional-order Lu system and its synchronization*, Phys. Lett. A, 354(4) (2006), 305–311.

[36] Lu, L. *Synchronization of a class of fractional-order chaotic systems via a scalar transmitted signal*, Chaos Solitons Fract. 27 (2006), 519–525.

[37] Muthukumar, P., Balasubramaniam, P. and Ratnavelu, K. *Synchronization of a novel fractional order stretch-twist-fold (STF) flow chaotic system and its application to a new authenticated encryption scheme (AES),* Nonlinear Dyn. 77(4) (2014), 1547–1559.

[38] Pecora, L.M. and Carroll, T.L. *Synchronization in chaotic systems*, Phys. Rev. Lett. 64(8) (1990), 821–824.

[39] Pham, V.T., Rahma, F., Frasca, M. and Fortuna, L. *Dynamics and synchronization of a novel hyperchaotic system without equilibrium*, Int. J. Bifurcat. Chaos, 24(6) (2014), 1450087.

[40] Pikovsky, A.S., Rosenblum, M.G., Osipov, G.V. and Kurths, J. *Phase synchronization of chaotic oscillators by external driving*, Phys. D, 104(3) (1997), 219–238.

[41] Rikitake, T. *Oscillations of a system of disk dynamos*, Math. Proc. Cambridge, 54(1) (1958), 89–105.

[42] Rossler, O. *An equation for hyperchaos*, Phys. Lett. A, 71(2-3) (1979), 155–157.

[43] Sau, N.H., Binh, T.N., Thanh, N.T. and Thuan, M.V. *Event-triggered H_∞ controller design for uncertain fractional-order systems with time-varying delays*, J. Appl. Math. Comput. 70(3) (2024), 1813–1835.

[44] Shao, S. and Chen, M. *Fractional-order control for a novel chaotic system without equilibrium*, J. Autom. Sin. 6(4) (2016), 1000–1009.

[45] Sheu, L. *A speech encryption using fractional chaotic systems*, Nonlinear Dyn. 65 (2011), 103–108.

[46] Tabasi, M., Hosseini, S.A. and Houshmand, M. *Stability analysis of two fractional-order Duffing-Holmes chaotic systems using a generalized predictive controller*, Int. J. Dyn. Control. (2024), 1–12.

[47] Vafaei, V., Jodayree Akbarfam, A. and Kheiri, H. *A new synchronisation method of fractional-order chaotic systems with distinct orders and dimensions and its application in secure communication*, Int. J. Syst. Sci. 52(16) (2021), 3437–3450.

[48] Vaidyanathan, S., Dolvis, L. G., Jacques, K., Lien, C. H. and Sambas, A. *A new five-dimensional four-wing hyperchaotic system with hidden attractor, its electronic circuit realisation and synchronisation via integral sliding mode control*, Int. J. Model. Identif. Control, 32(1) (2019), 30–45.

[49] Vaidyanathan, S., Volos, Ch.K. and Pham, V.T. *Analysis, control, synchronization and SPICE implementation of a novel $4-D$ hyperchaotic Rikitake Dynamo system without equilibrium*, J. Eng. Technol. Rev. 8(2) (2015), 232–244.

[50] Volos, Ch.K., Kyprianidis, I.M. and Stouboulos, I.N. *Image encryption process based on chaotic synchronization phenomena*, Signal Proc. 93(5) (2013), 1328–1340.

[51] Wang, Z., Cang, S., Ochola, E.O. and Sun, Y. *A hyperchaotic system without equilibrium*, Nonlinear Dyn. 69 (2012), 531–537.

[52] Wei, J., Zhang, C., Guo, Y. and Wang, F. *Cluster synchronization of stochastic two-layer delayed neural networks via pinning impulsive control*, J. Appl. Math. Comput. 70(2) (2024), 1193–1210.

[53] Wu, X. and Lu, Y. *Generalized projective synchronization of the fractional-order Chen hyperchaotic system*, Nonlinear Dyn. 57 (2009), 25–35.

[54] Yang, N. and Liu, C. *A novel fractional-order hyperchaotic system stabilization via fractional sliding-mode control*, Nonlinear Dyn. 74 (2013), 721–732.

[55] Zhang, C., Zhang, C., Zhang, X. and Liang, Y. *Sampling-based event-triggered control for cluster synchronization in two-layer nonlinear networks*, J. Appl. Math. Comput. 69(5) (2023), 3969–3986.

[56] Zhang, R. and Yang, S. *Adaptive synchronization of fractional-order chaotic systems via a single driving variable*, Nonlinear Dyn. 66 (2011), 831–837.

[57] Zhang, S. and Zeng, Y. *A simple Jerk-like system without equilibrium: Asymmetric coexisting hidden attractors, bursting oscillation and double full Feigenbaum remerging trees*, Chaos Solitons Fract. 120 (2019), 25–40.

[58] Zhou, J. and Bao, H. *Fixed-time synchronization for competitive neural networks with Gaussian-wavelet-type activation functions and discrete delays*, J. Appl. Math. Comput. 64(1) (2020), 103–118.

Iranian Journal of Numerical Analysis and Optimization
Vol. 15, No. 2, 2025, pp 531–569
https://doi.org/10.22067/ijnao.2025.89033.1482
https://ijnao.um.ac.ir/

Research Article

Exploring the dynamics of lymphatic filariasis through a mathematical model and analysis with Holling type II treatment functions

F.A. Oguntolu, O.J. Peter*, B.I. Omede, T.A. Ayoola and G.B. Balogun

*Corresponding author

Received 22 July 2024; revised 16 November 2024; accepted 17 Janurary 2025

Festus Abiodun Oguntolu[1], festus.tolu@futminna.edu.ng
Olumuyiwa James Peter[2], peterjames4real@gmail.com
Benjamin Idoko Omede[3], benjaminomede197@gmail.com
Tawakalt Abosede Ayoola[4], tawakalt.alade@uniosun.edu.ng
Ghaniyyat Bolanle Balogun[5], balogun.gb@unilorin.edu.ng

[1]Department of Mathematics, Federal University of Technology, Minna, Niger State, Nigeria.
[2]Department of Mathematical and Computer Sciences, University of Medical Sciences, Ondo State, Nigeria.
Department of Epidemiology and Biostatistics, School of Public Health, University of Medical Sciences, Ondo State, Nigeria.
[3]Department of Mathematical Sciences, Prince Abubakar Audu (Formerly Kogi State) University, Anyigba, Nigeria.
[4]Department of Mathematical Sciences, Osun State University Osogbo, Nigeria.
[5]Department of Computer Science, University of Ilorin, Nigeria.

How to cite this article
Oguntolu, F.A. Peter, O.J., Omede, B.I., Ayoola, T.A. and Balogun, G.B., Exploring the dynamics of lymphatic filariasis through a mathematical model and analysis with Holling type II treatment functions. *Iran. J. Numer. Anal. Optim.*, 2025; 15(2): 531-569. https://doi.org/10.22067/ijnao.2025.89033.1482

Abstract

This paper presents a robust deterministic mathematical model incorporating Holling type II treatment functions to comprehensively investigate the dynamics of Lymphatic filariasis. Through qualitative analysis, the model demonstrates the occurrence of backward bifurcation when the basic reproduction number is less than one. Moreover, numerical simulations are employed to illustrate and validate key analytical findings. These simulation results emphasize the significance of accessible medical resources and the efficacy of prophylactic drugs in eradicating Lymphatic filariasis. The findings show that, enhancing medical resource availability and implementing effective treatment strategies in rural areas and regions vulnerable to Lymphatic filariasis is crucial for combating the transmission and control of this disease.

AMS subject classifications (2020): 92B05; 92D30; 92-10; 34D20

Keywords: Lymphatic filariasis; Holling type II treatment function; Reproduction number; Stability; Bifurcation.

1 Introduction

Lymphatic filariasis (LF), also known as Elephantiasis, is a neglected tropical disease that is prevalent in developing countries and disadvantaged communities across Sub-Sahara Africa, Asia, South and Central America, and the Pacific Island nations [22]. It is a parasitic disease transmitted through vectors and caused by three types of nematode parasites that reside in the lymphatic system: Wuchereria bancrofti, Brugia Malayi, and B. timori. Wuchereria bancrofti is responsible for the majority of cases worldwide, while Brugia Malayi and B. timori play significant roles as local causes in Southeast Asia. The transmission of these nematode parasites occurs through various species of mosquito vectors, including Anopheles, Aedes, Culex, Mansonia, and Ochlerotatus [8, 5]. The adult parasites take up residence in the lymphatic system of the human host. Following mating, they produce first stage larvae known as microfilariae, which then migrate to the peripheral blood circulation. These microfilariae exhibit a diurnal periodicity, residing in the

deep veins during the day and migrating to the peripheral circulation at night. The peak concentration of microfilariae in the peripheral blood occurs during a 4-hour timeframe from 10 p.m. to 4 a.m. This adaptation aligns with the biting behavior of the mosquito vectors [15, 10].

The extrinsic life cycle commences when the mosquito ingests microfilariae along with human blood during a bite. The microfilariae traverse the gut wall of the mosquito and reach the thoracic muscles, where they undergo a transformation, becoming shorter and thicker. These microfilariae progress into the first stage larvae (L1). Within a span of 5 to 7 days, the L1 larvae grow and develop into the more active second stage (L2). By approximately 10 to 11 days, they mature into infective stage larvae (L3) [1]. After reaching full maturity, the majority of infective larvae (L3) relocate to the proboscis of the mosquito, poised to infect another human [1]. This intricate life cycle involving mosquitoes as intermediaries is the primary means of transmitting LF. When a mosquito bites a human host, the L3 larvae are deposited on the skin's surface. Upon retracting its proboscis, the larvae enter the wound and travel to the lymphatic system. Approximately 9 to 10 days after entry, the L3 larvae molt and transform into fourth stage larvae (L4). The L4 stage undergoes further development, taking several days to a few months before reaching adulthood [1, 7, 12, 2].

In April 2012, LF was endemic in 73 countries and territories, putting an estimated 1.39 billion people at risk of infection, with approximately 120 million already infected [34]. Most infected individuals remain asymptomatic and are unaware of their condition unless tested [35]. However, these asymptomatic infections gradually cause damage to the lymphatic system, kidneys, and disrupt the body's immune system. The severity of symptoms is directly influenced by the patient's immune response and may manifest as acute inflammation of the lymphatic vessels, accompanied by high temperatures, chills, body aches, and swollen lymph nodes [1]. Excessive fluid accumulation in the affected tissues can occur [35, 13]. In chronic cases, LF leads to conditions such as lymphedema (tissue swelling) or elephantiasis (thickening of the skin), as well as swelling in the limbs and hydrocele (scrotal and breast swelling). In some instances, filarial abscesses may develop [1, 7, 12, 2].

In 2000, the World Health Organization (WHO) initiated the Global Program to Eliminate Lymphatic Filariasis (GPELF) with the ambitious target of eradicating LF by 2030. The strategy employed by the GPELF focuses on interrupting transmission through mass drug administration (MDA), while also addressing the suffering and disability caused by chronic manifestations of the disease [20, 36, 19]. The approach to interrupting transmission involves administering a combination of two drugs annually to the entire at-risk population. In areas without onchocerciasis, the treatment consists of albendazole (400mg) combined with diethylcarbamazine (6mg/kg) in areas where onchocerciasis and LF coexist [32, 27]. The mass administration of drugs typically extends over a period of 4-6 years, allowing sufficient time for adult parasites to exhaust their reproductive lifespan. However, in areas with low treatment coverage or high transmission intensity, longer campaigns may be necessary to ensure the interruption of transmission [27, 33].

Mathematical modeling has played a crucial role in understanding infectious diseases, although only a limited number of models have been developed specifically for studying the transmission dynamics and control of LF. For instance, Bhunu and Mushayabasa [3] created an epidemiological model to analyze the spread of LF. Their analysis of the reproduction number indicated that treatment would contribute to a reduction in LF cases, but the magnitude of the reduction was not quantified. Their numerical simulations suggested that effective LF control might require treatment for symptomatic individuals and chemoprophylaxis for those at risk. Mwamtobe et al. [21] formulated and analyzed a mathematical model for LF, incorporating intervention strategies. Their findings suggested that treatment led to a faster reduction in LF cases compared to quarantine. However, the greatest reduction occurred when both intervention approaches were implemented simultaneously. Oguntolu et al. [24] developed and rigorously analyzed a mathematical model for the transmission dynamics of LF, accounting for asymptomatic and symptomatic infections, as well as treatment. Their simulations demonstrated that a treatment coverage level of 75% among the human population was necessary to control LF. The successful elimination of LF by 2030 relies on the availability of medical resources and the implementation of effective treatments. In traditional epidemic models, the treatment rate for infected

individuals is often assumed to be constant or proportional to the number of infected cases. However, it is important to consider the limited availability of treatment resources within the community when determining an appropriate treatment rate for the disease [17]. Wang and Ruan [30] consider an SIR epidemic model with a constant treatment rate, the authors investigated the stability of the model and showed that the model exhibits various bifurcation. Furthermore, Zhou and Fan [37] modified the treatment rate to Holling type II, by varying the amount of medical resources and their supply efficiently, the target model admits both backward and Hopf bifurcation.

In this paper, our study focuses on the dynamics of how LF spreads and its control. This is the first research to examine the initial phase of infection by taking into account susceptible individuals who are undergoing prophylactic treatment. Furthermore, we have integrated a Holling type II treatment function at each stage of infection to enhance our understanding of the treatment and control strategies for LF. The rest of the paper is organized as follows: the model formulation and basic properties of the model described in section 2, the model analysis in section 3, the numerical analysis in section 4, and the concluding remarks in section 5.

2 Model formulations

The total human population at time t, denoted by $N_H(t)$, is sub-divided into seven mutually exclusive compartments of susceptible individuals not on prophylactic treatments $S_H(t)$, susceptible individuals on prophylactic treatments $S_{HP}(t)$, exposed individuals $E_H(t)$, L3-larvae stage (initial stage) infected individuals $L(t)$, acute stage infected individuals $A(t)$, Chronic stage infected individuals $C(t)$ and treated individuals $T(t)$. So that

$$N_H(t) = S_H(t) + S_{HP}(t) + E_H(t) + L(t) + A(t) + C(t) + T(t).$$

The vector (mosquitoes) population at time t, denoted by $N_V(t)$, is sub-divided into three compartments of susceptible mosquitoes $S_V(t)$, exposed mosquitoes $E_V(t)$, and infectious mosquitoes $I_V(t)$. So that

$$N_V(t) = S_V(t) + E_V(t) + I_V(t).$$

The susceptible individuals not on prophylactic treatments $S_H(t)$ is generated by birth or immigration at a constant rate Π_H. Susceptible individuals not on prophylactic treatments acquire LF infection through effective contact with infectious mosquitoes at the rate λ_H, given by

$$\lambda_H = \frac{\phi\beta_H I_V}{N_H},$$

where ϕ is the mosquito-biting rate, and β_H is the transmission probability from $I_V(t)$ to $S_H(t)$. The population of susceptible individuals on prophylactic treatments $S_{HP}(t)$ is generated by susceptible individuals not on prophylactic treatments $S_{HP}(t)$ that started receiving prophylactic treatment at the rate α, the susceptible individuals on prophylactic treatments acquire LF infection through effective contact with infectious mosquitoes at the rate $(1-\omega)\lambda_H$, where ω is the efficacy of the prophylaxis drugs. Exposed individuals progressed to the L3-larvae (initial stage of infection) stage at the rate η, L3-larvae stage infected individuals progressed to acute stage at the rate σ, and acute stage infected individuals progressed to chronic stage at the rate v. The human natural death rate γ_H is the same in all human compartments. We considered the Holling type II treatment functions $f(L)$, $f(A)$, and $f(C)$ for the L3-larvae stage infected individuals, the Acute stage infected individuals, and the Chronic stage infected individuals respectively, given by

$$f(L) = \frac{\tau_1 L}{1+\varepsilon L}, \quad f(A) = \frac{\tau_2 A}{1+\varepsilon A}, \quad \textit{and} \quad f(C) = \frac{\tau_3 C}{1+\varepsilon C},$$

where τ_1, τ_2, and τ_3 are the treatment rates for L3-larvae stage, Acute stage, and Chronic stage infected individuals respectively, and ε is the limitation rate in medical resources availability. The population of the susceptible mosquitoes $S_V(t)$, is generated by the recruitment rate Π_V. Susceptible mosquitoes acquires LF infection through effective contact with infected individuals at the rate λ_V, given by

$$\lambda_V = \frac{\phi\beta_V(L+A+C)}{N_H},$$

where β_V is the transmission probability from infected individuals to susceptible mosquitoes and ϕ is the mosquito-biting rate. The exposed mosquitoes

progressed to being infectious at the rate ρ. The death rate of mosquitoes is given by γ_V.

Based on the above formulation and assumptions, the LF model is given by the following system of deterministic nonlinear ordinary differential equations in (1) while the schematic diagram is presented in Figure 1. The description of the model state variables and parameters is presented in Table 1.

$$\begin{aligned}
\frac{dS_H}{dt} &= \Pi_H - \frac{\phi\beta_H I_V S_H}{N_H} - (\alpha+\gamma_H) S_H + \kappa T,\\
\frac{dS_{HP}}{dt} &= \alpha S_H - \frac{(1-\omega)\phi\beta_H I_V S_{HP}}{N_H} - \gamma_H S_{HP},\\
\frac{dE_H}{dt} &= \frac{\phi\beta_H I_V S_H}{N_H} + \frac{(1-\omega)\phi\beta_H I_V S_{HP}}{N_H} - (\eta+\gamma_H) E_H,\\
\frac{dL}{dt} &= \eta E - \frac{\tau_1 L}{1+\varepsilon L} - (\sigma+\gamma_H) L,\\
\frac{dA}{dt} &= \sigma L - \frac{\tau_2 A}{1+\varepsilon A} - (v+\gamma_H) A,\\
\frac{dC}{dt} &= vA - \frac{\tau_3 C}{1+\varepsilon C} - \gamma_H C,\\
\frac{dT}{dt} &= \frac{\tau_1 L}{1+\varepsilon L} + \frac{\tau_2 A}{1+\varepsilon A} + \frac{\tau_3 C}{1+\varepsilon C} - (\kappa+\gamma_H) T,\\
\frac{dS_V}{dt} &= \Pi_V - \frac{\phi\beta_V (L+A+C) S_V}{N_H} - \gamma_V S_V,\\
\frac{dE_V}{dt} &= \frac{\phi\beta_V (L+A+C) S_V}{N_H} - (\rho+\gamma_V) E_V,\\
\frac{dI_V}{dt} &= \rho E_V - \gamma_V I_V.
\end{aligned} \tag{1}$$

2.1 Basic properties of the model

2.1.1 *Positivity and boundedness of solution*

In order for the model (1) to be biological and mathematically meaningful, it is pertinent to show that the state variables of the model are nonnegative for all time $t > 0$.

Theorem 1. Let the initial data be $S_H(0) \geq 0$, $S_{HP}(0) \geq 0$, $E_H(0) \geq 0$, $L(0) \geq 0$, $A(0) \geq 0$, $C(0) \geq 0$, $T(0) \geq 0$, $S_V(0) \geq 0$, $E_V(0) \geq 0$, and

Table 1: Description of the model state variables and parameters

Variable	Description
S_H	Susceptible individuals not on prophylactic treatment.
S_{HP}	Susceptible individuals on prophylactic treatment.
E_H	Exposed individuals.
L	L3-larvae stage (initial stage) infected individuals.
A	Acute stage infected individuals.
C	Chronic stage infected individuals.
T	Treated individuals.
S_V	Susceptible mosquitoes.
E_V	Exposed mosquitoes.
I_V	Infectious mosquitoes.
Parameter	**Description**
Π_H	Recruitment rate for humans.
Π_V	Recruitment rate for mosquitoes.
ϕ	Mosquito biting rate.
β_H	Transmission probability from I_V to S_H.
β_V	Transmission probability from L, A, and C to S_V.
γ_H	Natural death rate for humans.
γ_V	Death rate for mosquitoes.
α	Progression rate from S_H to S_{HP}.
σ	Progression rate from L to A.
ω	Efficacy of prophylaxis drugs.
η	Progression rate from E_H to L.
ν	Progression rate from A to C.
τ_1	Treatment rate of L3-larvae stage infection.
τ_2	Treatment rate of Acute stage infection.
τ_3	Treatment rate of Chronic stage infection.
ε	Limitation rate in medical resources availability.
κ	Rate at which treated individuals becomes susceptible again.
ρ	Progression rate from E_V to I_V.

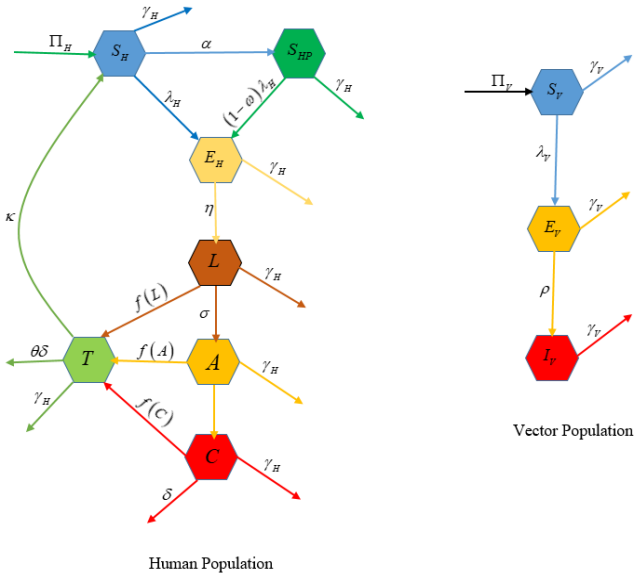


Figure 1: Schematic diagram of the flowchart

$I_V(0) \geq 0$. Then, the solution $(S_H, S_{HP}, E_H, L, A, C, T, S_V, E_V, I_V)$ to the model (1) is nonnegative for all $t > 0$.

Proof. Let $t_1 = \sup[t > 0 : S_H(t) > 0, S_{HP}(t) > 0, E_H(t) > 0, L(t) > 0, A(t) > 0, C(t) > 0, T(t) > 0, S_V(t) > 0, E_V(t) > 0, I_V(t) > 0 \in [0, t]]$. Then $t_1 > 0$. We have from the first equation of the model (1) that

$$\frac{dS_H}{dt} = \Pi_H - \lambda_H S_H - (\alpha + \gamma_H) S_H + \kappa T.$$

Solving the equation above, we have

$$\begin{aligned} &\frac{d}{dt}\left\{S_H(t)\left[\exp\left(\int_0^t \lambda_H(\zeta)\, d\zeta + (\alpha+\gamma_H)\, t\right)\right]\right\} \\ &= (\Pi_H + \kappa T) \times \exp\left(\int_0^t \lambda_H(\zeta)\, d\zeta + (\alpha+\gamma_H)\, t\right). \end{aligned}$$

Integrating the above equations at the range $[0, t_1]$, we obtained

$$\begin{aligned} &\left\{S_H(t_1)\exp\left[\int_0^{t_1} \lambda_H(\zeta)\, d\zeta + (\alpha+\gamma_H)\, t_1\right]\right\} - S_H(0) \\ &= (\Pi_H + \kappa T) \times \int_0^{t_1} \exp\left[\int_0^x \lambda_H(\zeta)\, d\zeta + (\alpha+\gamma_H)\, x\right] dx. \end{aligned}$$

So that

$$\begin{aligned} S_H(t_1) =& S_H(0) \exp\left[-\left(\int_0^{t_1} \lambda_H(\zeta)\, d\zeta + (\alpha+\gamma_H)\, t_1\right)\right] \\ &+ \exp\left[-\left(\int_0^{t_1} \lambda_H(\zeta)\, d\zeta + (\alpha+\gamma_H)\, t_1\right)\right] \\ &\times (\Pi_H + \kappa T) \int_0^{t_1} \exp\left[\int_0^{x} \lambda_H(\zeta)\, d\zeta + (\alpha+\gamma_H)\, x\right] dx > 0. \end{aligned}$$

Similarly, it can be shown that $S_{HP} > 0, E_H > 0, L > 0, A > 0, C > 0, T > 0, S_V > 0, E_V > 0, I_V > 0$. □

2.1.2 Invariant region

We considered that the region in which the solution of the LF model (1) is feasible, given by

$$\Omega = \Omega_H \times \Omega_V \subset \mathbb{R}_+^7 \times \mathbb{R}_+^3, \tag{2}$$

where

$$\Omega_H = \left\{(S_H, S_{HP}, E_H, L, A, C, T) \in \mathbb{R}_+^7 : N_H \leq \frac{\Pi_H}{\gamma_H}\right\},$$

and

$$\Omega_V = \left\{(S_V, E_V, I_V) \in \mathbb{R}_+^3 : N_V \leq \frac{\Pi_V}{\gamma_V}\right\}.$$

Lemma 1. The region $\Omega = \left\{(S_H, S_{HP}, E_H, L, A, C, T, S_V, E_V, I_V) \in \mathbb{R}_+^{10} : N_H \leq \frac{\Pi_H}{\gamma_H}, N_V \leq \frac{\Pi_V}{\gamma_V}\right\}$ is positively invariant for the model (1).

Proof. By adding the human and vector compartments of the model system (1), we obtained

$$\begin{aligned} \frac{dN_H}{dt} &= \Pi_H - \gamma_H N_H, \\ \frac{dN_V}{dt} &= \Pi_V - \gamma_V N_V. \end{aligned} \tag{3}$$

Thus, the standard comparison theorem [18] is used to show that

$$\begin{aligned} N_H(t) &\leq N_H(0)\, e^{-\gamma_H t} + \frac{\Pi_H}{\gamma_H}\left(1 - e^{-\gamma_H t}\right), \\ N_V(t) &\leq N_V(0)\, e^{-\gamma_V t} + \frac{\Pi_V}{\gamma_V}\left(1 - e^{-\gamma_V t}\right). \end{aligned} \tag{4}$$

In particular, if $N_H(0) \leq \frac{\Pi_H}{\gamma_H}$ and $N_V(0) \leq \frac{\Pi_V}{\gamma_V}$, then $N_H(t) \leq \frac{\Pi_H}{\gamma_H}$ and $N_V(t) \leq \frac{\Pi_V}{\gamma_V}$. Therefore, the region Ω is a positively-invariant (that is, the solutions to the model remain positive for all time t) and the model is well posed and biologically meaningful. □

3 Model analysis

In this section, we shall be analyzing the model for the stability of the disease-free equilibrium (DFE) and endemic equilibrium.

3.1 Disease-free equilibrium (DFE)

The DFE points are steady state solutions where there is no LF infection in the community. This is achieved by setting the infected compartments to zero (that is, $E_H = L = A = C = T = E_V = I_V = 0$) and setting the right-hand side of the model system (1) to zero. Therefore, the DFE point, (ξ_0), is given by

$$\xi_0 = (S_H^*, S_{HP}^*, E_H^*, L^*, A^*, C^*, T^*, S_V^*, E_V^*, I_V^*) = (S_H^*, S_{HP}^*, 0, 0, 0, 0, 0, S_V^*, 0, 0), \quad (5)$$

where

$$S_H^* = \frac{\Pi_H}{\alpha + \gamma_H}, S_{HP}^* = \frac{\alpha \Pi_H}{(\alpha + \gamma_H)\gamma_H}, \text{ and } S_V^* = \frac{\Pi_V}{\gamma_V}.$$

3.2 Basic reproduction number

The basic reproductive number R_0 is a measurement of the potential for the spreading disease in a population. Mathematically, R_0 is a threshold parameter for the stability of a DFE and is related to the peak and final size of an epidemic. It is defined as the expected number of secondary cases of infection, which would occur due to a primary case in a completely susceptible population [9, 29]. It is an important parameter that governs the spread of a disease.

The reproductive number (R_0) can be computed using the next generation operator method described in [9, 29]. Using the matrices F and V, for the new infection and the remaining transition terms, respectively,

$$F = \begin{bmatrix} 0 & 0 & 0 & 0 & 0\ 0 & \frac{\phi\beta_H(S_H^*+(1-\omega)S_{HP}^*)}{N_H} \\ 0 & 0 & 0 & 0 & 0\ 0 & 0 \\ 0 & 0 & 0 & 0 & 0\ 0 & 0 \\ 0 & 0 & 0 & 0 & 0\ 0 & 0 \\ 0 & 0 & 0 & 0 & 0\ 0 & 0 \\ 0 & \frac{\phi\beta_V S_V^*}{N_H} & \frac{\phi\beta_V S_V^*}{N_H} & \frac{\phi\beta_V S_V^*}{N_H} & 0\ 0 & 0 \\ 0 & 0 & 0 & 0 & 0\ 0 & 0 \end{bmatrix},$$

and

$$V = \begin{bmatrix} (\eta+\gamma_H) & 0 & 0 & 0 & 0 & 0 & 0 \\ -\eta & (\sigma+\tau_1+\gamma_H) & 0 & 0 & 0 & 0 & 0 \\ 0 & -\sigma & (\nu+\tau_2+\gamma_H) & 0 & 0 & 0 & 0 \\ 0 & 0 & -\nu & (\tau_3+\gamma_H) & 0 & 0 & 0 \\ 0 & -\tau_1 & -\tau_2 & -\tau_3 & (\kappa+\gamma_H) & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & (\rho+\gamma_V) & 0 \\ 0 & 0 & 0 & 0 & 0 & -\rho & \gamma_V \end{bmatrix}.$$

Hence, it follows from [9] that $R_0 = \rho\left(FV^{-1}\right)$, where ρ is the spectral radius or largest eigenvalues of $\left(FV^{-1}\right)$.

Therefore,

$$R_0 = \sqrt{\frac{\Pi_V \gamma_H \phi^2 \beta_V \beta_H \left(\gamma_H + (1-\omega)\alpha\right)\left(\sigma\left(v + B_5\right) + B_4 B_5\right)\eta\rho}{\Pi_H \gamma_V^2 B_1 B_2 B_3 B_4 B_5 B_7}}, \quad (6)$$

where

$B_1 = \alpha + \gamma_H$, $B_2 = \eta + \gamma_H$, $B_3 = \sigma + \tau_1 + \gamma_H$, $B_4 = v + \tau_2 + \gamma_H$, $B_5 = \tau_3 + \gamma_H$, $B_6 = \kappa + \gamma_H$, and $B_7 = \rho + \gamma_V$.

3.3 Local stability of the DFE

Theorem 2. The DFE (ξ_0) of the model (1) is locally asymptotically stable if $R_0 < 1$, and unstable if $R_0 > 1$.

Proof. The local stability of the DFE (ξ_0) of the model (1) is analyzed by the Jacobian matrix of the model system (1) evaluated at the DFE (ξ_0), given by

$$J(\xi_0) = \begin{bmatrix}
-B_1 & 0 & 0 & 0 & 0 & 0 & \kappa & 0 & 0 & -\frac{\phi\beta_H\gamma_H}{B_1} \\
\alpha & -\gamma_H & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{\phi\beta_H D_1}{B_1} \\
0 & 0 & -B_2 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{\phi\beta_H D_2}{B_1} \\
0 & 0 & \eta & -B_3 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \sigma & -B_4 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \nu & -B_5 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & \tau_1 & \tau_2 & \tau_3 & -B_6 & 0 & 0 & 0 \\
0 & 0 & 0 & -D_3 & -D_3 & -D_3 & 0 & -\gamma_V & 0 & 0 \\
0 & 0 & 0 & D_3 & D_3 & D_3 & 0 & 0 & -B_7 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \rho & -\gamma_V
\end{bmatrix},$$

where $B_1 = \alpha + \gamma_H$, $B_2 = \eta + \gamma_H$, $B_3 = \sigma + \tau_1 + \gamma_H$, $B_4 = \upsilon + \tau_2 + \gamma_H$, $B_5 = \tau_3 + \gamma_H$, $B_6 = \kappa + \gamma_H$, $B_7 = \rho + \gamma_V$, $D_1 = (1-\omega)\,\alpha$, $D_2 = \gamma_H + (1-\omega)\,\alpha$, and $D_3 = \frac{\phi\beta_V \Pi_V \gamma_H}{\Pi_H \gamma_V}$.

The eigenvalues of the Jacobian matrix $J(\xi_0)$ are $\lambda_1 = -\gamma_H$, $\lambda_2 = -B_1$, $\lambda_3 = -B_6$, $\lambda_4 = -\gamma_V$ and the roots of the characteristic polynomial below:

$$P(\lambda) = \lambda^6 + a_1\lambda^5 + a_2\lambda^4 + a_3\lambda^3 + a_4\lambda^2 + a_5\lambda + a_6, \tag{7}$$

where

$$\begin{aligned}
a_1 =& B_2 + B_3 + B_4 + B_5 + B_7 + \gamma_V,\\
a_2 =& B_2(B_3 + B_4 + B_5 + B_7 + \gamma_V) + B_3(B_4 + B_5 + B_7 + \gamma_V) + B_7\gamma_V\\
&+ B_4(B_5 + B_7 + \gamma_V) + B_5(B_7 + \gamma_V),\\
a_3 =& B_2B_3(B_4 + B_5 + B_7 + \gamma_V) + (B_2B_4 + B_3B_4)(B_5 + B_7 + \gamma_V)\\
&+ (B_7 + \gamma_V)(B_2B_5 + B_3B_5) + B_7\gamma_V(B_2 + B_3 + B_4 + B_5)\\
&+ B_4B_5(B_7 + \gamma_V),\\
a_4 =& \gamma_V(B_2B_3(B_4 + B_5 + B_7) + B_4(B_2 + B_3)(B_5 + B_7))\\
&+ \gamma_V B_5B_7(B_2 + B_3 + B_4) + B_2B_3B_4(B_5 + B_7)\\
&+ B_5B_7(B_2(B_3 + B_4) + B_3B_4) - \frac{\Pi_V\gamma_H\phi^2\beta_H\beta_V D_2\eta\rho}{\Pi_H\gamma_V B_1},\\
a_5 =& \gamma_V(B_2B_3(B_4B_5 + B_4B_7 + B_5B_7) + B_4B_5B_7(B_2 + B_3))\\
&+ B_2B_3B_4B_5B_7 - \frac{\Pi_V\gamma_H\phi^2\beta_H\beta_V D_2\eta\rho(\sigma + B_4 + B_5)}{\Pi_H\gamma_V B_1},\\
a_6 =& \gamma_V B_2B_3B_4B_5B_7\left(1 - R_0^2\right).
\end{aligned}$$

Applying the Routh–Hurwitz criterion [23], which states that all roots of the polynomial (7) have negative real parts if and only if the coefficient of $a_i > 0$, for $i = 1, 2, 3, 4, 5, 6$. Clearly, for $a_6 > 0$ then $R_0 < 1$. Therefore, all the eigenvalues of the Jacobian matrix $J(\xi_0)$ have negative real parts when $R_0 < 1$ and the DFE point is locally asymptotically stable. □

Biologically, the implication of Theorem 2 is that a small influx of LF infected individuals in the population will not cause an LF outbreak as long as the reproduction number (R_0) is less than unity [26]. It is important to

note that this result depends on the initial sizes of the infected individuals in the population.

3.4 Existence of an endemic equilibrium

In this section, we shall investigate the existence of an endemic equilibrium point. The endemic equilibrium is a steady state where the disease persists in the population. In this case, the infected variables are nonzero (that is, $E_H \neq 0$, $L \neq 0$, $A \neq 0$, $C \neq 0$, $T \neq 0$, $E_V \neq 0$, $I_V \neq 0$). We shall be investigating the endemic equilibrium using the "*forces of infection*", given by

$$\lambda_H^{**} = \frac{\phi\beta_H I_V^{**}}{N_H^{**}}, and \quad \lambda_V^{**} = \frac{\phi\beta_V \left(L^{**} + A^{**} + C^{**}\right)}{N_H^{**}}, \tag{8}$$

where

$$N_H^{**} = S_H^{**} + S_{HP}^{**} + E_H^{**} + L^{**} + A^{**} + C^{**} + T^{**}.$$

The equations of the model system (1) are equated to zero and solved to get an expression of each of the model variables in terms of the forces of infection. For convenience, we let $\varepsilon L^{**} = Z_1$, $\varepsilon A^{**} = Z_2$, and $\varepsilon C^{**} = Z_3$. So that

$$\begin{aligned}
S_H^{**} &= \frac{\Pi_H B_2 B_6 \left((1-\omega)\lambda_H^{**} + \gamma_H\right) Q_1}{B_2 B_6 \left(\lambda_H^{**} + B_1\right)\left((1-\omega)\lambda_H^{**} + \gamma_H\right) Q_1 - \lambda_H^{**}\kappa\eta\left((1-\omega)\left(\lambda_H^{**} + \alpha\right) + \gamma_H\right) Q_2},\\
S_{HP}^{**} &= \frac{\alpha\Pi_H B_2 B_6 Q_1}{B_2 B_6 \left(\lambda_H^{**} + B_1\right)\left((1-\omega)\lambda_H^{**} + \gamma_H\right) Q_1 - \lambda_H^{**}\kappa\eta\left((1-\omega)\left(\lambda_H^{**} + \alpha\right) + \gamma_H\right) Q_2},
\end{aligned} \tag{9}$$

$$\begin{aligned}
E_H^{**} &= \frac{\lambda_H^{**}\Pi_H B_6 \left((1-\omega)\left(\lambda_H^{**} + \alpha\right) + \gamma_H\right) Q_1}{B_2 B_6 \left(\lambda_H^{**} + B_1\right)\left((1-\omega)\lambda_H^{**} + \gamma_H\right) Q_1 - \lambda_H^{**}\kappa\eta\left((1-\omega)\left(\lambda_H^{**} + \alpha\right) + \gamma_H\right) Q_2},\\
L^{**} &= \frac{\lambda_H^{**}\Pi_H B_6\eta\left(1+Z_1\right)\left((1-\omega)\left(\lambda_H^{**} + \alpha\right) + \gamma_H\right) Q_3}{B_2 B_6 \left(\lambda_H^{**} + B_1\right)\left((1-\omega)\lambda_H^{**} + \gamma_H\right) Q_1 - \lambda_H^{**}\kappa\eta\left((1-\omega)\left(\lambda_H^{**} + \alpha\right) + \gamma_H\right) Q_2},\\
A^{**} &= \frac{\lambda_H^{**}\Pi_H B_6\eta\sigma\left(1+Z_1\right)\left(1+Z_2\right)\left((1-\omega)\left(\lambda_H^{**} + \alpha\right) + \gamma_H\right)\left(B_5 + \gamma_H Z_3\right)}{B_2 B_6 \left(\lambda_H^{**} + B_1\right)\left((1-\omega)\lambda_H^{**} + \gamma_H\right) Q_1 - \lambda_H^{**}\kappa\eta\left((1-\omega)\left(\lambda_H^{**} + \alpha\right) + \gamma_H\right) Q_2},\\
C^{**} &= \frac{\lambda_H^{**}\Pi_H B_6\eta\sigma v\left(1+Z_1\right)\left(1+Z_2\right)\left(1+Z_3\right)\left((1-\omega)\left(\lambda_H^{**} + \alpha\right) + \gamma_H\right)}{B_2 B_6 \left(\lambda_H^{**} + B_1\right)\left((1-\omega)\lambda_H^{**} + \gamma_H\right) Q_1 - \lambda_H^{**}\kappa\eta\left((1-\omega)\left(\lambda_H^{**} + \alpha\right) + \gamma_H\right) Q_2},\\
T^{**} &= \frac{\lambda_H^{**}\Pi_H\eta\left((1-\omega)\left(\lambda_H^{**} + \alpha\right) + \gamma_H\right) Q_2}{B_2 B_6 \left(\lambda_H^{**} + B_1\right)\left((1-\omega)\lambda_H^{**} + \gamma_H\right) Q_1 - \lambda_H^{**}\kappa\eta\left((1-\omega)\left(\lambda_H^{**} + \alpha\right) + \gamma_H\right) Q_2},\\
S_V^{**} &= \frac{\Pi_V}{\lambda_V^{**} + \gamma_V}, E_V^{**} = \frac{\lambda_V^{**}\Pi_V}{\left(\lambda_V^{**} + \gamma_V\right) B_7}, \; and \; I_V^{**} = \frac{\lambda_V^{**}\Pi_V\rho}{\left(\lambda_V^{**} + \gamma_V\right) B_7\gamma_V},
\end{aligned} \tag{10}$$

where

$B_1 = \alpha + \gamma_H$, $B_2 = \eta + \gamma_H$, $B_3 = \sigma + \tau_1 + \gamma_H$, $B_4 = \upsilon + \tau_2 + \gamma_H$, $B_5 = \tau_3 + \gamma_H$, $B_6 = \kappa + \gamma_H$, $B_7 = \rho + \gamma_V$,

$Q_1 = (B_3 + (\sigma + \gamma_H) Z_1)(B_4 + (\upsilon + \gamma_H) Z_2)(B_5 + \gamma_H Z_3)$,
$Q_2 = \tau_1 (B_4 + (\upsilon + \gamma_H) Z_2)(B_5 + \gamma_H Z_3) + \tau_2\sigma (1 + Z_1)(B_5 + \gamma_H Z_3) + \tau_3\sigma\upsilon (1 + Z_1)(1 + Z_2)$, and $Q_3 = (B_4 + (\upsilon + \gamma_H) Z_2)(B_5 + \gamma_H Z_3)$.
Substituting (9) into (8), we obtained

$$f(\lambda_H^{**}) = P_1\lambda_H^{**4} + P_2\lambda_H^{**3} + P_3\lambda_H^{**2} + P_4\lambda_H^{**} + P_5, \tag{11}$$

where $P_1 = \Pi_H^2 B_7\gamma_V (1-\omega)^2 (\phi\beta_V B_6 e_1 + \gamma_V e_2) e_2$, $P_2 = K_1 + K_2 - K_3$, $P_3 = K_4 + K_5 + K_6 - (K_7 + K_8)$, $P_4 = K_9 + K_{10} - K_{11}$, and $P_5 = \Pi_H^2\gamma_V^2 B_1^2 B_6^2 B_7 y_1^2 y_2^2 y_3^2 \left(1 - \frac{\Pi_H\Pi_V\phi^2\beta_H\beta_V B_1 B_2 B_6^2\gamma_H\rho Q_1 e_1 (B_1 - \alpha\omega)}{\Pi_H^2\gamma_V^2 B_1^2 B_6^2 B_7 y_1^2 y_2^2 y_3^2}\right)$, with

$$\begin{aligned}
K_1 =& \Pi_H^2 B_7\gamma_V e_2 (B_1 - \alpha\omega)(1-\omega)(\phi\beta_V B_6 e_1 + \gamma_V e_2) \\
&+ \Pi_H^2 B_2 B_6 B_7\gamma_V y_1 y_2 y_3 (1-\omega)^2 (\phi\beta_V B_6 e_1 + \gamma_V e_2), \\
K_2 =& \Pi_H^2 B_7\gamma_V (\phi\beta_V B_6 e_1 (B_1 - \alpha\omega) + \gamma_V (1-\omega) e_2)(1-\omega) e_2, \\
K_3 =& \Pi_H\Pi_V\phi^2\beta_V\beta_H B_6\rho e_1 e_3 (1-\omega), \\
K_4 =& \Pi_H^2 B_7\gamma_V (1-\omega)(\phi\beta_V B_6 e_1 + \gamma_V e_2) B_1 B_2 B_6 y_1 y_2 y_3, \\
K_5 =& \Pi_H^2 B_2 B_6 B_7\gamma_V y_1 y_2 y_3 (1-\omega)(\phi\beta_V B_6 e_1 (B_1 - \alpha\omega) + \gamma_V e_2 (1-\omega)) \\
&+ \Pi_H^2 B_7\gamma_V e_2 (B_1 - \alpha\omega)(\phi\beta_V B_6 e_1 (B_1 - \alpha\omega) + \gamma_V e_2 (1-\omega)), \\
K_6 =& \Pi_H^2\gamma_V^2 B_1 B_2 B_6 B_7 e_2 y_1 y_2 y_3 (1-\omega), \\
K_7 =& \Pi_H\Pi_V\phi^2\beta_V\beta_H B_6\rho e_1 e_4 (1-\omega), \\
K_8 =& \Pi_H\Pi_V\phi^2\beta_V\beta_H B_6\rho e_1 e_3 (B_1 - \alpha\omega), \\
K_9 =& \Pi_H^2 B_7\gamma_V (\phi\beta_V B_6 e_1 (B_1 - \alpha\omega) + \gamma_V (1-\omega) e_2) B_1 B_2 B_6 y_1 y_2 y_3, \\
K_{10} =& \Pi_H^2\gamma_V^2 (e_2 (B_1 - \alpha\omega) + B_2 B_6 y_1 y_2 y_3 (1-\omega)) B_1 B_2 B_6 B_7 y_1 y_2 y_3, \\
K_{11} =& \Pi_H\Pi_V\phi^2\beta_V\beta_H B_6\rho e_1 (B_1 B_2 B_6\gamma_H Q_1 (1-\omega) + e_4 (B_1 - \alpha\omega)), \\
e_1 =& \eta y_4 (y_2 y_3 + \sigma y_5 (y_3 + \upsilon y_6)), \\
e_2 =& B_6 (y_1 y_2 y_3 + \eta y_4 (y_2 y_3 + \sigma y_5 (y_3 + \upsilon y_6))) + \eta Q_2, \\
e_3 =& (1-\omega)(B_2 B_6 Q_1 - \kappa\eta Q_2), \\
e_4 =& B_2 B_6 Q_1 ((1-\omega) B_1 + \gamma_H) - \kappa\eta Q_2 (B_1 - \alpha\omega), \\
y_1 =& B_3 + (\sigma + \gamma_H) Z_1, y_2 = B_4 + (\upsilon + \gamma_H) Z_2, y_3 = B_5 + \gamma_H Z_3,
\end{aligned}$$

$$y_4 = 1 + Z_1, y_5 = 1 + Z_2, y_6 = 1 + Z_3.$$

For the case when $\varepsilon = 0$, the polynomial (11) is reduced to

$$f_{\varepsilon=0}\left(\lambda_H^{**}\right) = M_1\lambda_H^{**4} + M_2\lambda_H^{**3} + M_3\lambda_H^{**2} + M_4\lambda_H^{**} + M_5, \tag{12}$$

where

$$\begin{aligned}
M_1 &= \Pi_H^2 B_7\gamma_V\left(1-\omega\right)^2\left(\phi\beta_V B_6 g_1 + \gamma_V g_2\right) g_2,\\
M_2 &= H_1 + H_2 - H_3,\\
M_3 &= H_4 + H_5 + H_6 - \left(H_7 + H_8\right),\\
M_4 &= H_9 + H_{10} - H_{11},\\
M_5 &= \Pi_H^2\gamma_V^2 B_1^2 B_2^2 B_3^2 B_4^2 B_5^2 B_6^2 B_7\left(1 - R_0^2\right),
\end{aligned}$$

with

$$\begin{aligned}
H_1 =& \Pi_H^2 B_2B_3B_4B_5B_6B_7\gamma_V\left(1-\omega\right)^2\left(\phi\beta_V B_6 g_1 + \gamma_V g_2\right)\\
&+ \Pi_H^2 B_7\gamma_V g_2\left(B_1 - \alpha\omega\right)\left(1-\omega\right)\left(\phi\beta_V B_6 g_1 + \gamma_V g_2\right),\\
H_2 =& \Pi_H^2 B_7\gamma_V\left(\phi\beta_V B_6\left(B_1 - \alpha\omega\right) g_1 + \gamma_V\left(1-\omega\right) g_2\right)\left(1-\omega\right) g_2,\\
H_3 =& \Pi_H\Pi_V\phi^2\beta_V\beta_H B_6\rho g_1 g_3\left(1-\omega\right),\\
H_4 =& \Pi_H^2 B_7\gamma_V\left(1-\omega\right)\left(\phi\beta_V B_6 g_1 + \gamma_V g_2\right) B_1B_2B_3B_4B_5B_6,\\
H_5 =& \Pi_H^2 B_7\gamma_V g_2\left(B_1 - \alpha\omega\right)\left(\phi\beta_V B_6 g_1\left(B_1 - \alpha\omega\right) + \gamma_V g_2\left(1-\omega\right)\right)\\
&+ \Pi_H^2 B_2B_3B_4B_5B_6B_7\gamma_V\left(1-\omega\right)\left(\phi\beta_V B_6 g_1\left(B_1 - \alpha\omega\right) + \gamma_V g_2\left(1-\omega\right)\right),\\
H_6 =& \Pi_H^2\gamma_V^2 B_1B_2B_3B_4B_5B_6B_7 g_2\left(1-\omega\right),\\
H_7 =& \Pi_H\Pi_V\phi^2\beta_V\beta_H B_6\rho g_1 g_4\left(1-\omega\right),\\
H_8 =& \Pi_H\Pi_V\phi^2\beta_V\beta_H B_6\rho g_1 g_3\left(B_1 - \alpha\omega\right),\\
H_9 =& \Pi_H^2 B_7\gamma_V\left(\phi\beta_V B_6 g_1\left(B_1 - \alpha\omega\right) + \gamma_V\left(1-\omega\right) g_2\right) B_1B_2B_3B_4B_5B_6,\\
H_{10} =& \Pi_H^2\gamma_V^2\left(g_2\left(B_1 - \alpha\omega\right) + B_2B_3B_4B_5B_6\left(1-\omega\right)\right) B_1B_2B_3B_4B_5B_6B_7,\\
H_{11} =& \Pi_H\Pi_V\phi^2\beta_V\beta_H B_6\rho g_1\left(B_1B_2B_3B_4B_5B_6\gamma_H\left(1-\omega\right) + g_4\left(B_1 - \alpha\omega\right)\right).\\
g_1 =& \eta\left(B_4B_5 + \sigma\left(B_5 + v\right)\right),\\
g_2 =& B_6\left(B_3B_4B_5 + \eta\left(B_4B_5 + \sigma\left(B_5 + v\right)\right)\right) + \eta\left(\tau_1 B_4B_5 + \sigma\left(\tau_2 B_5 + \tau_3 v\right)\right),\\
g_3 =& \left(1-\omega\right)\left(B_2B_3B_4B_5B_6 - \kappa\eta\left(\tau_1 B_4B_5 + \sigma\left(\tau_2 B_5 + \tau_3 v\right)\right)\right),\\
g_4 =& B_2B_3B_4B_5B_6\left(\left(1-\omega\right) B_1 + \gamma_H\right)
\end{aligned}$$

$$- \kappa\eta \left(B_1 - \alpha\omega\right)\left(\tau_1 B_4 B_5 + \sigma\left(\tau_2 B_5 + \tau_3 v\right)\right).$$

Clearly, in the polynomial (12) $M_1 > 0$ (since all the model parameters are positive) and $M_5 > 0$ whenever $R_0^2 < 1\,(R_0 < 1)$. Therefore, the number of possible positive real roots of the polynomial (12) can be determined depending on the signs of M_2, M_3, and M_4. Using the Descartes' rule of signs [31] on the polynomial $f_{\varepsilon=0}\left(\lambda_H^{**}\right) = M_1\lambda_H^{**4} + M_2\lambda_H^{**3} + M_3\lambda_H^{**2} + M_4\lambda_H^{**} + M_5$, the following result is established.

Theorem 3. The LF model (1)

(a) has a unique endemic equilibrium if $R_0 > 1$ and either of the following conditions holds:

(i) $M_2 > 0$, $M_3 > 0$, $M_4 > 0$,

(ii) $M_2 < 0$, $M_3 < 0$, $M_4 < 0$,

(iii) $M_2 > 0$, $M_3 < 0$, $M_4 < 0$,

(iv) $M_2 > 0$, $M_3 > 0$, $M_4 < 0$;

(b) Could have more than one endemic equilibrium if $R_0 > 1$ and either of the following conditions holds:

(i) $M_2 < 0$, $M_3 > 0$, $M_4 < 0$,

(ii) $M_2 < 0$, $M_3 < 0$, $M_4 > 0$,

(iii) $M_2 > 0$, $M_3 < 0$, $M_4 > 0$,

(iv) $M_2 < 0$, $M_3 > 0$, $M_4 > 0$;

(c) Could have two or more endemic equilibria if $R_0 < 1$ and any, or all of M_2, M_3, and M_4 are negative.

In case (c) of Theorem 3, multiple endemic equilibria exist when the reproduction number (R_0) is less than one (which is a feature of the phenomenon of backward bifurcation). The phenomenon of backward bifurcation is characterized by the co-existence of a stable DFE and a stable endemic equilibrium when the basic reproduction number of the model is less than unity (as observed in numerous epidemiological models, such as [16, 14, 25, 11]). Biologically, the implication of the backward bifurcation is that the necessary

condition for the effective control of LF in the population when the basic reproduction number is less than unity ($R_0 < 1$) is no longer sufficient [26]. The presence of the phenomenon of backward bifurcation in the LF model (1) is investigated in the subsequent section.

3.5 Bifurcation analysis

Theorem 4 (Castillo-Chavez and Song theorem)**.** Let us consider a general system of ordinary differential equations with a parameter φ,

$$\frac{dx}{dt} = f(x, \varphi), f : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n, f \in \mathbf{C}^2(\mathbb{R}^n \times \mathbb{R}), \tag{13}$$

where $x = 0$ is an equilibrium point for the system in (11). That is, $f(0, \varphi) \equiv 0$ for all φ. Assume the following conditions:
H_1: $A = D_x f(0,0) = \left[\frac{\partial f}{\partial x}(0,0)\right]$ is the linearization matrix of the system given by (13) around the equilibrium 0 with φ evaluated at 0. Zero is a simple eigenvalue of A and other eigenvalues of A have negative real parts.
H_2: Matrix A has a nonnegative right eigenvector w and a left eigenvector v corresponding to the zero eigenvalue.
Let f_k be the kth component of f and let

$$\begin{aligned} a &= \sum_{k,i,j=1}^{n} v_k w_i w_j \frac{\partial^2 f_k}{\partial x_i \partial x_j}(0,0), \\ b &= \sum_{k,i=1}^{n} v_k w_i \frac{\partial^2 f_k}{\partial x_i \partial \varphi}(0,0). \end{aligned} \tag{14}$$

The local dynamics of (13) around 0 are totally determined by the sign of a and b.

1. $a > 0$, $b > 0$; When $\varphi < 0$ with $|\varphi| << 1$, 0 is locally asymptotically stable and there exists a positive unstable equilibrium; when $0 < \varphi << 1$, 0 is unstable and there exists a negative, locally asymptotically stable equilibrium.

2. $a < 0$, $b < 0$; When $\varphi < 0$ with $|\varphi| << 1$, 0 is unstable; when $0 < \varphi << 1$, 0 is locally asymptotically stable equilibrium, and there exists a positive unstable equilibrium.

3. $a > 0$, $b < 0$; When $\varphi < 0$ with $|\varphi| << 1$, 0 is unstable, and there exists a locally asymptotically stable negative equilibrium; when $0 < \varphi << 1$, 0 is stable and a positive unstable equilibrium appears.

4. $a < 0$, $b > 0$; When φ changes from negative to positive, 0 changes its stability from stable to unstable. Correspondingly, a negative unstable equilibrium becomes positive and locally asymptotically stable.

In particular, if $a < 0$ and $b > 0$, then the bifurcation is forward; if $a > 0$ and $b > 0$, then the bifurcation is backward. Using this approach, the following result may be obtained.

Theorem 5. The model of system (1) exhibits backward bifurcation at $R_0 = 1$.

Proof. We explore the nature of the bifurcation using the center manifold theory [16, 6]. To apply this theory, it is important to carry out the following change of variables. Let $S_H = x_1$, $S_{HP} = x_2$, $E_H = x_3$, $L = x_4$, $A = x_5$, $C = x_6$, $T = x_7$, $S_V = x_8$, $E_V = x_9$, $I_V = x_{10}$.
So that

$$N_H = \sum_{i=1}^{7} x_i.$$

Furthermore, using the vector notation $x = (x_1, x_2, x_3, \ldots, x_{10})^T$ and $\frac{dx}{dt} = F(x)$, with $F = (f_1, f_2, f_3, \ldots, f_{10})^T$. The LF model (1) can be rewritten as follows:

$$\begin{aligned}
\frac{dx_1}{dt} &\equiv f_1 = \Pi_H - \frac{\phi\beta_H x_{10} x_1}{x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7} - (\alpha + \gamma_H)\, x_1 + \kappa x_7,\\
\frac{dx_2}{dt} &\equiv f_2 = \alpha x_1 - \frac{(1-\omega)\,\phi\beta_H x_{10} x_2}{x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7} - \gamma_H x_2,\\
\frac{dx_3}{dt} &\equiv f_3 = \frac{\phi\beta_H x_{10}\,(x_1 + (1-\omega)\, x_2)}{x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7} - (\eta + \gamma_H)\, x_3,\\
\frac{dx_4}{dt} &\equiv f_4 = \eta x_3 - \frac{\tau_1 x_4}{1 + \varepsilon x_4} - (\sigma + \gamma_H)\, x_4,\\
\frac{dx_5}{dt} &\equiv f_5 = \sigma x_4 - \frac{\tau_2 x_5}{1 + \varepsilon x_5} - (\upsilon + \gamma_H)\, x_5,
\end{aligned}$$

$$\begin{aligned}
\frac{dx_6}{dt} &\equiv f_6 = vx_5 - \frac{\tau_3 x_6}{1+\varepsilon x_6} - \gamma_H x_6,\\
\frac{dx_7}{dt} &\equiv f_7 = \frac{\tau_1 x_4}{1+\varepsilon x_4} + \frac{\tau_2 x_5}{1+\varepsilon x_5} + \frac{\tau_3 x_6}{1+\varepsilon x_6} - (\kappa+\gamma_H)\, x_7,\\
\frac{dx_8}{dt} &\equiv f_8 = \Pi_V - \frac{\phi\beta_V\,(x_4+x_5+x_6)\, x_8}{x_1+x_2+x_3+x_4+x_5+x_6+x_7} - \gamma_V x_8,\\
\frac{dx_9}{dt} &\equiv f_9 = \frac{\phi\beta_V\,(x_4+x_5+x_6)\, x_8}{x_1+x_2+x_3+x_4+x_5+x_6+x_7} - (\rho+\gamma_V)\, x_9,\\
\frac{dx_{10}}{dt} &\equiv f_{10} = \rho x_9 - \gamma_V x_{10}.
\end{aligned} \tag{15}$$

Consider the transmission probability β_H as the bifurcation parameter. Solving for $\beta_H = \beta_H^*$ from $R_0 = 1$ gives

$$\beta_H = \beta_H^* = \frac{\Pi_H \gamma_V B_1 B_2 B_3 B_4 B_5 B_7}{\Pi_V \gamma_H \phi^2 \beta_V\,(\gamma_H + (1-\omega)\,\alpha)\,(\sigma\,(v+B_5)+B_4 B_5)\,\eta\rho}.$$

Evaluating the Jacobian of the transformed system (15) at DFE (ξ_0) with $\beta_H = \beta_H^*$, is given by

$$J(\xi_0)|_{\beta_H=\beta_H^*} = \begin{bmatrix} -B_1 & 0 & 0 & 0 & 0 & 0 & \kappa & 0 & 0 & -\frac{\phi\beta_H^*\gamma_H}{B_1} \\ \alpha & -\gamma_H & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{\phi\beta_H^* D_1}{B_1} \\ 0 & 0 & -B_2 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{\phi\beta_H^* D_2}{B_1} \\ 0 & 0 & \eta & -B_3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma & -B_4 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \nu & -B_5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \tau_1 & \tau_2 & \tau_3 & -B_6 & 0 & 0 & 0 \\ 0 & 0 & 0 & -D_3 & -D_3 & -D_3 & 0 & -\gamma_V & 0 & 0 \\ 0 & 0 & 0 & D_3 & D_3 & D_3 & 0 & 0 & -B_7 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \rho & -\gamma_V \end{bmatrix},$$

where

$B_1 = \alpha + \gamma_H$, $B_2 = \eta + \gamma_H$, $B_3 = \sigma + \tau_1 + \gamma_H$, $B_4 = \upsilon + \tau_2 + \gamma_H$, $B_5 = \tau_3 + \gamma_H$, $B_6 = \kappa + \gamma_H$, $B_7 = \rho + \gamma_V$, $D_1 = (1-\omega)\,\alpha$, $D_2 = \gamma_H + (1-\omega)\,\alpha$, and $D_3 = \frac{\phi\beta_V \Pi_V \gamma_H}{\Pi_H \gamma_V}$.

The right eigenvector, $w = (w_1, w_2, w_3, w_4, w_5, w_6, w_7, w_8, w_9, w_{10})^T$, associated with the simple zero eigenvalue can be obtained from $J(\xi_0)|_{\beta_H=\beta_H^*} w = 0$, given by

$$\begin{aligned} &-B_1 w_1 + \kappa w_7 - \frac{\phi\beta_H^*\gamma_H}{B_1} w_{10} = 0, \\ &\alpha w_1 - \gamma_H w_2 - \frac{\phi\beta_H^* D_1}{B_1} w_{10} = 0, \\ &-B_2 w_3 + \frac{\phi\beta_H^* D_2}{B_1} w_{10} = 0, \\ &\eta w_3 - B_3 w_4 = 0, \\ &\sigma w_4 - B_4 w_5 = 0, \end{aligned}$$

$$
\begin{aligned}
&vw_5 - B_5w_6 = 0,\\
&\tau_1w_4 + \tau_2w_5 + \tau_3w_6 - B_6w_7 = 0,\\
&-\frac{\phi\beta_V\Pi_V\gamma_H}{\Pi_H\gamma_V}w_4 - \frac{\phi\beta_V\Pi_V\gamma_H}{\Pi_H\gamma_V}w_5 - \frac{\phi\beta_V\Pi_V\gamma_H}{\Pi_H\gamma_V}w_6 - \gamma_Vw_8 = 0,\\
&\frac{\phi\beta_V\Pi_V\gamma_H}{\Pi_H\gamma_V}w_4 + \frac{\phi\beta_V\Pi_V\gamma_H}{\Pi_H\gamma_V}w_5 + \frac{\phi\beta_V\Pi_V\gamma_H}{\Pi_H\gamma_V}w_6 - B_7w_9 = 0,\\
&\rho w_9 - \gamma_Vw_{10} = 0.
\end{aligned} \tag{16}
$$

From (16), we obtained

$$
\begin{aligned}
w_1 &= \frac{(\kappa\eta D_2(\tau_1B_4B_5 + \tau_2\sigma B_5 + \tau_3v\sigma) - \gamma_HB_2B_3B_4B_5)}{B_1B_3B_4B_5D_2}w_3,\\
w_2 &= \frac{(\alpha\kappa\eta D_2(\tau_1B_4B_5 + \tau_2\sigma B_5 + \tau_3v\sigma) - B_2B_3B_4B_5(\alpha\gamma_H + B_1D_1))}{\gamma_HB_1B_3B_4B_5D_2}w_3,\\
w_3 &= w_3 > 0, \quad w_4 = \frac{\eta}{B_3}w_3, \quad w_5 = \frac{\eta\sigma}{B_3B_4}w_3, \quad w_6 = \frac{\eta v\sigma}{B_3B_4B_5}w_3,\\
w_7 &= \frac{\eta(\tau_1B_4B_5 + \tau_2\sigma B_5 + \tau_3v\sigma)}{B_3B_4B_5}w_3,\\
w_8 &= -\frac{\phi\beta_V\Pi_V\gamma_H\eta(B_4B_5 + \sigma B_5 + v\sigma)}{\Pi_H\gamma_V^2B_3B_4B_5}w_3, w_9 = \frac{\gamma_VB_1B_2}{\phi\beta_H^*D_2\rho}w_3,\\
w_{10} &= \frac{B_1B_2}{\phi\beta_H^*D_2}w_3.
\end{aligned}
$$

Similarly, the left eigenvector, $\nu = (\nu_1, \nu_2, \nu_3, \nu_4, \nu_5, \nu_6, \nu_7, \nu_8, \nu_9, \nu_{10})$, satisfying $\nu.w = 1$, associated with the simple zero eigenvalue can be obtained from $\nu J(\xi_0)|_{\beta_H=\beta_H^*} = 0$, given by

$$
\begin{aligned}
&-B_1\nu_1+\alpha\nu_2=0,\\
&-\gamma_H\nu_2=0,\\
&-B_2\nu_3+\eta\nu_4=0,\\
&-B_3\nu_4+\sigma\nu_5+\tau_1\nu_7-\frac{\phi\beta_V\Pi_V\gamma_H}{\Pi_H\gamma_V}\nu_8+\frac{\phi\beta_V\Pi_V\gamma_H}{\Pi_H\gamma_V}\nu_9=0,\\
&-B_4\nu_5+\upsilon\nu_6+\tau_2\nu_7-\frac{\phi\beta_V\Pi_V\gamma_H}{\Pi_H\gamma_V}\nu_8+\frac{\phi\beta_V\Pi_V\gamma_H}{\Pi_H\gamma_V}\nu_9=0,\\
&-B_5\nu_6+\tau_3\nu_7-\frac{\phi\beta_V\Pi_V\gamma_H}{\Pi_H\gamma_V}\nu_8+\frac{\phi\beta_V\Pi_V\gamma_H}{\Pi_H\gamma_V}\nu_9=0,\\
&\kappa\nu_1-B_6\nu_7=0,\\
&-\gamma_V\nu_8=0,\\
&-B_7\nu_9+\rho\nu_{10}=0,\\
&-\frac{\phi\beta_H^*\gamma_H}{B_1}\nu_1-\frac{\phi\beta_H^*D_1}{B_1}\nu_2+\frac{\phi\beta_H^*D_2}{B_1}\nu_3-\gamma_V\nu_{10}=0.
\end{aligned}
\tag{17}
$$

From (17), we obtained

$$
\begin{aligned}
&\nu_1=\nu_2=\nu_7=\nu_8=0,\quad \nu_3=\nu_3>0,\quad \nu_4=\frac{B_2}{\eta}\nu_3,\\
&\nu_5=\frac{\Pi_H\gamma_V B_2B_3\nu_3-\phi\beta_V\Pi_V\gamma_H\eta\nu_9}{\Pi_H\gamma_V\eta\sigma},\\
&\nu_6=\frac{\phi\beta_V\Pi_V\gamma_H}{\Pi_H\gamma_V B_5}\nu_9,\nu_9=\nu_9>0,\nu_{10}=\frac{B_7}{\rho}\nu_9.
\end{aligned}
$$

Computation of a and b

Since $\nu_1=\nu_2=\nu_7=\nu_8=0$ for $k=1,2,3,\ldots,10$, the only nonzero partial derivatives are

$$
\begin{aligned}
&\frac{\partial^2 f_3}{\partial x_1\partial x_{10}}=\frac{\partial^2 f_3}{\partial x_{10}\partial x_1}=\frac{\phi\beta_H^*\gamma_H\left(B_1-\gamma_H-(1-\omega)\,\alpha\right)}{\Pi_H B_1},\\
&\frac{\partial^2 f_3}{\partial x_2\partial x_{10}}=\frac{\partial^2 f_3}{\partial x_{10}\partial x_2}=\frac{\phi\beta_H^*\gamma_H\left((1-\omega)\,(B_1-\alpha)-\gamma_H\right)}{\Pi_H B_1},\\
&\frac{\partial^2 f_3}{\partial x_3\partial x_{10}}=\frac{\partial^2 f_3}{\partial x_{10}\partial x_3}=\frac{\partial^2 f_3}{\partial x_4\partial x_{10}}=\frac{\partial^2 f_3}{\partial x_{10}\partial x_4}=\frac{\partial^2 f_3}{\partial x_5\partial x_{10}}=\frac{\partial^2 f_3}{\partial x_{10}\partial x_5}\\
&\qquad=\frac{\partial^2 f_3}{\partial x_6\partial x_{10}}=\frac{\partial^2 f_3}{\partial x_{10}\partial x_6}=\frac{\partial^2 f_3}{\partial x_7\partial x_{10}}=\frac{\partial^2 f_3}{\partial x_{10}\partial x_7}\\
&\qquad=-\frac{\phi\beta_H^*\gamma_H\left(\gamma_H-(1-\omega)\,\alpha\right)}{\Pi_H B_1},\\
&\frac{\partial^2 f_9}{\partial x_1\partial x_4}=\frac{\partial^2 f_9}{\partial x_4\partial x_1}=\frac{\partial^2 f_9}{\partial x_2\partial x_4}=\frac{\partial^2 f_9}{\partial x_4\partial x_2}=\frac{\partial^2 f_9}{\partial x_3\partial x_4}=\frac{\partial^2 f_9}{\partial x_4\partial x_3}=\frac{\partial^2 f_9}{\partial x_4\partial x_7}
\end{aligned}
\tag{18}
$$

$$
\begin{aligned}
&= \frac{\partial^2 f_9}{\partial x_7 \partial x_4} = -\frac{\phi\beta_V \Pi_V \gamma_H^2}{\Pi_H^2 \gamma_V},\\
\frac{\partial^2 f_9}{\partial x_4^2} &= \frac{\partial^2 f_9}{\partial x_4 \partial x_5} = \frac{\partial^2 f_9}{\partial x_5 \partial x_4} = \frac{\partial^2 f_9}{\partial x_4 \partial x_6} = \frac{\partial^2 f_9}{\partial x_6 \partial x_4} = -\frac{2\phi\beta_V \Pi_V \gamma_H^2}{\Pi_H^2 \gamma_V},\\
\frac{\partial^2 f_9}{\partial x_1 \partial x_5} &= \frac{\partial^2 f_9}{\partial x_5 \partial x_1} = \frac{\partial^2 f_9}{\partial x_2 \partial x_5} = \frac{\partial^2 f_9}{\partial x_5 \partial x_2} = \frac{\partial^2 f_9}{\partial x_3 \partial x_5} = \frac{\partial^2 f_9}{\partial x_5 \partial x_3} = \frac{\partial^2 f_9}{\partial x_5 \partial x_7}\\
&= \frac{\partial^2 f_9}{\partial x_7 \partial x_5} = -\frac{\phi\beta_V \Pi_V \gamma_H^2}{\Pi_H^2 \gamma_V}, \frac{\partial^2 f_9}{\partial x_5^2} = \frac{\partial^2 f_9}{\partial x_6^2} = \frac{\partial^2 f_9}{\partial x_5 \partial x_6}\\
&= \frac{\partial^2 f_9}{\partial x_6 \partial x_5} = -\frac{2\phi\beta_V \Pi_V \gamma_H^2}{\Pi_H^2 \gamma_V},\\
\frac{\partial^2 f_9}{\partial x_1 \partial x_6} &= \frac{\partial^2 f_9}{\partial x_6 \partial x_1} = \frac{\partial^2 f_9}{\partial x_2 \partial x_6} = \frac{\partial^2 f_9}{\partial x_6 \partial x_2} = \frac{\partial^2 f_9}{\partial x_3 \partial x_6} = \frac{\partial^2 f_9}{\partial x_6 \partial x_3} = \frac{\partial^2 f_9}{\partial x_6 \partial x_7}\\
&= \frac{\partial^2 f_9}{\partial x_7 \partial x_6} = -\frac{\phi\beta_V \Pi_V \gamma_H^2}{\Pi_H^2 \gamma_V},\\
\frac{\partial^2 f_9}{\partial x_4 \partial x_8} &= \frac{\partial^2 f_9}{\partial x_8 \partial x_4} = \frac{\partial^2 f_9}{\partial x_5 \partial x_8} = \frac{\partial^2 f_9}{\partial x_8 \partial x_5} = \frac{\partial^2 f_9}{\partial x_6 \partial x_8} = \frac{\partial^2 f_9}{\partial x_8 \partial x_6} = \frac{\partial^2 f_9}{\partial x_6 \partial x_7}\\
&= \frac{\partial^2 f_9}{\partial x_7 \partial x_6} = \frac{\phi\beta_V \gamma_H}{\Pi_H}, \quad \frac{\partial^2 f_3}{\partial x_{10} \partial \beta_H^*} = \frac{\phi\left(\gamma_H + (1-\omega)\,\alpha\right)}{B_1},
\end{aligned} \tag{19}
$$

but

$$
a = \sum_{k,i,j=1}^{n} v_k w_i w_j \frac{\partial^2 f_k}{\partial x_i \partial x_j}\left(0,0\right),
$$

$$
\begin{aligned}
a = &-\frac{2\nu_3 w_3^2 \gamma_H \left(\gamma_H - (1-\omega)\,\alpha\right) B_2 \Theta_1}{\Pi_H B_3 B_4 B_5 D_2}\\
&+\frac{2\nu_3 w_3^2 \gamma_H \left(B_1 - \gamma_H - (1-\omega)\,\alpha\right) B_2 \Theta_2}{\Pi_H B_1 B_3 B_4 B_5 D_2^2}\\
&+\frac{2\nu_3 w_3^2 \left((1-\omega)\left(B_1 - \alpha\right) - \gamma_H\right) B_2 \Theta_3}{\Pi_H B_1 B_3 B_4 B_5 D_2^2}\\
&-\frac{2\nu_9 w_3^2 \phi\beta_V \Pi_V \gamma_H^2 \eta \left(B_4 B_5 + \sigma B_5 + v\sigma\right) \Theta_1}{\Pi_H^2 \gamma_V B_3^2 B_4^2 B_5^2}\\
&-\frac{2\nu_9 w_3^2 \phi\beta_V \Pi_V \gamma_H^2 \eta \left(B_4 B_5 + \sigma B_5 + v\sigma\right) \Theta_2}{\Pi_H^2 \gamma_V B_1 B_3^2 B_4^2 B_5^2 D_2}\\
&-\frac{2\nu_9 w_3^2 \phi\beta_V \Pi_V \gamma_H \eta \left(B_4 B_5 + \sigma B_5 + v\sigma\right) \Theta_3}{\Pi_H^2 \gamma_V B_1 B_3^2 B_4^2 B_5^2 D_2}\\
&-\frac{2\nu_9 w_3^2 \phi\beta_V \Pi_V \gamma_H^2 \eta^2 \left(B_4 B_5 + \sigma B_5 + v\sigma\right)^2}{\Pi_H^2 \gamma_V^2 B_3^2 B_4^2 B_5^2},
\end{aligned} \tag{20}
$$

where

$$\Theta_1 = B_3B_4B_5 + \eta\left(B_4B_5\left(1+\tau_1\right) + \sigma B_5\left(1+\tau_2\right) + v\sigma\left(1+\tau_3\right)\right),$$
$$\Theta_2 = \kappa\eta D_2\left(\tau_1 B_4B_5 + \tau_2\sigma B_5 + \tau_3 v\sigma\right) - \gamma_H B_2B_3B_4B_5,$$
$$\Theta_3 = \alpha\kappa\eta D_2\left(\tau_1 B_4B_5 + \tau_2\sigma B_5 + \tau_3 v\sigma\right) - B_2B_3B_4B_5\left(\alpha\gamma_H + B_1D_1\right).$$

Similarly, we calculate for b, given by

$$b = \sum_{k,i=1}^{n} v_k w_i \frac{\partial^2 f_k}{\partial x_i \partial \beta^*}\left(0,0\right),$$

$$b = \frac{\nu_3 w_3 B_2}{\beta_H^*} > 0, \tag{21}$$

Hence, since the bifurcation coefficient b is positive and the sign of the coefficient of a is positive, it follows, from Theorem 5 that the LF model exhibits a backward bifurcation at $R_0 = 1$. Figure 2 shows the backward bifurcation diagram of the LF model (1). □

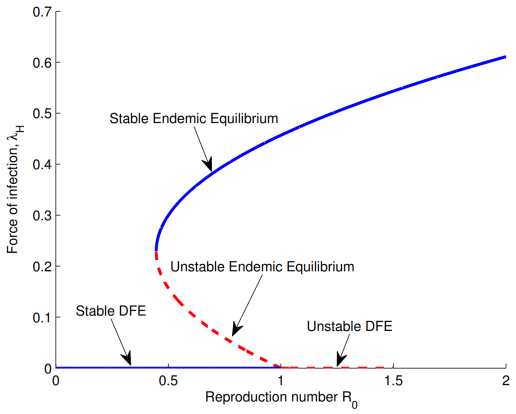


Figure 2: Backward bifurcation diagram of the LF model (1), showing the profile of the force of infection $\left(\lambda_H^{**}\right)$ as a function of the reproduction number (R_0) The parameter values used are as in Table 2, (so that $a = 1.4006 \times 10^{-4}$).

4 Numerical simulations

In this section, we carry out the numerical simulations of the LF model (1) using parameter values in Table 2, so as to illustrate some of the analytic results obtained in this study.

Table 2: Parameters value of the LF model (1)

Paramter	**Value**	**Source**
Π_H	0.01747	[4]
Π_V	100	[16]
ϕ	0.45	[4]
β_H	0.01	[28]
β_V	0.1	[28]
γ_H	0.000039	[21]
γ_V	0.1439	[21]
α	0.1	Assumed
σ	0.45	Assumed
ω	0.2	Assumed
η	0.00238	[21]
ν	0.48	Assumed
τ_1	0.2	Assumed
τ_2	0.3	Assumed
τ_3	0.5	Assumed
ε	0.5	Assumed
κ	0.135	Assumed
ρ	0.0555	[21]

4.1 Discussion

In Figure 3, each surface represents the value of the basic reproduction number as a function of two parameters. In Figure 3e, it is shown that the basic reproduction number increases very fast as the mosquito-biting rate

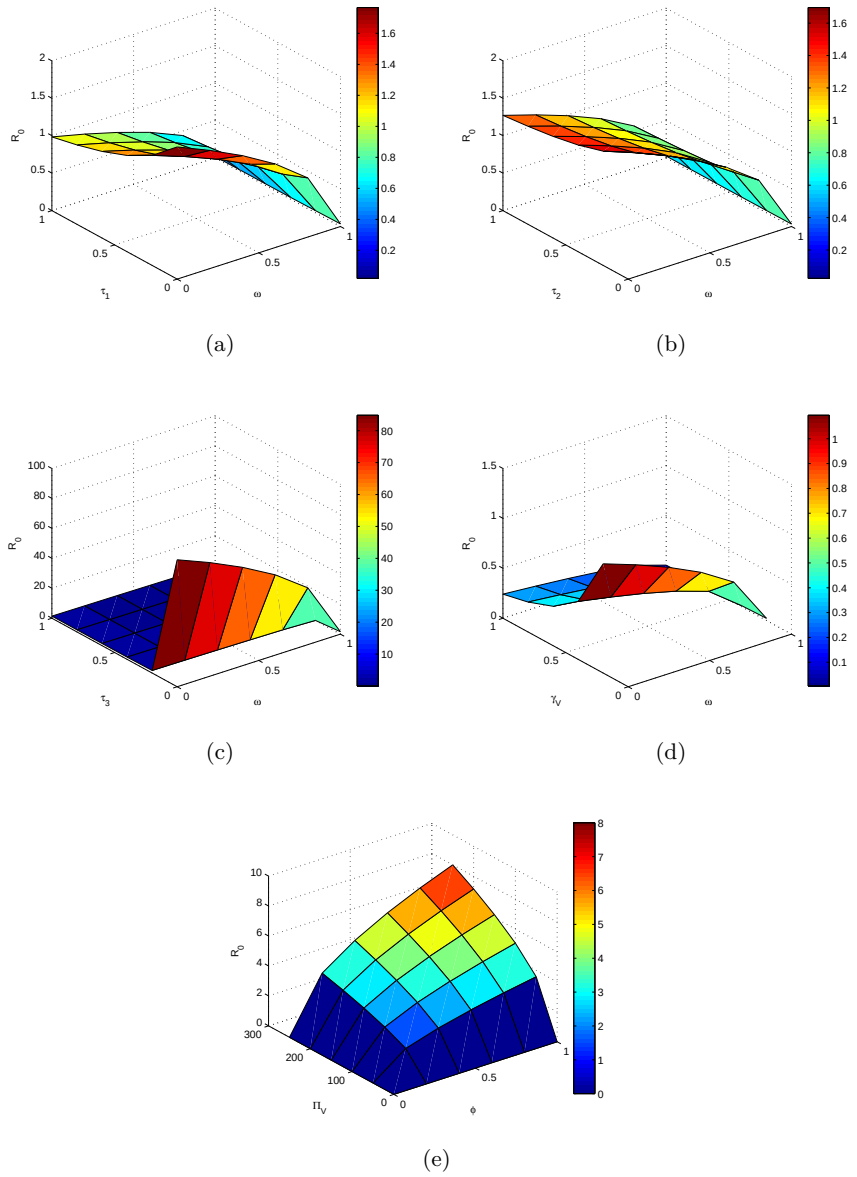
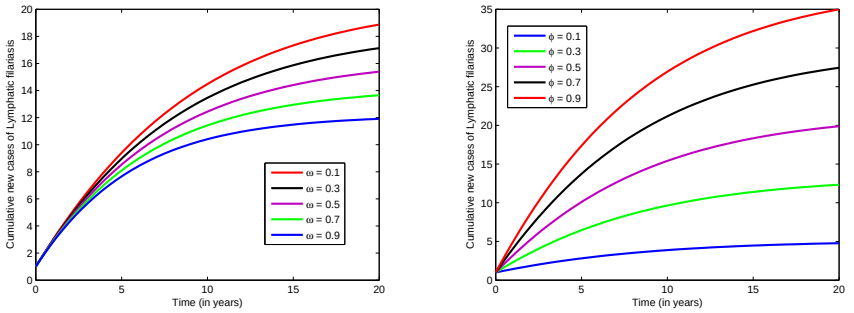


Figure 3: Surface plots of the basic reproduction number with respect to (a) ω and τ_1 (b) ω and τ_2 (c) ω and τ_3 (d) ω and γ_V (e) ϕ and Π_V.

and vector recruitment rate increase. In Figure 3c, the surface of the basic reproduction number diminishes as the value of the treatment rate for

Chronic stage infected individuals and the efficacy of the prophylaxis drugs increases. It is observed in Figure 3a and 3b that the surface of the basic reproduction number increases as the treatment rates and the efficacy of prophylaxis drugs decrease. In Figure 3d, the surface of the basic reproduction number decreases as the vector death rate and efficacy of prophylaxis drugs increases.



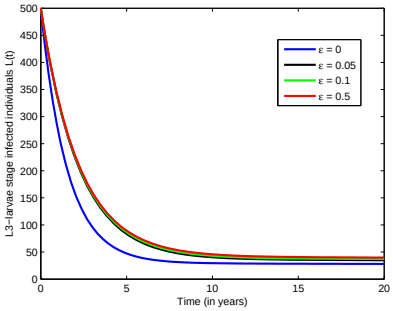
(a) The effect of ω on the Cumulative new cases of LF.

(b) The effect of ϕ on the Cumulative new cases of LF.

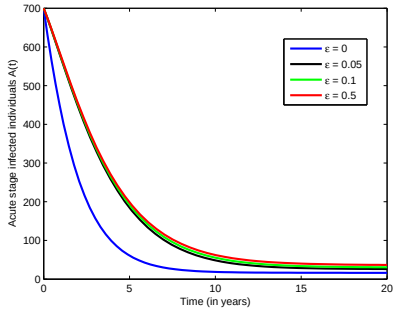
Figure 4: Effect of ω and ϕ on the Cumulative new cases of LF

Figure 4a depicts the simulation of the effect of efficacy of prophylaxis drugs (ω) on the cumulative new cases of LF. It is observed that as the rate of the efficacy of prophylaxis drugs (ω) increases, there is a decrease in the number of the cumulative new cases of LF. Figure 4b is the simulation of the effect of the mosquito-biting rate (ϕ) on the cumulative new cases of LF, it is shown that as the mosquito-biting rate (ϕ) increases, the number of the cumulative new cases of LF increases as well.

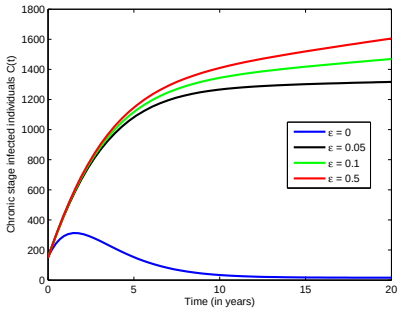
Figure 5 depicts the simulations of the effect of the limitation rate in medical resources availability (ε) on the infected classes respectively while the treatment rates are constant. It is observed that the limitation rate in medical resources availability has a significant impact on the increase and decrease of infected individuals, especially in the chronic stage of infection. In Figure 5a, when $\varepsilon = 0$, the number of L3-larvae stage infected individuals was low compared to when $\varepsilon = 0.05$, $\varepsilon = 0.1$, and $\varepsilon = 0.5$, we observed a similar result in Figure 5b. In Figure 5c, when $\varepsilon = 0$, the chronic stage



(a) The effect of ε on L3-larvae stage infected individuals.



(b) The effect of ε on Acute stage infected individuals.

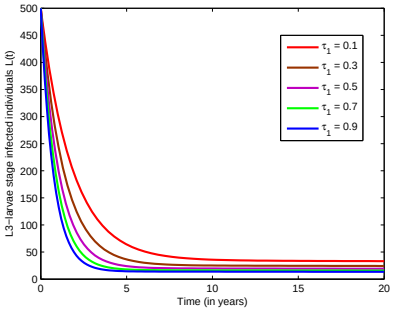


(c) The effect of ε on Chronic stage infected individuals.

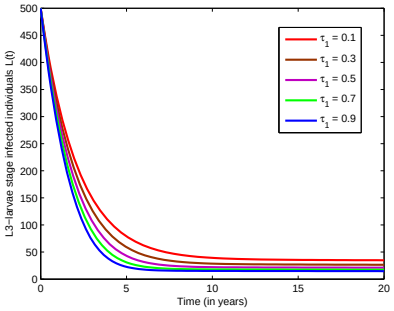
Figure 5: Effect of ε on the infected compartments.

infected individuals reduce from 200 to almost less than 10 in eight years, but when $\varepsilon = 0.05$, $\varepsilon = 0.1$, and $\varepsilon = 0.5$, it was observed that the chronic stage infected individuals increase tremendously.

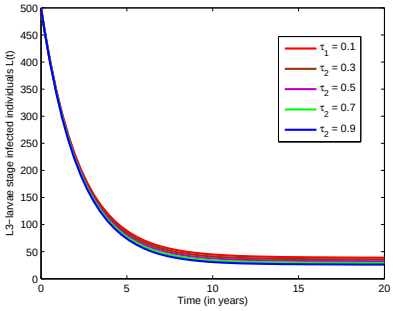
In Figure 6, we simulate the effect of the treatment rate (τ_1) on the L3-larvae stage infected individuals with various values for the limitation rate in medical resources availability (ε). In Figure 6a, it is observed that when $\varepsilon = 0$, and the treatment rate is varied (τ_1) to 0.9 the number of L3-larvae stage infected individuals reduced from 500 to less than 50 in 2 years. Whereas, in Figure 6b, when $\varepsilon = 0.01$, it took about 3.2 years for the



(a) When $\varepsilon = 0$.



(b) When $\varepsilon = 0.01$.

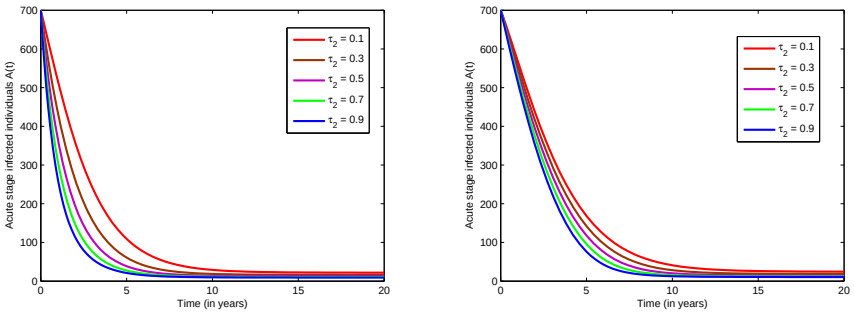


(c) When $\varepsilon = 0.1$.

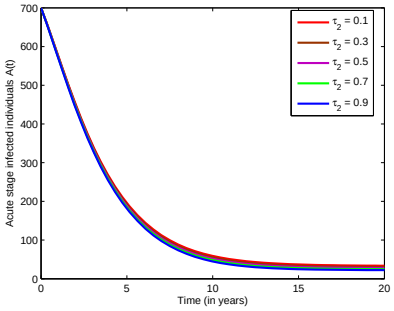
Figure 6: The effect of the treatment rate (τ_1) on the L3-larvae stage infected individuals.

infected individuals to be less than 50. In Figure 6c it took about 6 years for the number of infected individuals to be less than 50, when $\varepsilon = 0.1$.

Similarly, Figure 7 depicts simulations of the effect of treatment rate (τ_2) on the Acute stage infected individuals with various values for the limitation rate in medical resources availability (ε). It is observed in Figure 7a that when $\varepsilon = 0$, and the treatment rate (τ_2) is varied to 0.9 the number of Acute stage infected individuals reduced from 700 to less than 100 in 2 years. Whereas, in Figure 7b, when $\varepsilon = 0.01$, it took about 4 years for the infected individuals to be less than 100. In Figure 7c it took about 6.5 years for the number of infected individuals to be less than 100, when $\varepsilon = 0.1$. Figure 8 depicts simulations of the effect of treatment rate (τ_3) on the Chronic stage infected individuals with various values for the limitation rate in medical resources



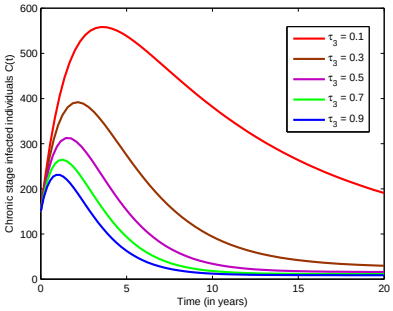
(a) When $\varepsilon = 0$. (b) When $\varepsilon = 0.01$.



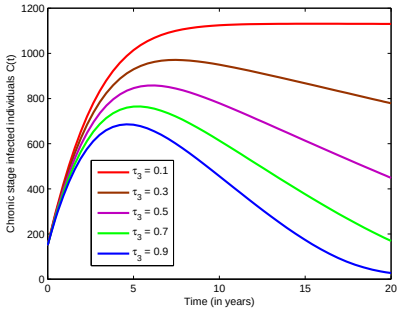
(c) When $\varepsilon = 0.1$.

Figure 7: The effect of the treatment rate (τ_2) on the Acute stage infected individuals.

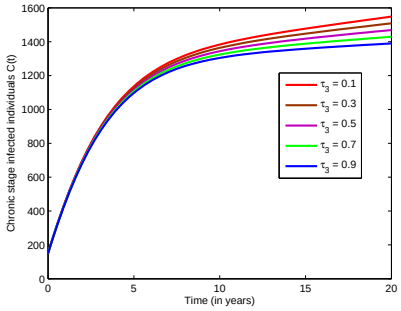
availability (ε). In Figure 8a, it is observed that when the treatment rate (τ_3) is varied to 0.9 and $\varepsilon = 0$, the number of chronic stage infected individual reduces from 200 to zero within 10 years, but in Figure 8b it took up to 20 years to reduce the number of infected to be less than 10. It is also observed that in Figure 8a when $\tau_3 = 0.1$ and $\varepsilon = 0$, the number of infected individuals is lower as compared to Figure 8b where $\tau_3 = 0.1$ and $\varepsilon = 0.01$. In Figure 8c, it is observed that when $\varepsilon = 0.1$, the treatment rate has minimal effect in reducing the number of infected individuals, and even if the treatment rate is increased to 0.9 the disease will persist and the number of chronic stage infected individuals will keep on increasing.



(a) When $\varepsilon = 0$.



(b) When $\varepsilon = 0.01$.



(c) When $\varepsilon = 0.1$.

Figure 8: The effect of the treatment rate (τ_3) on the Chronic stage infected individuals.

5 Conclusions

In this paper, a deterministic mathematical model with Holling type II treatment functions is presented and rigorously analyzed in order to understand the dynamics of LF. Qualitative analysis of the model shows that the model has a locally asymptotically stable DFE when the basic reproduction number (R_0) is less than unity. The study also uncovers that the LF model exhibits the phenomenon of backward bifurcation, characterized by the coexistence of a stable DFE and a stable endemic equilibrium when the basic reproduction number of the model is less than one. From a biological standpoint, the implication of backward bifurcation is that the necessary condition for effectively controlling LF in the population is no longer sufficient when the basic re-

production number is less than one. The numerical simulations demonstrate that increasing treatment rates, availability of medical resources, and efficacy of prophylaxis drugs significantly reduce the burden of LF. It is recommended that policy makers in the healthcare sector should increase the rate of MDA and availability of medical resources as well as mosquito bed nets in rural areas and regions at risk of LF. Sensitization should also be carried out on the importance of testing and sanitation. In future research, there is potential to reformulate the model and include the diurnal periodicity of microfilariae, thus incorporating the periodic nature of disease transmission pathways. Additionally, extending the model to introduce time-dependent optimal control strategies could be explored, allowing for a more nuanced understanding of disease management. Furthermore, investigating the co-dynamics of LF with other infectious diseases would provide valuable insights into the interplay between multiple diseases and their control measures.

Abbreviations

LF Lymphatic filariasis
DFE Disease-Free Equilibrium
R_0 Basic Reproduction Number

Ethics approval and consent to participate

No ethical approval was required for this manuscript.

Consent for publication

Not applicable

Availability of data and material

Not applicable

Competing interests

The authors declare that they have no competing interests, financial or non-financial, that could be perceived as influencing the content or conclusions of this paper.

Funding

No funding received for this work

Authors' contributions

All the authors contributed equally, and approved the final draft of the manuscript

Acknowledgements

Not applicable

References

[1] Al-Tameemi, K. and Kabakli, R.A.I.A.N. *Lymphatic filariasis: an overview*, Asian J. Pharm. Clin. Res., 12 (2019), 1–5.

[2] Bakowski, M.A. and McNamara, C.W. *Advances in antiwolbachial drug discovery for treatment of parasitic filarial worm infections*, Trop. Med. Infect. Dis., 4 (2019), 108.

[3] Bhunu, C.P. and Mushayabasa, S. *Transmission dynamics of lymphatic filariasis: a mathematical approach*, Int. Scholarly Res. Notices, 2012 (2012), 930130.

[4] Biswas, S.K., Ghosh, U. and Sarkar, S. *Mathematical model of Zika virus dynamics with vector control and sensitivity analysis*, Infect. Dis. Model., 5 (2020), 23–41.

[5] Cano, J., Rebollo, M.P., Golding, N., Pullan, R.L., Crellen, T., Soler, A., Hope, L.A.K., Lindsay, S.W., Hay, S.I., Bockarie, M.J. and Brooker, S.J. *The global distribution and transmission limits of lymphatic filariasis: past and present*, Parasites Vectors, 7 (2014), 1–19.

[6] Castillo-Chavez, C. and Song, B. *Dynamical models of tuberculosis and their applications*, Math. Biosci. Eng., 1 (2004), 361–404.

[7] Chandy, A., Thakur, A.S., Singh, M.P. and Manigauha, A. *A review of neglected tropical diseases: filariasis*, Asian Pac. J. Trop. Med., 4 (2011), 581–586.

[8] Das, P.K., Pani, S.P. and Krishnamoorthy, K.*Prospects of elimination of lymphatic filariasis in India*, ICMR Bull., 32 (2002), 41–54.

[9] Diekmann, O. and Heesterbeek, J.A.P. *Mathematical epidemiology of infectious diseases: model building, analysis and interpretation*, Vol. 5, John Wiley & Sons, 2000.

[10] Fimbo, A.M., Minzi, O.M., Mmbando, B.P., Barry, A., Nkayamba, A.F., Mwamwitwa, K.W., Malishee, A., Seth, M.D., Makunde, W.H., Gurumurthy, P. and Lusingu, J.P. *Prevalence and correlates of lymphatic filariasis infection and its morbidity following mass ivermectin and albendazole administration in Mkinga District, North-Eastern Tanzania*, J. Clin. Med., 9 (2020), 1550.

[11] Garba, S.M., Gumel, A.B. and Bakar, M.A. *Backward bifurcations in dengue transmission dynamics*, Math. Biosci., 215 (2008), 11–25.

[12] Gordon, C.A., Jones, M.K. and McManus, D.P. *The history of Bancroftian lymphatic filariasis in Australasia and Oceania: is there a threat of re-occurrence in mainland Australia?*, Trop. Med. Infect. Dis., 3 (2018), 58.

[13] Gravis, M. *National Organization for Rare Disorders (NORD)*, Available here, accessed March 2022.

[14] Gumel, A.B. *Causes of backward bifurcations in some epidemiological models*, J. Math. Anal. Appl., 395 (2012), 355–365.

[15] Hoerauf, A., Pfarr, K., Mand, S., Debrah, A. Y. and Specht, S. *Filariasis in Africa—treatment challenges and prospects*, Clin. Microbiol. Infect., 17 (2011), 977–985.

[16] Jose, S.A., Raja, R., Omede, B.I., Agarwal, R.P., Alzabut, J., Cao, J. and Balas, V.E. *Mathematical modeling on co-infection: transmission dynamics of Zika virus and Dengue fever*, Nonlinear Dyn., 111 (2023), 4879–4914.

[17] Kumar, A. and Nilam *Mathematical analysis of a delayed epidemic model with nonlinear incidence and treatment rates*, J. Eng. Math., 115 (2019), 1–20.

[18] Lakshmikantham, V., Leela, S. and Martynyuk, A.A. *Stability analysis of nonlinear systems*, M. Dekker, New York, (1989) 249–275.

[19] Medeiros, Z.M., Vieira, A.V., Xavier, A.T., Bezerra, G.S., Lopes, M.D.F.C., Bonfim, C.V. and Aguiar-Santos, A.M. *Lymphatic filariasis: a systematic review on morbidity and its repercussions in countries in the Americas*, Int. J. Environ. Res. Public Health, 19 (2021), 316.

[20] Melrose, W.D. *Lymphatic filariasis: new insights into an old disease*, Int. J. Parasitol., 32 (2002), 947–960.

[21] Mwamtobe, P.M., Simelane, S.M., Abelman, S. and Tchuenche, J.M. *Mathematical analysis of a lymphatic filariasis model with quarantine and treatment*, BMC Public Health, 17 (2017), 1–13.

[22] Negasa, A. and Dufera, M. *Assessment of etiology of elephantiasis and its associated risk factors in Jeldu district, West Shoa, Ethiopia*, J. Trop. Med., (2021) 5551637.

[23] Oguntolu, F.A., Peter, O.J., Yusuf, A., Omede, B.I., Bolarin, G. and Ayoola, T.A. *Mathematical model and analysis of the soil-transmitted helminth infections with optimal control*, Model. Earth Syst. Environ., 10 (2024), 883–897.

[24] Oguntolu, F.A., Yavalah, D.W., Udom, C.F., Ayoola, T.A. and Victor, A.A. *A mathematical modelling of lymphatic filariasis and malaria co-infection*, 2022.

[25] Omede, B.I., Jose, S.A., Anuwat, J. and Park, T. *Mathematical analysis on the transmission dynamics of delta and omicron variants of COVID-19 in the United States*, Model. Earth Syst. Environ., (2024) 1–38.

[26] Omede, B.I., Odionyenma, U.B., Ibrahim, A.A. and Bolaji, B. *Third wave of COVID-19: mathematical model with optimal control strategy for reducing the disease burden in Nigeria*, Int. J. Dyn. Control, 11 (2023), 411–427.

[27] Stocks, M.E., Freeman, M.C. and Addiss, D.G. *The effect of hygiene-based lymphedema management in lymphatic filariasis-endemic areas: a systematic review and meta-analysis*, PLoS Negl. Trop. Dis., 9 (2015), e0004171.

[28] Supriatna, A.K. and Anggriani, N. *Lymphatic filariasis transmission and control: a mathematical modelling approach*, in Current Tropics in Tropical Medicine, (2012) 425–442.

[29] Van den Driessche, P. and Watmough, J. *Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission*, Math. Biosci., 180 (2002), 29–48.

[30] Wang, W. and Ruan, S. *Bifurcations in an epidemic model with constant removal rate of the infectives*, J. Math. Anal. Appl., 291 (2004), 775–793.

[31] Wang, X. *A simple proof of Descartes's rule of signs*, Am. Math. Mon., 111 (2004), 525–526.

[32] World Health Organization. *Preventive chemotherapy in human helminthiasis, Geneva: World Health Organization*, 2006 [online].

[33] World Health Organization. *Monitoring and epidemiological assessment of mass drug administration in the global programme to eliminate lymphatic filariasis: a manual for national elimination programmes*, World Health Organization, 2011.

[34] World Health Organization. *Global programme to eliminate lymphatic filariasis: progress report, 2011*, Wkly. Epidemiol. Rec., 87 (2012), 346–356.

[35] World Health Organization. *Lymphatic filariasis: a handbook of practical entomology for national lymphatic filariasis elimination programmes (No. WHO/HTM/NTD/PCT/2013.10)*, World Health Organization, 2013.

[36] World Health Organization. *Ending the neglect to attain the Sustainable Development Goals: a road map for neglected tropical diseases 2021–2030*, World Health Organization, 2020.

[37] Zhou, L. and Fan, M. *Dynamics of an SIR epidemic model with limited medical resources revisited*, Nonlinear Anal. Real World Appl., 13 (2012), 312–324.

Iranian Journal of Numerical Analysis and Optimization
Vol. 15, No. 2, 2025, pp 570–599
https://doi.org/10.22067/ijnao.2025.90287.1538
https://ijnao.um.ac.ir/

Research Article

Bilinear optimal control of a reaction-diffusion equation: overcoming boundedness constraints on controls

R. El Mezegueldy*, Y. Ouakrim and M. Ouzahra

Abstract

We study an optimal control problem for bilinear reaction-diffusion equations. The novelty of our approach lies in considering controls from the space of essentially bounded functions without imposing a priori bounds on the admissible set. We introduce an auxiliary problem with controls in L^p spaces ($p > 1 + N/2$, where N is the spatial dimension) and demonstrate that both formulations are equivalent in terms of optimal solutions.

*Corresponding author

Received 10 October 2024; revised 8 Junurary 2025; accepted 12 Junurary 2025

Redouane El Mezegueldy
Laboratory MMPA, Department of Mathematics and Informatics, ENS, University of Sidi Mohamed Ben Abdellah, Fez, Morocco. e-mail: redouane.elmezegueldy@usmba.ac.ma

Youssef Ouakrim
Laboratory MMPA, Department of Mathematics and Informatics, ENS, University of Sidi Mohamed Ben Abdellah, Fez, Morocco. e-mail: yosf.ouakrim@gmail.com

Mohamed Ouzahra
Laboratory MMPA, Department of Mathematics and Informatics, ENS, University of Sidi Mohamed Ben Abdellah, Fez, Morocco. e-mail: mohamed.ouzahra@usmba.ac.ma

How to cite this article
El Mezegueldy, R., Ouakrim, Y. and Ouzahra, M., Bilinear optimal control of a reaction-diffusion equation: overcoming boundedness constraints on controls. *Iran. J. Numer. Anal. Optim.*, 2025; 15(2): 570-599. https://doi.org/10.22067/ijnao.2025.90287.1538

Under suitable assumptions on the system parameters, we establish the existence of optimal controls and derive first-order optimality conditions. Our theoretical findings are supported by numerical simulations that validate the practical effectiveness of the proposed approach. This work provides a new framework for handling bilinear control problems without artificial constraints, offering potential applications in population dynamics and ecological management.

AMS subject classifications (2020): 49J20, 35K57, 49K20, 35K58

Keywords: Bilinear optimal control; Reaction-diffusion equations; Essentially bounded controls; Optimality conditions.

1 Introduction

In this paper, we address the control problem of bilinear reaction-diffusion equations with essentially bounded controls. Specifically, we consider the following evolution problem:

$$\begin{cases} \dot{y}(x,t) + Ay(x,t) = u(x,t)L(y(x,t)) + f(x,t) \ \text{ in } \ \Omega \times (0,T), \\ \dfrac{\partial y}{\partial \mathbf{n}} = 0 \text{ on } \partial\Omega, \\ y(0) = y_0. \end{cases} \tag{1}$$

where $\Omega \subset \mathbb{R}^N$ is an open bounded Lipschitz domain, $A := -\Delta$ is the Laplacian operator, $T > 0$ is a fixed final time, $Q = \Omega \times (0,T)$, and $\Sigma = \partial\Omega \times (0,T)$. The Neumann boundary condition $\frac{\partial y}{\partial \mathbf{n}} = 0$ ensures that there is no flux across the boundary $\partial\Omega$, where $\mathbf{n}$ denotes the outward unit normal vector to $\partial\Omega$.

The optimization problem we study is as follows:

$$\text{Find } \ \bar{u} \ \text{ such that } \ J(\bar{u}) = \inf_{u \in L^\infty(Q)} J(u), \tag{P}$$

where the cost function is expressed as

$$J(u) := \frac{1}{2} \iint_Q (y(x,t) - y_d(x,t))^2 \, \mathrm{d}x\mathrm{d}t + \frac{\varepsilon}{p} \iint_Q |u(x,t)|^p \, \mathrm{d}x\mathrm{d}t, \tag{2}$$

with y_d representing a target state, y being the solution to (1) associated with u, and $\varepsilon > 0$ being a regularization parameter.

Throughout this paper, we assume the following assumptions:

Assumption 1. *The spatial dimension N satisfies $N \geq 2$.*
This condition ensures that the inclusion $H^1(\Omega) \subset L^2(\Omega)$ is compact. However, for the one-dimensional system, we can also consider homogeneous Dirichlet boundary condition where the injection $H_0^1(\Omega) \subset L^2(\Omega)$ is already compact.

Assumption 2. *The exponent p is chosen such that $1 + N/2 < p < \infty$. Additionally, y_d, $f \in L^p(Q)$, and $y_0 \in C(\bar{\Omega})$.*
These conditions are essential for deriving L^∞-norm bounds for both the state and adjoint state. The problem is also discussed under weaker assumptions (see Section 5).

Assumption 3. *The operator $L : L^{p^*}(\Omega) \to L^\infty(\Omega)$ is assumed to be bounded and linear, where p^* is the conjugate exponent of p, that is, $p^* = \frac{p}{p-1}$.*
Examples of such operators include spectral projections [32], Bochner–Riesz means and spectral measures [30, 10], as well as spectral multipliers of self-adjoint operators [29, 25]. Spectral multipliers, in particular, enable us to control the behavior of the original operator by modifying its spectrum through a specific function.

Under these assumptions, (1) serves as a fundamental model in population dynamics, describing the spatio-temporal evolution of a single species within a geographically isolated domain. The equation captures the combined effects of diffusive movement, represented by the term $Ay(x,t)$, and population growth dynamics, influenced by the bilinear control term $u(x,t)L(y(x,t))$, where $y(x,t)$ denotes the species density at position x and time t. The bilinear control structure is particularly notable, as the control $u(x,t)$ acts as a spatial distribution kernel that modulates reproduction rates, offering significant applications in ecological management and optimization problems. The non-local nature of the reaction term, $u(x,t)L(y(x,t))$, reflects the realistic scenario where population changes at a given point depend on the entire species distribution. Furthermore, the Neumann boundary conditions impose

population conservation across the domain, ensuring no flux at the boundary. This model's formulation is applicable in various ecological contexts and control theory frameworks, with further details discussed in [17, 3, 20, 28, 2].

Traditional approaches for optimization with essentially bounded controls require imposing a priori bounds on the set of admissible controls to guarantee the existence of solutions (e.g., [4, 23, 13, 16, 14, 1, 26]). However, this practice presents significant drawbacks. Determining appropriate bounds can be challenging in practical applications. Moreover, choosing an upper bound that is not large enough can lead to suboptimal solutions, potentially compromising the effectiveness of the control strategy. The methods used in these references are therefore not suitable for solving the problem at hand.

Our work aims to overcome these limitations by developing a framework that does not require such constraints. We formulate the control problem with admissible controls in $L^\infty(Q)$ without directly imposing any explicit boundedness conditions on the controls. To tackle this challenge, we introduce an auxiliary optimal control problem, where the control set is the space $L^p(Q)$, as defined at the beginning of Section 3. Through rigorous analysis, we establish in Proposition 3 that these problems are equivalent in terms of optimality: A control solves the auxiliary problem if and only if it solves the original problem in $L^\infty(Q)$. This fundamental equivalence enables us to solve the original optimization problem without requiring a priori bounds.

This equivalence holds when both the optimal state $\bar{y}$ and optimal adjoint state $\bar{\varphi}$ belong to $L^\infty(Q)$. Notably, when $\bar{u} \in L^p(Q)$ solves the auxiliary problem, the necessary optimality condition yields

$$|\bar{u}(x,t)| = C|L(\bar{y}(x,t))\bar{\varphi}(x,t)|^{\frac{1}{p-1}} \quad \text{a.e. in } Q,$$

from which we deduce $\bar{u} \in L^\infty(Q)$. While the well-posedness of system (1) in $C([0,T];L^2(\Omega))$ is known [22, p. 66], our optimal control framework requires enhanced regularity properties. We establish the existence and uniqueness of solutions in $L^\infty(Q)$ for both state and adjoint equations through fixed-point analysis under Assumptions (2) and (3), where the interplay between L and p plays a crucial role in achieving this improved regularity.

Related problems have been studied in the literature for systems governed by an additive control operator [9, 8]. For instance, Casas and Kunisch [8]

investigated the minimization of a quadratic cost function ($p = 2$) subject to an infinite-horizon semilinear system of dimension up to $N = 3$. For a bilinear control operator, when considering a quadratic cost function, the problem can only be addressed for one-dimensional systems (see section 5 for more details). Therefore, strong assumptions are required to handle bilinear optimal control, compared to additive control within the framework of minimization over $L^\infty(Q)$.

Considerable attention has been devoted in the literature to systems governed by semilinear equations with multiplicative controls. For instance, optimization problems involving purely time-dependent controls have been extensively explored [5, 34, 35], as have problems involving distributed controls, as discussed in [18, 31]. However, the methods developed in these studies are not directly applicable to the bilinear control problem addressed here, as they typically consider controls belonging to Hilbert spaces, while our framework necessitates a different approach. Additionally, studies on the controllability of such systems can be found in [27, 7].

The cost function analyzed in this work has been studied in various contexts. For instance, similar functionals have been explored in bilinear control problems involving chemotaxis-consumption systems in bounded domains [19, 11]. However, the methods used in these studies are not suitable for the specific bilinear optimal control problem considered here, as they rely on the assumption that optimal controls exist within Banach spaces of the form $L^r(Q)$, for some $r > 1$. Additionally, analogous functionals have been examined in optimal control problems involving purely time-dependent controls within bilinear infinite-dimensional systems [36]. This method is not applicable to our problem, as the admissible control set in our framework extends beyond pure time-dependent controls.

The primary contribution of this work is the formulation of a new optimal control problem with essentially bounded controls, where no pre-defined upper bounds are placed on the admissible controls. We also demonstrate the existence of optimal controls and derive their structure using first-order optimality conditions. Numerical simulations are provided to verify the theoretical findings and emphasize the practical relevance of the proposed approach.

The organization of the paper is as follows: In Section 2, we establish the well-posedness of the state equation. Section 3 proves the existence of an optimal control and derives the first-order necessary conditions for optimality. Section 4 presents numerical examples to illustrate the theoretical results. Section 5 offers additional remarks and discussions. Finally, section 6 concludes the paper.

2 Well-posedness of the state equation

This section focuses on establishing the well-posedness of the state equation. We start by defining the appropriate function spaces and discussing their properties. Next, we introduce the variational formulation and provide the definition of a weak solution to (1).

Let $W(0,T)$ denote the Hilbert space

$$W(0,T) = \left\{ y \in L^2(0,T; H^1(\Omega)) : \dot{y} \in H^1(\Omega)^* \right\},$$

endowed with the norm

$$\|y\|_{W(0,T)} = \left(\int_0^T \left(\|y(t)\|^2_{L^2(\Omega)} + \sum_{i=1}^N \left\| \frac{\partial y(t)}{\partial x_i} \right\|^2_{L^2(\Omega)} + \|\dot{y}(t)\|^2_{H^1(\Omega)^*} \right) \mathrm{d}t \right)^{1/2}.$$

This space admits a continuous embedding into $C([0,T]; L^2(\Omega))$ [24], with embedding constant M_0:

$$\|y\|_{L^\infty(0,T,L^2(\Omega))} \leq M_0 \|y\|_{W(0,T)}. \tag{3}$$

Moreover, by the Rellich–Kondrachov Theorem [15] and the Aubin–Lions Lemma [24], we have the compact embedding $W(0,T) \subset L^2(Q)$. For the variational formulation, we introduce the space $W_2^{1,1}(Q)$:

$$W_2^{1,1}(Q) = \left\{ y \in L^2(Q) : \dot{y} \in L^2(Q) \text{ and } \frac{\partial y}{\partial x_i} \in L^2(Q), \quad \text{for all } i = 1, \ldots, N \right\},$$

endowed with the norm

$$\|y\|_{W_2^{1,1}(Q)} = \left(\int_0^T \left(\|y(t)\|^2_{L^2(\Omega)} + \sum_{i=1}^N \left\| \frac{\partial y(t)}{\partial x_i} \right\|^2_{L^2(\Omega)} + \|\dot{y}(t)\|^2_{L^2(\Omega)} \right) \mathrm{d}t \right)^{1/2}.$$

The test function space $\mathcal{W}$ is defined as

$$\mathcal{W} = \left\{ v \in W_2^{1,1}(Q) : v(x,T) = 0 \text{ for almost all } x \in \Omega \right\}.$$

For the operator L, we define

$$|L| = \max\left\{ \|L\|_{\mathcal{L}(L^2(\Omega),L^\infty(\Omega))}, \|L\|_{\mathcal{L}(L^{p^*}(\Omega),L^\infty(\Omega))} \right\}.$$

The variational formulation is obtained by multiplying the state equation with a test function $v \in \mathcal{W}$ and integrating over Q:

$$\iint_Q \Big(-y\dot{v} + \nabla y \nabla v \Big) \mathrm{d}x\mathrm{d}t \quad = \iint_Q \Big(uL(y)v + fv \Big) \mathrm{d}x\mathrm{d}t + \int_\Omega y_0\, v(0)\, \mathrm{d}x. \quad (4)$$

Definition 1. A function $y \in W(0,T)$ is called a weak solution to the state equation (1) if it satisfies the variational equality (4) for all test functions $v \in \mathcal{W}$.

The following enhanced regularity result holds for linear systems.

Lemma 1. [33, Ch.V] Let $g \in L^p(Q)$. For any $z_0 \in C(\bar{\Omega})$, the system

$$\begin{cases} \dot{z} + Az = g & \text{in} \quad Q, \\ \frac{\partial z}{\partial \mathbf{n}} = 0 & \text{on} \quad \Sigma, \\ z(0) = z_0 & \text{in} \quad \Omega, \end{cases} \quad (5)$$

admits a unique solution $z \in W(0,T) \cap C(\bar{Q})$ satisfying

$$\|z\|_{W(0,T)} + \|z\|_{L^\infty(Q)} \leq c_\infty \Big(\|z_0\|_{L^\infty(\Omega)} + \|g\|_{L^p(Q)} \Big), \quad (6)$$

where c_∞ is a positive constant, non-decreasing with respect to $|Q|$ and independent of g and z_0.

Remark 1. The constant c_∞ in Lemma 1 depends only on $|Q|$, as detailed in [21, Ch.III]. Moreover, the estimate (6) remains valid with the same constant c_∞ when considering (5) over any subinterval $(\tau_1, \tau_2) \subseteq (0,T)$. This time-independence of c_∞ will be crucial in our subsequent analysis.

Building on this result, we establish the well-posedness of the state equation.

Proposition 1. For every $u \in L^p(Q)$, the state equation (1) possesses a unique weak solution $y \in W(0,T) \cap C(\bar{Q})$. In addition, this solution satisfies the following priori bound:

$$\|y\|_{W(0,T)} + \|y\|_{L^\infty(Q)} \leq C_u \Big(\|y_0\|_{L^\infty(\Omega)} + \|f\|_{L^p(Q)} \Big), \tag{7}$$

where C_u is a constant that does not depend on y_0 and f, and non-decreasing with respect to $\|u\|_{L^p(Q)}$.

Proof. Fix an arbitrary $\tau \in (0,T]$. Let $Q_\tau := \Omega \times (0,\tau)$, $\Sigma_\tau := \partial\Omega \times (0,\tau)$, and define the Banach space $X_\tau := W(0,\tau) \cap C(\bar{Q}_\tau)$ equipped with the norm $\|y\|_{X_\tau} = \|y\|_{W(0,\tau)} + \|y\|_{L^\infty(Q_\tau)}$.
Consider the mapping $\mathcal{Y} : X_\tau \to X_\tau$, where $\mathcal{Y}(w) = y$ solves

$$\begin{cases} \dot{y} + Ay = uL(w) + f & \text{in} \quad Q_\tau, \\ \frac{\partial y}{\partial \mathbf{n}} = 0 & \text{on} \quad \Sigma_\tau, \\ y(0) = y_0 & \text{in} \quad \Omega. \end{cases} \tag{8}$$

By Lemma 1, $\mathcal{Y}$ is well-defined since $uL(w) + f \in L^p(Q_\tau)$. To show $\mathcal{Y}$ is a contraction, take $w_1, w_2 \in X_\tau$ and let $z = \mathcal{Y}(w_1) - \mathcal{Y}(w_2)$. Then z satisfies

$$\begin{cases} \dot{z} + Az = uL(w_1 - w_2) & \text{in} \quad Q_\tau, \\ \frac{\partial z}{\partial \mathbf{n}} = 0 & \text{on} \quad \Sigma_\tau, \\ z(0) = 0 & \text{in} \quad \Omega. \end{cases} \tag{9}$$

Applying Lemma 1 and using (3), we obtain

$$\begin{aligned} \|\mathcal{Y}(w_1) - \mathcal{Y}(w_2)\|_{X_\tau} &\leq c_\infty \Big\| \big\| u(t) L(w_1(t) - w_2(t)) \big\|_{L^p(\Omega)} \Big\|_{L^p(0,\tau)} \\ &\leq c_\infty |L| \, \Big\| \|u(t)\|_{L^p(\Omega)} \, \|w_1(t) - w_2(t)\|_{L^2(\Omega)} \Big\|_{L^p(0,\tau)} \\ &\leq c_\infty |L| \, \|u\|_{L^p(Q_\tau)} \|w_1 - w_2\|_{L^\infty(0,\tau;L^2(\Omega))} \\ &\leq c_\infty M_0 |L| \|u\|_{L^p(Q_\tau)} \|w_1 - w_2\|_{X_\tau}. \end{aligned}$$

Therefore, $\mathcal{Y}$ is a contraction mapping if

$$c_\infty M_0 |L| \|u\|_{L^p(Q_\tau)} < 1.$$

If the above condition is not met, we subdivide the interval $[0,\tau]$ into subintervals $[0,\tau_1],[\tau_1,\tau_2],\ldots,[\tau_{I-1},\tau_I]=[\tau_{I-1},\tau]$ such that

$$c_\infty M_0|L|\|u\|_{L^p(\tau_i,\tau_{i+1};L^p(\Omega))}\leq 1/2,\qquad \text{for all } i=0,\ldots,I-1,$$

and we consider the following systems:

$$(\mathcal{S}_0)\begin{cases}\dot{y}^0+Ay^0=uL(y^0)+f & \text{in}\quad \Omega\times(0,\tau_1),\\ \frac{\partial y^0}{\partial \mathbf{n}}=0 & \text{on}\quad \partial\Omega\times(0,\tau_1),\\ y^0(0)=y_0 & \text{in}\quad \Omega,\end{cases}$$

and for $i=1,\ldots,I-1$,

$$(\mathcal{S}_i)\begin{cases}\dot{y}^i+Ay^i=uL(y^i)+f & \text{in}\quad \Omega\times(\tau_i,\tau_{i+1}),\\ \frac{\partial y^i}{\partial \mathbf{n}}=0 & \text{on}\quad \partial\Omega\times(\tau_i,\tau_{i+1}),\\ y^i(\tau_i)=y^{i-1}(\tau_i) & \text{in}\quad \Omega.\end{cases}$$

We then define solutions on each subinterval using a similar approach as before. This leads to a solution y for the entire interval $[0,\tau]$ by piecing together the solutions on subintervals.
Moreover, this solution satisfies

$$\|y\|_{X_\tau}\leq c_\infty\Big(\|y_0\|_{L^\infty(\Omega)}+\|uL(y)+f\|_{L^p(Q_\tau)}\Big). \tag{10}$$

This estimate can be further refined to obtain

$$\|y\|_{X_\tau}\leq c_\infty\Big(\|y_0\|_{L^\infty(\Omega)}+\|f\|_{L^p(Q_\tau)}+|L|\sqrt{|\Omega|}\,\|u\|_{L^p(Q_\tau)}\,\|y\|_{L^\infty(Q_\tau)}\Big). \tag{11}$$

We derive also from (10)

$$\begin{aligned}\|y(\tau)\|_{L^\infty(\Omega)}&\leq c_\infty\Bigg(\|y_0\|_{L^\infty(\Omega)}+\|f\|_{L^p(Q_\tau)}\\ &\qquad+|L|\left(\int_0^\tau\|u(t)\|^p_{L^p(\Omega)}\|y(t)\|^p_{L^2(\Omega)}\,\mathrm{d}t\right)^{\frac{1}{p}}\Bigg)\\ &\leq c_\infty\Bigg(\|y_0\|_{L^\infty(\Omega)}+\|f\|_{L^p(Q_\tau)}\\ &\qquad+|L|\sqrt{|\Omega|}\left(\int_0^\tau\|u(t)\|^p_{L^p(\Omega)}\|y(t)\|^p_{L^\infty(\Omega)}\,\mathrm{d}t\right)^{\frac{1}{p}}\Bigg).\end{aligned}$$

Taking both sides of this inequality to the p-th power, and using the fact $(a+b)^p \leq 2^{p-1}(a^p+b^p)$ for any non-negative a and b, we get

$$\|y(\tau)\|^p_{L^\infty(\Omega)} \leq 2^{p-1}c^p_\infty \Bigg(\|y_0\|^p_{L^\infty(\Omega)} + \|f\|^p_{L^p(Q_\tau)} + \Big(|L|\sqrt{|\Omega|}\Big)^p \int_0^\tau \|u(t)\|^p_{L^p(\Omega)} \|y(t)\|^p_{L^\infty(\Omega)} \,\mathrm{d}t \Bigg).$$

By applying Gronwall's lemma, we derive

$$\begin{aligned} \|y(\tau)\|^p_{L^\infty(\Omega)} &\leq 2^{p-1}c^p_\infty \Big(\|y_0\|^p_{L^\infty(\Omega)} + \|f\|^p_{L^p(Q_\tau)} \Big) e^{2^{p-1}c^p_\infty \left(|L|\sqrt{|\Omega|}\right)^p \|u\|^p_{L^p(Q_\tau)}} \\ &\leq 2^{p-1}c^p_\infty \Big(\|y_0\|_{L^\infty(\Omega)} + \|f\|_{L^p(Q)} \Big)^p e^{2^{p-1}\left(c_\infty|L|\sqrt{|\Omega|}\right)^p \|u\|^p_{L^p(Q)}}. \end{aligned}$$

This inequality is valid for all $\tau \in [0,T]$. Consequently, we have

$$\|y\|_{L^\infty(Q)} \leq 2^{\frac{p-1}{p}} c_\infty \Big(\|y_0\|_{L^\infty(\Omega)} + \|f\|_{L^p(Q)} \Big) e^{\frac{2^{p-1}\left(c_\infty|L|\sqrt{|\Omega|}\right)^p}{p} \|u\|^p_{L^p(Q)}}. \quad (12)$$

Combining estimates (11) and (12), we get

$$\begin{aligned} &\|y\|_{W(0,T)} + \|y\|_{L^\infty(Q)} \\ &\leq c_\infty \Big(\|y_0\|_{L^\infty(\Omega)} + \|f\|_{L^p(Q)} \Big) \\ &\quad + c^2_\infty |L| \sqrt{|\Omega|}\, \|u\|_{L^p(Q)} 2^{\frac{p-1}{p}} \Big(\|y_0\|_{L^\infty(\Omega)} + \|f\|_{L^p(Q)} \Big) e^{\frac{2^{p-1}\left(c_\infty|L|\sqrt{|\Omega|}\right)^p}{p} \|u\|^p_{L^p(Q)}}. \end{aligned}$$

Finally, this estimate leads to the desired estimate (7) with C_u defined as

$$C_u = c_\infty + c^2_\infty |L| \sqrt{|\Omega|}\, \|u\|_{L^p(Q)} 2^{\frac{p-1}{p}} e^{\frac{2^{p-1}\left(c_\infty|L|\sqrt{|\Omega|}\right)^p}{p} \|u\|^p_{L^p(Q)}}.$$

□

3 Existence and characterization of optimal controls

In this section, we analyze the optimal control problem. We introduce an auxiliary optimization problem with controls from the function space $L^p(Q)$. Using standard methods from the calculus of variations, we establish the existence of a solution and provide its characterization via an optimality

system. Finally, we show that any solution to the auxiliary problem is also a solution to the original problem (P).

Consider the following auxiliary problem:

$$\text{Find } \bar{u} \text{ such that } J(\bar{u}) = \inf_{u \in L^p(Q)} J(u). \tag{P_A}$$

In the following proposition, we show that this problem has a solution.

Proposition 2. The minimization problem (P_A) admits at least one solution.

Proof. Let us begin by selecting a minimizing sequence (u_n) in $L^p(Q)$. The coercivity property of J over $L^p(Q)$ ensures that this sequence is bounded. By the Banach-Alaoglu theorem, we can extract a subsequence that converges weakly to some element $\bar{u}$ in $L^p(Q)$.
For each control u_n, let y_n denote the corresponding state solution to (1). From the boundedness properties established in Proposition 1, we know that (y_n) forms a bounded sequence in $W(0,T)$. The reflexivity of this space allows us to extract a subsequence that converges weakly to some $\bar{y}$ in $W(0,T)$. Moreover, the compact embedding theorem yields strong convergence of a subsequence to $\bar{y}$ in $L^2(Q)$.
We must now verify that $\bar{y}$ is indeed the state associated with the control $\bar{u}$. For any test function $v \in \mathcal{W}$, we have the weak convergences

$$\begin{aligned} \dot{y}_n &\rightharpoonup \dot{\bar{y}} \quad \text{weakly in } L^2(0,T;H^1(\Omega)^*), \\ Ay_n &\rightharpoonup A\bar{y} \text{ weakly in } L^2(0,T;H^1(\Omega)^*). \end{aligned}$$

These convergences imply

$$\iint_Q (-y_n \dot{v} + \nabla y_n \nabla v)\,\mathrm{d}x\mathrm{d}t \underset{n\to\infty}{\longrightarrow} \iint_Q (-\bar{y}\dot{v} + \nabla \bar{y}\nabla v)\,\mathrm{d}x\mathrm{d}t. \tag{13}$$

For the remaining terms, we have the following decomposition:

$$u_n L(y_n) - \bar{u}L(\bar{y}) = (u_n - u)L(\bar{y}) + u_n L(y_n - \bar{y}). \tag{14}$$

The mapping $u \mapsto uL(\bar{y})$ is continuous from $L^p(Q)$ to $L^2(Q)$, yielding (see [6, Theorem III-3.10])

$$(u_n - \bar{u})L(\bar{y}) \rightharpoonup 0 \quad \text{weakly in} \quad L^2(Q). \tag{15}$$

Furthermore, we derive

$$\begin{aligned}
&\iint_Q u_n L(y_n - \bar{y}) v \, \mathrm{d}x\mathrm{d}t \\
&\leq \int_0^T \|u_n(t)L(y_n(t) - \bar{y}(t))\|_{L^2(\Omega)} \|v(t)\|_{L^2(\Omega)} \, \mathrm{d}t \\
&\leq |L||\Omega|^{\frac{p+2}{2p}} \int_0^T \|u(t)\|_{L^p(\Omega)} \|y_n(t) - \bar{y}(t)\|_{L^2(\Omega)} \|v(t)\|_{L^2(\Omega)} \, \mathrm{d}t \\
&\leq |L|(|\Omega|T)^{\frac{p+2}{2p}} \|u_n\|_{L^p(Q)} \|v\|_{L^\infty(0,T;L^2(\Omega))} \|y_n - \bar{y}\|_{L^2(Q)} \underset{n\to\infty}{\longrightarrow} 0.
\end{aligned}$$

Combining this with (13) and (15), we conclude that $\bar{y}$ is indeed the solution to (1) corresponding to the control $\bar{u}$.
Finally, the lower semicontinuity of norms yields

$$J(\bar{u}) \leq \liminf_n J(u_n) \leq \inf_{u \in L^p(Q)} J(u),$$

which proves the optimality of $\bar{u}$ in $L^p(Q)$. □

We introduce the control-to-state map G, defined as $G : L^p(Q) \longrightarrow W(0,T)$, where $G(u) = y$ represents the solution to (1) corresponding to u. The following lemma establishes the differentiability property of G.

Lemma 2. The operator G is Fréchet differentiable. Its derivative at $u \in L^p(Q)$ in the direction of $h \in L^p(Q)$, denoted by z_h, is defined as follows:

$$\begin{cases}
\dot{z}_h + Az_h = uL(z_h) + hL(G(u)), \\
\frac{\partial z_h}{\partial \mathbf{n}} = 0, \\
z_h(0) = 0.
\end{cases} \tag{16}$$

Proof. Let us consider arbitrary controls u and h in $L^p(Q)$. According to Proposition 1, both $G(u)$ and $G(u+h)$ exist in $W(0,T)$ and satisfy the following inequalities:

$$\begin{aligned}
\|G(u)\|_{W(0,T)} &\leq C_u \left(\|y_0\|_{L^\infty(\Omega)} + \|f\|_{L^p(Q)} \right), \\
\|G(u+h)\|_{W(0,T)} &\leq C_{u+h} \left(\|y_0\|_{L^\infty(\Omega)} + \|f\|_{L^p(Q)} \right).
\end{aligned}$$

The mapping $h \mapsto z_h$ is well-defined and bounded:

$$\begin{aligned}\|z_h\|_{W(0,T)} &\leq C_u\|hL(G(u))\|_{L^p(Q)}\\ &\leq C_u|L|\|y\|_{L^\infty(0,T;L^2(\Omega))}\|h\|_{L^p(Q)}\\ &\leq C_u|L|M_0\|y\|_{W(0,T)}\|h\|_{L^p(Q)}\\ &\leq c_1\|h\|_{L^p(Q)}.\end{aligned}$$

Here, $c_1 = C_u^2|L|M_0\left(\|y_0\|_{L^\infty(\Omega)} + \|f\|_{L^p(Q)}\right)$.
We define $w_1 = G(u+h) - G(u)$. Then, w_1 satisfies the following system:

$$\begin{cases}\dot{w}_1 + Aw_1 = uL(w_1) + hL(G(u+h)),\\ \frac{\partial w_1}{\partial \mathbf{n}} = 0,\\ w_1(0) = 0.\end{cases} \tag{17}$$

This system has a solution satisfying

$$\begin{aligned}\|w_1\|_{W(0,T)} &\leq C_u\|hL(G(u+h))\|_{L^p(Q)}\\ &\leq c_2\|h\|_{L^p(Q)}.\end{aligned}$$

Here, $c_2 = C_uC_{u+h}|L|M_0\left(\|y_0\|_{L^\infty(\Omega)} + \|f\|_{L^p(Q)}\right)$.
Next, we define $w_2 = w_1 - z_h$. Then, w_2 satisfies

$$\begin{cases}\dot{w}_2 + Aw_2 = uL(w_2) + hL(w_1),\\ \frac{\partial w_2}{\partial \mathbf{n}} = 0,\\ w_2(0) = 0.\end{cases} \tag{18}$$

Furthermore, we have

$$\begin{aligned}\|w_2\|_{W(0,T)} &\leq C_u\|hL(w_1)\|_{L^p(Q)}\\ &\leq c_3\|h\|^2_{L^p(Q)}.\end{aligned}$$

Here, $c_3 = C_u|L|M_0c_2$. Consequently,

$$\frac{\|G(u+h) - G(u) - z_h\|_{W(0,T)}}{\|h\|_{L^p(Q)}} \leq c_3\|h\|_{L^p(Q)} \longrightarrow 0 \quad \text{as} \quad \|h\|_{L^p(Q)} \to 0.$$

□

The Fréchet differentiability of the control-to-state map and the smoothness of the L^2 and L^p norms ensure the differentiability of the cost function.

Lemma 3. The cost function is Fréchet differentiable. Its derivative at u in the direction of h is given by

$$J'(u)h = \iint_Q \Big((y - y_d)z_h \ + \varepsilon \operatorname{sgn}(u)\,|u|^{p-1}h\Big)\,\mathrm{d}x\mathrm{d}t, \tag{19}$$

where z_h is the solution to (16), and sgn stands for the signum function.

Since the operator $L : L^{p^*}(\Omega) \to L^{\infty}(\Omega)$ is bounded and linear, we define its adjoint $L^* : L^1(\Omega) \to L^p(\Omega)$ as follows:

$$\langle L(y_1), y_2\rangle_{L^{\infty}(\Omega),L^1(\Omega)} = \langle y_1, L^*(y_2)\rangle_{L^{p^*}(\Omega),L^p(\Omega)}, \quad \text{for all } y_1 \in L^{p^*}(\Omega),\ y_2 \in L^1(\Omega).$$

This adjoint operator L^* is also bounded and linear.

We now prove the existence of optimal solutions and provide their characterization for the auxiliary problem (P_A).

Theorem 1. Problem (P_A) has at least one solution. Moreover, for every solution $\bar{u}$, there exist $\bar{y}, \bar{\varphi} \in W(0,T) \cap C(\bar{Q})$, such that the following conditions hold:

$$\begin{cases} \dot{\bar{y}} + A\bar{y} = \bar{u}L(\bar{y}) + f & \text{in} \quad Q, \\ \frac{\partial \bar{y}}{\partial \mathbf{n}} = 0 & \text{on} \quad \Sigma, \\ \bar{y}(0) = y_0 & \text{in} \quad \Omega, \end{cases} \tag{20}$$

$$\begin{cases} \dot{\bar{\varphi}} - A\bar{\varphi} = -L^*(\bar{u}\bar{\varphi}) - (\bar{y} - y_d) & \text{in } Q, \\ \frac{\partial \bar{\varphi}}{\partial \mathbf{n}} = 0 & \text{on } \Sigma, \\ \bar{\varphi}(T) = 0 & \text{in } \ \Omega, \end{cases} \tag{21}$$

$$\iint_Q \Big(L(\bar{y})\bar{\varphi} + \varepsilon \operatorname{sgn}(\bar{u})\,|\bar{u}|^{p-1}\Big)(u - \bar{u})\,\mathrm{d}x\mathrm{d}t \geq 0, \quad \text{for all } u \in L^p(Q), \tag{22}$$

and

$$\bar{u}(x,t) = -\frac{\operatorname{sgn}(L(\bar{y})\bar{\varphi})}{\varepsilon^{\frac{1}{p-1}}}\,|L(\bar{y})\bar{\varphi}|^{\frac{1}{p-1}} \quad \text{a.e. in} \quad Q. \tag{23}$$

Proof. By Proposition 2, the problem (P_A) has a solution $\bar{u}$. Let $\bar{y}$ denote the corresponding trajectory that satisfies (20). We will first prove the existence of $\bar{\varphi}$ that satisfies (21). Then derive the variational inequality (22), and conclude by obtaining the optimal control characterization in (23).

Step 1: Existence and uniqueness of $\bar{\varphi}$.

We proceed to prove that (21) admits a unique solution. To this end, we

consider the transformation

$$q(t) = \bar{\varphi}(T-t), \quad v(t) = \bar{u}(T-t), \quad g(t) = \bar{y}(T-t) - y_d(T-t),$$

for almost all $t \in [0,T]$. This transforms (21) into

$$\begin{cases} \dot{q} + Aq = L^*(vq) + g & \text{in } Q, \\ \frac{\partial q}{\partial \mathbf{n}} = 0 & \text{on } \Sigma, \\ q(0) = 0 & \text{in } \Omega. \end{cases}$$

Since $\bar{y} - y_d \in L^p(Q)$, Proposition 1 ensures a unique solution $q \in W(0,T) \cap C(\bar{Q})$. Thus, there exists a unique $\bar{\varphi} \in W(0,T) \cap C(\bar{Q})$ satisfying (21).

Step 2: Derivation of the inequality (22).

As $\bar{u}$ is optimal for problem (P_A), we have $J'(\bar{u})(u-\bar{u}) \geq 0$ for all $u \in L^p(Q)$. Let $h = u - \bar{u}$. Using Lemma 3, we can express this as

$$J'(\bar{u})h = \iint_Q \Big((\bar{y} - y_d) z_h + \varepsilon \operatorname{sgn}(\bar{u}) \, |\bar{u}|^{p-1} h \Big) \, \mathrm{d}x\mathrm{d}t \geq 0, \tag{24}$$

where z_h satisfies (16).

Now, let us focus on simplifying the term $\iint_Q (\bar{y} - y_d) z_h \, \mathrm{d}x\mathrm{d}t$:

$$\begin{aligned} \iint_Q (\bar{y} - y_d) z_h \, \mathrm{d}x\mathrm{d}t &= \iint_Q (-\dot{\bar{\varphi}} + A\bar{\varphi} - L^*(\bar{u}\bar{\varphi})) z_h \, \mathrm{d}x\mathrm{d}t \quad \text{(using (21))} \\ &= \iint_Q (-\dot{\bar{\varphi}} z_h + \bar{\varphi} A z_h - L^*(\bar{u}\bar{\varphi}) z_h) \, \mathrm{d}x\mathrm{d}t \\ &= \iint_Q (\bar{\varphi} \dot{z}_h + \bar{\varphi} A z_h - L^*(\bar{u}\bar{\varphi}) z_h) \, \mathrm{d}x\mathrm{d}t \\ &\quad - \int_\Omega \bar{\varphi}(T) z_h(T) \mathrm{d}x \\ &\quad + \int_\Omega \bar{\varphi}(0) z_h(0) \, \mathrm{d}x. \qquad \text{(integration by parts)} \end{aligned}$$

Using the equation for z_h, and noting that $\bar{\varphi}(T) = 0$ and $z_h(0) = 0$, we get

$$\iint_Q (\bar{y} - y_d) z_h \, \mathrm{d}x\mathrm{d}t = \iint_Q (\bar{\varphi}\bar{u} L(z_h) + \bar{\varphi} h L(\bar{y}) - L^*(\bar{u}\bar{\varphi}) z_h) \, \mathrm{d}x\mathrm{d}t.$$

By definition of L^*, we infer

$$\iint_Q (\bar{y} - y_d) z_h \, \mathrm{d}x\mathrm{d}t = \iint_Q h \bar{\varphi} L(\bar{y}) \, \mathrm{d}x\mathrm{d}t.$$

Substituting this back into (24), we obtain

$$\iint_Q \left(h\bar{\varphi}L(\bar{y}) + \varepsilon \operatorname{sgn}(\bar{u})\,|\bar{u}|^{p-1}h\right) \mathrm{d}x\mathrm{d}t \geq 0,$$

which gives the variational inequality (22).

Step 3: Derivation of characterization (23).

From the variational inequality (22), we can deduce the following pointwise relationship:

$$L(\bar{y})\bar{\varphi} + \varepsilon \operatorname{sgn}(\bar{u})\,|\bar{u}|^{p-1} = 0 \quad \text{a.e. in } Q. \tag{25}$$

The derivation of the pointwise condition (25) from the variational inequality (22) is standard, see for instance the method outlined in [33, Section 2.8]. Rearranging this equation and taking the $(p-1)$-th root, we obtain

$$\bar{u} = -\frac{\operatorname{sgn}(L(\bar{y})\bar{\varphi})}{\varepsilon^{\frac{1}{p-1}}} |L(\bar{y})\bar{\varphi}|^{\frac{1}{p-1}} \quad \text{a.e. in } Q,$$

which is precisely the optimal control characterization (23). □

Remark 2. It is worth noting that for $p = 2$, a variant of the previous result can be obtained under weaker assumptions: Replacing (2) with $y_d, f \in L^2(Q)$ and $y_0 \in L^2(\Omega)$, as discussed in Zerrik and Boukhari [36] using a semigroup framework. However, our subsequent main result maintains the original assumptions to address the problem at hand.

Now, we state our main result.

Theorem 2. Problem (P) admits a solution $\bar{u}$. Moreover, for any solution $\bar{u}$, there exist $\bar{y}, \bar{\varphi} \in W(0,T) \cap C(\bar{Q})$ satisfying the following conditions:

$$\begin{cases} \dot{\bar{y}} + A\bar{y} = \bar{u}L(\bar{y}) + f & \text{in} \quad Q, \\ \frac{\partial \bar{y}}{\partial \mathbf{n}} = 0 & \text{on} \quad \Sigma, \\ \bar{y}(0) = y_0 & \text{in} \quad \Omega, \end{cases} \tag{26}$$

$$\begin{cases} \dot{\bar{\varphi}} - A\bar{\varphi} = -L^*(\bar{u}\bar{\varphi}) - (\bar{y} - y_d) & \text{in } Q, \\ \frac{\partial \bar{\varphi}}{\partial \mathbf{n}} = 0 & \text{on } \Sigma, \\ \bar{\varphi}(T) = 0 & \text{in } \Omega, \end{cases} \tag{27}$$

$$\iint_Q \Big(L(\bar{y})\bar{\varphi} + \varepsilon \operatorname{sgn}(\bar{u})\,|\bar{u}|^{p-1}\Big)(h-\bar{u})\,\mathrm{d}x\mathrm{d}t \geq 0, \quad \text{for all } h \in L^\infty(Q), \tag{28}$$

and

$$\bar{u}(x,t) = -\frac{\operatorname{sgn}(L(\bar{y})\bar{\varphi})}{\varepsilon^{\frac{1}{p-1}}}\,|L(\bar{y})\bar{\varphi}|^{\frac{1}{p-1}} \quad \text{a.e. in } \ Q. \tag{29}$$

Proof. Let $\bar{u}$ be a solution to problem (P_A), with associated state and adjoint variables $\bar{y}, \bar{\varphi} \in W(0,T) \cap C(\bar{Q})$. Let us prove that $\bar{u}$ is also a solution to problem (P).

First, let $\overset{\circ}{y}$ denote the solution to (26) under the null control. By Lemma 1, we conclude that

$$\|\overset{\circ}{y}\|_{W(0,T)} + \|\overset{\circ}{y}\|_{L^\infty(Q)} \leq c_\infty(\|y_0\|_{L^\infty(\Omega)} + \|f\|_{L^p(Q)}),$$

which implies

$$\|\overset{\circ}{y}\|_{L^2(Q)} \leq K_0 := c_\infty(\|y_0\|_{L^\infty(\Omega)} + \|f\|_{L^p(Q)}).$$

Since $\bar{u}$ is optimal for problem (P_A), we have $J(\bar{u}) \leq J(0)$, leading to

$$\|\bar{u}\|^p_{L^p(Q)} \leq \frac{p}{2\varepsilon}\left(\|\overset{\circ}{y} - y_d\|^2_{L^2(Q)}\right) \leq \frac{p}{\varepsilon}(K_0^2 + \|y_d\|^2_{L^2(Q)}),$$

which gives

$$\|\bar{u}\|_{L^p(Q)} \leq K_1 := \left(\frac{p}{\varepsilon}(K_0^2 + \|y_d\|^2_{L^2(Q)})\right)^{\frac{1}{p}}.$$

Using Proposition 1, we estimate $\bar{y}$ as

$$\|\bar{y}\|_{W(0,T)} + \|\bar{y}\|_{L^\infty(Q)} \leq C_{K_1}(\|y_0\|_{L^\infty(\Omega)} + \|f\|_{L^p(\Omega)}),$$

where C_{K_1} depends on K_1. The embedding $C(\bar{Q}) \subset L^p(Q)$ yields

$$\|\bar{y}\|_{L^p(Q)} \leq |Q|^{\frac{1}{p}} C_{K_1}(\|y_0\|_{L^\infty(\Omega)} + \|f\|_{L^p(\Omega)}).$$

For the operator L, we have

$$\|L(\bar{y})\|_{L^\infty(Q)} \leq K_2 := |L| M_0 C_{K_1}(\|y_0\|_{L^\infty(\Omega)} + \|f\|_{L^p(\Omega)}).$$

For $\bar{\varphi}$ satisfying (27), Theorem 1 implies

$$\|\bar{\varphi}\|_{L^\infty(Q)} \leq K_3 := C_{K_1}\left(|Q|^{\frac{1}{p}} C_{K_1}(\|y_0\|_{L^\infty(\Omega)} + \|f\|_{L^p(\Omega)}) + \|y_d\|_{L^p(Q)}\right).$$

Finally, using (29) and the bounds on $L(\bar{y})$ and $\bar{\varphi}$, we conclude

$$|\bar{u}(x,t)| \leq \frac{1}{\varepsilon^{\frac{1}{p-1}}}(K_2K_3)^{\frac{1}{p-1}} =: K_4 \quad \text{a.e. in } Q.$$

This shows that $\bar{u} \in L^\infty(Q)$ with $\|\bar{u}\|_{L^\infty(Q)} \leq K_4$. Since $L^\infty(Q) \subset L^p(Q)$, we conclude that $\bar{u}$ is also a solution of problem (P). □

The next result shows that both optimization problems share identical solution sets.

Proposition 3. Problems (P) and (P_A) are equivalent; that is, they have the same solutions.

Proof. Suppose that $\bar{u}$ is a solution of problem (P_A). By using the proof of Theorem 2, we conclude that $\bar{u}$ is also a solution of problem (P).

For the reverse implication, let $\bar{u}$ be a solution to problem (P), and consider any $u \in L^p(Q)$. We construct a sequence by defining $u_n = \Pi_{[-n,n]}(u)$ for each $n \in \mathbb{N}$, where $\Pi_{[-n,n]}$ represents the pointwise projection onto the interval $[-n,n]$. This sequence has two key properties:

- $u_n \in L^\infty(Q)$ for each n,
- $u_n \to u$ in $L^p(Q)$ as $n \to \infty$.

Since $\bar{u}$ is optimal for problem (P), we have

$$J(\bar{u}) \leq J(u_n) \quad \text{for all } n \in \mathbb{N}.$$

The continuity of J allows us to pass to the limit as $n \to \infty$, yielding

$$J(\bar{u}) \leq J(u).$$

As u was chosen arbitrarily in $L^p(Q)$, this proves that $\bar{u}$ solves problem (P_A). □

4 Simulations and numerical results

In this section, we provide numerical examples to illustrate our theoretical results and examine the behavior of the bilinear optimal control problem in both one- and two-dimensional settings.

4.1 Problem setup

A practical scenario motivating this study arises in ecological management, where the objective is to regulate the population of a species within a geographically bounded region Ω (where Ω could be a linear habitat in one-dimensional, or a surface in two-dimensional). The species density $y(x,t)$ evolves according to the dynamics:

$$\frac{\partial y}{\partial t} - \Delta y = u(x,t) \int_\Omega y(x,t)\,\mathrm{d}x, \quad (x,t) \in \Omega \times (0,T),$$

with the initial distribution $y(x,0) = y_0(x)$. Here, the control $u(x,t)$ represents interventions, such as resource allocation or habitat modification, which influence the reproduction dynamics. The goal is to steer the population density $y(x,t)$ to track a desired trajectory distribution $y_d(x,t)$ over a specified period T (e.g., one week). No explicit constraints are imposed on the control input $u(x,t)$, as it is not known whether such constraints exist, and the focus is on determining the optimal control strategy while minimizing interventions.

This problem can be reformulated as problem (P), with the operator $L : L^1(\Omega) \to L^\infty(\Omega)$ defined as

$$L(y) = \int_\Omega y(x)\,\mathrm{d}x.$$

and $f = 0$. We also consider the parameters $T = 1$ and $\varepsilon = 10^{-4}$.

A solution $\bar{u}$ to this problem must satisfy the following optimality system:

$$\frac{\partial \bar{y}}{\partial t} - \Delta \bar{y} = \bar{u} L(\bar{y}), \quad \bar{y}(0) = y_0, \tag{30}$$

$$-\frac{\partial \bar{\varphi}}{\partial t} - \Delta \bar{\varphi} = (\bar{y} - y_d) + L(\bar{u}\bar{\varphi}), \quad \bar{\varphi}(T) = 0, \tag{31}$$

and

$$\bar{u}(x,t) = -\frac{\operatorname{sgn}(L(\bar{y})\bar{\varphi})}{\varepsilon^{\frac{1}{p-1}}} \left|L(\bar{y})\bar{\varphi}\right|^{\frac{1}{p-1}}. \tag{32}$$

The boundary conditions for (30) and (31) will be specified later. We choose p such that $1 + \frac{N}{2} < p < \infty$, in accordance with Assumption (2) of our theoretical framework.

4.2 Numerical method

We employ the thresholding operator

$$u_{m+1} = S_\alpha(u_m, L(y_m)\varphi_m), \tag{33}$$

where

$$S_\alpha(u,v) = \frac{1}{\alpha + \varepsilon^{\frac{1}{p-1}}} \left(\alpha u - \operatorname{sgn}(v)|v|^{\frac{1}{p-1}}\right). \tag{34}$$

Here, $\alpha > 0$ is an accuracy threshold. The convergence of this algorithm has been rigorously established in the literature (see, for instance, [12]).

The iterative procedure for solving the optimization problem is as follows:

Algorithm 1 Optimal Control Algorithm

1: Initialize $u_0 = 0$, set $m = 0$, choose $\alpha > 0$, and tol > 0 a chosen tolerance
2: **repeat**
3: Compute y_m by solving forward the state equation.
4: Determine φ_m by solving backward the adjoint equation
5: Update control via the thresholding operator:
6: $u_{m+1} = S_\alpha(u_m, L(y_m)\varphi_m)$,
7: $m \leftarrow m + 1$
8: **until** $\|u_m - u_{m-1}\|_{L^\infty(Q)} <$ tol.

4.3 One-dimensional case

For the one-dimensional case, we examine a system on the unit interval $\Omega = (0,1)$ with $N = 1$ and $p = 2$. The system is governed by homogeneous Dirichlet boundary conditions, representing a scenario where the population density vanishes at the domain boundaries. The complete system is formulated as

$$\begin{cases} \frac{\partial y}{\partial t} - \frac{\partial^2 y}{\partial x^2} = u(x,t) \int_0^1 y(x,t)\,\mathrm{d}x, \\ y(0,t) = y(1,t) = 0, \\ y(x,0) = x - x^2. \end{cases} \tag{35}$$

We choose a target state $y_d(x,t) = \frac{1}{2}\sin(\pi x)$, which represents a desired sinusoidal distribution pattern for the population density. This choice of target state is particularly relevant for ecological applications as it maintains zero density at the boundaries while achieving maximum population density at the center of the domain.

The numerical results, presented in Figure 1, reveal several key aspects of the control system's behavior. The optimal control profile demonstrates smooth temporal evolution while maintaining essential boundedness, despite no explicit bounds being imposed. This natural boundedness validates our theoretical framework's core prediction. The corresponding state trajectory shows effective tracking of the target profile, successfully counteracting the system's natural dissipative tendency that would otherwise drive the population to extinction ($u = 0$ case). The tracking error exhibits monotonic decrease, indicating stable convergence of the control algorithm, while the cost functional's evolution demonstrates the effective balance between tracking accuracy and control effort maintained by our chosen cost structure. These results provide strong numerical evidence for the practical viability of our theoretical framework in the one-dimensional setting.

4.4 Two-dimensional case

For the two-dimensional case, we consider a more complex setting with $\Omega = (0,1)^2$ and parameters $N = 2$ and $p = 3$, where the higher value of p satisfies our theoretical requirement $p > 1 + N/2$. The system is subject to Neumann boundary conditions, representing no-flux conditions at the domain boundaries, a natural choice for population dynamics in enclosed habitats. The complete system takes the form:

$$\begin{cases} \frac{\partial y}{\partial t} - (\frac{\partial^2 y}{\partial x_1^2} + \frac{\partial^2 y}{\partial x_2^2}) = u(x_1,x_2,t)\int_\Omega y(x_1,x_2,t)\,\mathrm{d}x_1\mathrm{d}x_2, \\ \frac{\partial y}{\partial n} = 0 \text{ on } \partial\Omega, \\ y(x_1,x_2,0) = 2x_1^3 - 3x_1^2. \end{cases} \tag{36}$$

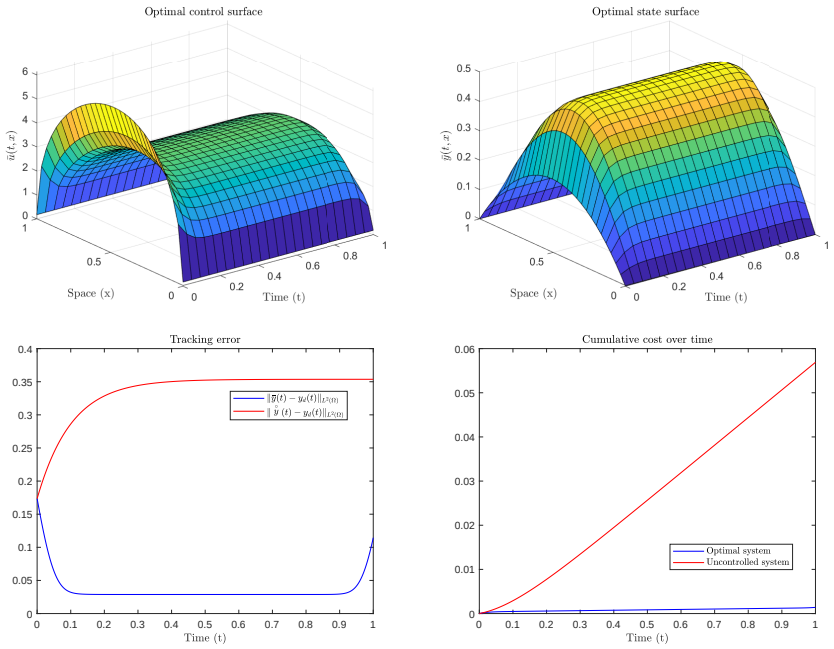


Figure 1: Optimal control (top left), corresponding state (top right), evolution of tracking error (bottom left), and cost functional (bottom right) for the one-dimensional case.

The target state is chosen as $y_d(x_1, x_2, t) = \frac{1}{2}\cos(\pi x_1)\cos(\pi x_2)\sin(\pi t)$, representing a time-varying spatial pattern that tests the system's ability to track complex spatio-temporal distributions.

The numerical results, displayed in Figure 2, demonstrate the framework's robustness in higher dimensions. The evolution of control norms confirms the maintenance of essential boundedness without explicit constraints, even in this more complex setting. The state and adjoint state norms exhibit stable behavior throughout the time horizon, indicating well-regulated dynamics. The tracking error and cumulative cost profiles show consistent convergence, though with slightly different characteristics compared to the one-dimensional case due to the increased spatial dimension. The success in maintaining the desired state pattern, particularly notable given the time-varying nature of the target, validates our theoretical framework's applicability to higher-dimensional scenarios. The effective control performance

without a priori bounds demonstrates the framework's practical utility for real-world applications in spatial population dynamics.

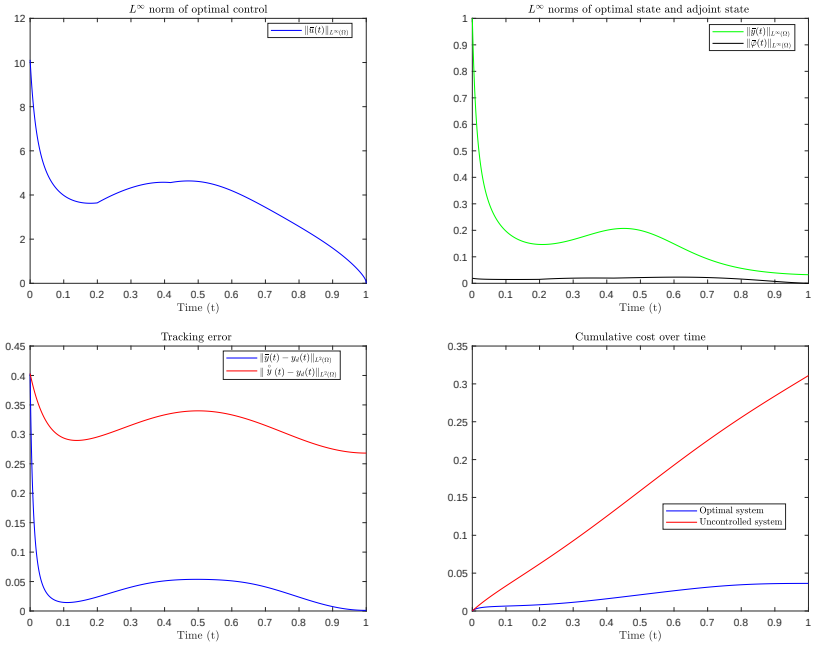


Figure 2: Results for the two-dimensional case: Optimal control norms (top left), state and adjoint state norms (top right), tracking error (bottom left), and cumulative cost (bottom right).

Both the one-dimensional and two-dimensional results demonstrate our algorithm's capability to compute essentially bounded controls that successfully achieve the desired state tracking objectives, while maintaining numerical stability and convergence properties predicted by the theoretical framework.

5 Further comments

Using an additive control operator instead of a multiplicative one can allow for weaker assumptions in solving the optimization problem. To illustrate, let's consider a quadratic cost function (i.e., $p = 2$) and investigate in which spatial dimensions the problem can be addressed for both additive and multiplicative

control operators.

To bypass the restriction imposed in (1), we consider state equations with homogeneous Dirichlet boundary conditions while maintaining Assumption (3) unchanged.

For the optimal control of a linear system, we consider the problem

$$\min_{u\in L^\infty(Q)} J(u) := \frac{1}{2}\|y-y_d\|^2_{L^2(Q)} + \frac{\varepsilon}{2}\|u\|^2_{L^2(Q)}, \tag{PL}$$

subject to

$$\begin{cases} \dot{y} + Ay = u \quad \text{in} \quad Q, \\ y = 0 \quad \text{on} \quad \partial\Omega, \\ y(0) = y_0. \end{cases} \tag{37}$$

It is well-known that the cost function J possesses a minimizer $\bar{u}$ in $L^2(Q)$, with a corresponding state $\bar{y} \in C([0,T];L^2(\Omega))$ and an adjoint state $\bar{\varphi} \in C([0,T];L^2(\Omega))$ satisfying

$$\begin{cases} -\dot{\bar{\varphi}} + A\bar{\varphi} = \bar{y} - y_d \quad \text{in} \quad Q, \\ \bar{\varphi} = 0 \quad \text{on} \quad \partial\Omega, \\ \bar{\varphi}(T) = 0, \end{cases} \tag{38}$$

$$\bar{u}(x,t) = \frac{-1}{\varepsilon}\bar{\varphi}(x,t) \quad \text{a.e. in } Q. \tag{39}$$

Assuming $y_d \in L^\infty(0,T;L^2(Q))$, we have $\bar{y} - y_d \in L^\infty(0,T;L^2(Q))$. According to Ladyženskaja et al., if the dimension N satisfies $N/4 < 1$, then $\bar{\varphi} \in L^\infty(Q)$. Consequently, $\bar{u} \in L^\infty(Q)$, and problem (PL) admits an optimal triple $(\bar{u},\bar{y},\bar{\varphi})$ satisfying (37), (38), and (39) for dimensions $N = 1,2,3$, as established by Casas.

In contrast, for the optimal control of a bilinear system, as discussed in Section 3, we have an optimal triple $(\bar{u},\bar{y},\bar{\varphi}) \in L^2(Q) \times C([0,T];L^2(\Omega))^2$ solving problem (P_A). To prove that $\bar{u}$ solves problem (P), we must show that $\bar{\varphi} \in L^\infty(Q)$.

The dynamics of $\bar{\varphi}$ are given by $-\dot{\bar{\varphi}} = A\bar{\varphi} + L^*(\bar{u}\bar{\varphi}) + (\bar{y} - y_d)$. For a given $\varphi \in L^\infty(Q)$, we have $L^*(\bar{u}\varphi) + (\bar{y} - y_d) \in L^2(Q)$. Again, by Ladyženskaja et al., a solution to (38) has the regularity $\bar{\varphi} \in L^\infty(Q)$ if $1/2 + N/4 < 1$. Therefore, we can solve problem (P) only in the one-dimensional case.

Our study has focused on optimal bilinear control with a basic differential operator (Laplacian) and a trivial external forcing f. However, the problem can be explored for a wider class of reaction-diffusion systems. Specifically, consider a differential operator:

$$Ay := -\sum_{i,j=1}^{N} \partial_{x_j} \left(a_{ij}(x)\partial_{x_i} y\right),$$

satisfying

$$a_{ij} \in L^\infty(\Omega) \quad \text{and there exists } m > 0 \text{ such that}$$
$$\sum_{i,j=1}^{N} a_{ij}(x)\xi_i\xi_j \geq m|\xi|^2, \quad \text{for all } \xi \in \mathbb{R}^N, \text{ for all } x \in \Omega,$$

and a Carathéodory function $f = f(x,t,y) : Q \times \mathbb{R}$ of class C^1 with respect to y, where $f(\cdot,\cdot,0) \in L^p(Q)$, satisfying

$$\text{there exists } C_f \in \mathbb{R} \text{ s.t. } \frac{\partial f}{\partial y}(x,t,y) \leq C_f, \quad \text{for all } y \in \mathbb{R},$$

$$\text{for all } M > 0, \text{ there exists } C_{f,M} > 0 \text{ s.t. } \left|\frac{\partial f}{\partial y}(x,t,y)\right| \leq C_{f,M}, \quad \text{for all } |y| \leq M.$$

We can then consider the evolution problem

$$\begin{cases} \dot{y} + Ay = uL(y) + f(x,t,y) \quad \text{in} \quad Q, \\ \frac{\partial y}{\partial \mathbf{n}} = 0 \quad \text{on} \quad \partial\Omega, \\ y(0) = y_0, \end{cases}$$

with the associated variational formulation:

$$\iint_Q \Big(-y\dot{v} + \sum_{i,j=1}^{N} a_{ij}(x)\partial_{x_i} y \partial_{x_j} v \Big) \mathrm{d}x\mathrm{d}t = \iint_Q \Big(uL(y)v + f(y)v \Big) \mathrm{d}x\mathrm{d}t + \int_\Omega y_0\, v(0)\, \mathrm{d}x.$$

Assuming the other conditions remain unchanged, we can follow the same approach to deduce that this state equation has a unique solution $y_u \in W(0,T) \cap C(\bar{Q})$. Referring to Tröltzsch [33, Ch.V] and Theorem 2, we can establish the existence of an optimal control. Furthermore, we can deduce

that for each solution $\bar{u}$, there exists a pair $\bar{y}, \bar{\varphi} \in W(0,T) \cap C(\bar{Q})$ satisfying the following system:

$$\begin{cases} \dot{\bar{y}} + A\bar{y} = \bar{u}L(\bar{y}) + f(\cdot,\cdot,\bar{y}) \quad \text{in} \quad Q, \\ \frac{\partial \bar{y}}{\partial \mathbf{n}} = 0 \quad \text{on} \quad \partial\Omega, \\ \bar{y}(0) = y_0, \end{cases}$$

$$\begin{cases} -\dot{\bar{\varphi}} + A^*\bar{\varphi} = L^*(\bar{u}\bar{\varphi}) + f_y(\cdot,\cdot,\bar{y})\bar{\varphi} + (\bar{y} - y_d) \text{ in } Q, \\ \frac{\partial \bar{\varphi}}{\partial \mathbf{n}} = 0 \quad \text{on} \quad \partial\Omega, \\ \bar{\varphi}(T) = 0, \end{cases}$$

$$\iint_Q \Big(L(\bar{y})\bar{\varphi} + \varepsilon \operatorname{sgn}(\bar{u}) |\bar{u}|^{p-1} \Big)(h - \bar{u}) \, \mathrm{d}x\mathrm{d}t \geq 0,$$

and

$$\bar{u}(x,t) = -\frac{\operatorname{sgn}(L(\bar{y})\bar{\varphi})}{\varepsilon^{\frac{1}{p-1}}} |L(\bar{y})\bar{\varphi}|^{\frac{1}{p-1}} \quad \text{a.e. in} \quad Q.$$

Here, $f_y(\cdot,\cdot,\bar{y})$ represents the derivative of f with respect to the second variable evaluated at $\bar{y}$.

These generalizations demonstrate the versatility of our approach, allowing for application to a wider class of reaction-diffusion systems while maintaining the key insights gained from the simpler model.

6 Conclusion

In this paper, we have presented a comprehensive study of the optimal control of reaction-diffusion systems subject to essentially bounded functions. Our primary findings include the proof of existence of optimal controls and the derivation of their first-order optimality conditions. We also illustrated the theoretical results with numerical examples for both one- and two-dimensional cases, demonstrating the applicability of our approach.

A key contribution of this work is the introduction of a novel formulation for the optimal control problem that does not require a priori bounds on the admissible controls. This approach effectively addresses a significant challenge in practical applications, where defining control bounds in advance is often difficult and potentially leads to suboptimal solutions. By working

in the $L^\infty(Q)$ space, we have established an equivalence between the original problem and an auxiliary optimization problem in $L^p(Q)$, proving that solutions to the auxiliary problem are inherently essentially bounded.

The results presented here contribute significantly to the theory of bilinear optimal control. They provide a robust framework for practical applications, particularly in scenarios where traditional bounded control approaches are limiting. Our findings are expected to have broad relevance, particularly in areas where control bounds are not easily determined.

This work also opens several promising directions for future research. One potential extension is to analyze the problem in an infinite horizon setting. Additionally, exploring the sensitivity of optimal control strategies to changes in system parameters or initial conditions could yield valuable insights. Investigating alternative cost functionals may offer the opportunity to tailor control characteristics for specific applications. Furthermore, the development of more efficient numerical methods to solve the optimality system could enhance the practical implementation of this approach. Lastly, this framework could be adapted and applied to other types of partial differential equations involving bilinear control terms, broadening its applicability across various domains.

Acknowledgements

Authors are grateful to the anonymous referees and the Editor for their valuable comments and suggestions to improve the quality of the paper.

References

[1] Allegretto, W. and Nistri, P. *Existence and optimal control for periodic parabolic equations with nonlocal terms*, IMA J. Math. Control Inf., 16(1) (1999), 43–58.

[2] Banerjee, M., Kuznetsov, M., Udovenko, O. and Volpert, V. *Nonlocal reaction–diffusion equations in biomedical applications*, Acta Biotheor., 70(2) (2022), 12.

[3] Biccari, U. and Hernández-Santamaría, V. *Null controllability of linear and semilinear nonlocal heat equations with an additive integral kernel*, SIAM J. Control Optim., 57(4) (2019), 2924–2938.

[4] Bradley, M.E. and Lenhart, S. *Bilinear optimal control of a Kirchhoff plate*, Syst. Control Lett., 22(1) (1994), 27–38.

[5] Breiten, T., Kunisch, K., and Pfeiffer, L. *Infinite-horizon bilinear optimal control problems: Sensitivity analysis and polynomial feedback laws*, SIAM J. Control Optim., 56(5) (2018), 3184–3214.

[6] Brezis, H. *Analyse fonctionnelle*, Théorie et applications (1983).

[7] Cannarsa, P., Floridia, G. and Khapalov, A.Y. *Multiplicative controllability for semilinear reaction–diffusion equations with finitely many changes of sign*, J. Math. Pures Appl., 108(4) (2017), 425–458.

[8] Casas, E. and Kunisch, K. *Infinite horizon optimal control problems for a class of semilinear parabolic equations*, SIAM J. Control Optim., 60(4) (2022), 2070–2094.

[9] Casas, E. and Kunisch, K. *Optimal control of semilinear parabolic equations with non-smooth pointwise-integral control constraints in time-space*, Appl. Math. Optim., 85(1) (2022), 12.

[10] Colbrook, M.J. *Computing spectral measures and spectral types*, Commun. Math. Phys., 384 (2021), 433–501.

[11] Corrêa Vianna Filho, A.L. and Guillén-González, F. *A Review on the Analysis and Optimal Control of Chemotaxis-Consumption Models*, SeMA J. (2024), 1–34.

[12] Daubechies, I., Defrise, M. and De Mol, C. *An iterative thresholding algorithm for linear inverse problems with a sparsity constraint*, Commun. Pure Appl. Math., 57(11) (2004), 1413–1457.

[13] El Boukhari, N. and Zerrik, E.H. *Optimal control of distributed semilinear systems with essentially bounded measurable control functions*, Int. J. Dyn. Control, 11(6) (2023), 2809–2819.

[14] El Kabouss, A. *Fractional Optimal Control Problem of Parabolic Bilinear Systems with Bounded Controls*, Vibration Control of Structures (2023).

[15] Evans, L.C. *Partial differential equations*, American Mathematical Society, 19 (2022).

[16] Fleig, A. and Guglielmi, R. *Optimal control of the Fokker–Planck equation with space-dependent controls*, J. Optim. Theory Appl., 174 (2017), 408–427.

[17] Furter, J. and Grinfeld, M. *Local vs. non-local interactions in population dynamics*, J. Math. Biol., 27 (1989), 65–80.

[18] Glowinski, R., Song, Y., Yuan, X., and Yue, H. *Bilinear optimal control of an advection-reaction-diffusion system*, SIAM Rev., 64(2) (2022), 392–421.

[19] Guillén-González, F. and Corrêa Vianna Filho, A.L. *An optimal control problem subject to strong solutions of chemotaxis-consumption models*, SIAM J. Control Optim., 61(5) (2023), 3156–3182.

[20] Hillen, T., Enderling, H. and Hahnfeldt, P. *The tumor growth paradox and immune system-mediated selection for cancer stem cells*, Bull. Math. Biol., 75 (2013), 161–184.

[21] Ladyzhenskaya, O.A. *Linear and quasilinear elliptic equations*, translated by Scripta Technica (1968).

[22] Li, X. and Yong, J. *Optimal control theory for infinite dimensional systems*, Springer Science & Business Media (2012).

[23] Liang, M. *Bilinear optimal control for a wave equation*, Math. Models Methods Appl. Sci., 9(01) (1999), 45–68.

[24] Lions, J.L. *Optimal control of systems governed by partial differential equations*, Springer, Berlin, 170 (1971).

[25] Martini, A., Müller, D. and Golo, S.N. *Spectral multipliers and wave equation for sub-Laplacians: lower regularity bounds of Euclidean type*, J. Eur. Math. Soc., 25(3) (2022), 785–843.

[26] Mophou, G., Kenne, C. and Warma, M. *Bilinear optimal control for a fractional diffusive equation*, J. Optim., (2024), 1–27.

[27] Ouzahra, M. *Approximate controllability of the semilinear reaction-diffusion equation governed by a multiplicative control*, Discrete Contin. Dyn. Syst. Ser. B, 27(2) (2022), 1075–1090.

[28] Pal, S. and Melnik, R. *Nonlocal Models in Biology and Life Sciences: Sources, Developments, and Applications*, arXiv preprint arXiv:2401.14651 (2024).

[29] Sikora, A. *On the $L^2 \to L^\infty$ norms of spectral multipliers of "quasi-homogeneous" operators on homogeneous groups*, Trans. Amer. Math. Soc., 351(9) (1999), 3743–3755.

[30] Sikora, A., Yan, L. and Yao, X. *Spectral multipliers, Bochner–Riesz means and uniform Sobolev inequalities for elliptic operators*, Int. Math. Res. Not., 2018(10) (2018), 3070–3121.

[31] Saporiti, R., Sinigaglia, C., Manzoni, A., and Braghin, F. *An optimal control strategy to design passive thermal cloaks of arbitrary shape*, Proc. R. Soc. A, 480(2294) (2024), 20230646.

[32] Sogge, C.D. *Eigenfunction and Bochner Riesz estimates on manifolds with boundary*, arXiv preprint math/0202032 (2002).

[33] Tröltzsch, F. *Optimal control of partial differential equations: theory, methods and applications*, American Mathematical Society, 112 (2010).

[34] Yahyaoui, S. and Ouzahra, M. *Quadratic optimal control and feedback stabilization of bilinear systems*, Optim. Control Appl. Methods, 42(4) (2021), 878–890.

[35] Yahyaoui, S. and Ouzahra, M. *Optimal control problem for unbounded bilinear systems and applications*, J. Control Decis. (2023), 1–13.

[36] Zerrik, E.H. and Boukhari, N.E. *Constrained optimal control for a class of semilinear infinite dimensional systems*, J. Dyn. Control Syst., 24 (2018), 65–81.

Iranian Journal of Numerical Analysis and Optimization
Vol. 15, No. 2, 2025, pp 600–624
https://doi.org/10.22067/ijnao.2025.89520.1499
https://ijnao.um.ac.ir/

Research Article

Multi-objective portfolio optimization using real coded genetic algorithm based support vector machines

B. Surja, L. Chin and F. Kusnadi*,

Abstract

Investors need to grasp how liquidity affects both risk and return in order to optimize their portfolio performance. There are three classes of stocks that accommodate those criteria: Liquid, high-yield, and less-risky. Classifying stocks help investors build portfolios that align with their risk profiles and investment goals, in which the model was constructed using the one-versus-one support vector machines method with a radial basis function kernel.

*Corresponding author

Received 21 August 2024; revised 28 Janurary 2025; accepted 23 February 2025

Bianca Surja
Center for Mathematics and Society, Department of Mathematics, Faculty of Science, Parahyangan Catholic University, Bandung, Indonesia. e-mail: bianca-surja78@gmail.com

Liem Chin
Center for Mathematics and Society, Department of Mathematics, Faculty of Science, Parahyangan Catholic University, Bandung, Indonesia. e-mail: chin@unpar.ac.id

Felivia Kusnadi
Center for Mathematics and Society, Department of Mathematics, Faculty of Science, Parahyangan Catholic University, Bandung, Indonesia. e-mail: felivia@unpar.ac.id

How to cite this article
Surja, B., Chin, L. and Kusnadi, F., Multi-objective portfolio optimization using real coded genetic algorithm based support vector machines. *Iran. J. Numer. Anal. Optim.*, 2025; 15(2): 600-624. https://doi.org/10.22067/ijnao.2025.89520.1499

This model was trained using a combination of the Kompas100 index and the Indonesian industrial sectors stocks data. Single optimal portfolios were created using the real coded genetic algorithm based on different sets of objectives: Maximizing short-term and long-term returns, maximizing liquidity, and minimizing risk. In conclusion, portfolios with a balance on all these four investment objectives yielded better results compared to those focused on partial objectives. Furthermore, our proposed method for selecting portfolios of top-performing stocks across all criteria outperformed the approach of choosing top stocks based on a single criterion.

AMS subject classifications (2020): Primary 65K10; Secondary 68T05, 62P05.

Keywords: Genetic algorithm; Liquidity; Multi-objective optimization; One-versus-one support vector machines; Radial basis functions.

1 Introduction

Portfolio optimization has long been a challenging, lucrative, and prominent research topic in investment management. Achieving the right balance between risk and return while maintaining liquidity has consistently been essential for effective portfolio selection [7, 3]. The stock market often overlooked changes in liquidity which can yield monthly returns (0.7-1.2%) in the short term [2, 6]. Santoso et al. [25] also suggested that liquidity can act as an early warning indicator for financial instability of an entity.

By taking all of these factors into account, selecting the top-performing stocks has always been a crucial task in successful portfolio optimization. Contemporary research utilized machine learning and artificial intelligence algorithms. Many researchers had noted that artificial neural network converged slowly. For example, finding the optimal number of neurons in hidden layers was often tiresome, as it involved trial-and-error [24].

Research by Huang [15] employed genetic algorithm (GA) for feature selection and parameter optimization. Subsequently, support vector regression (SVR) predicted and identified stocks with high returns serving as surrogates of actual returns, without considering liquidity. Similarly, Ma, Han, and Wang [20] predicted return of stocks by combining machine learning models,

namely random forest (RF) and SVR, and deep learning models, such as LSTM neural network, deep multilayer perceptron, and convolutional neural network, with mean-variance (MV) and omega portfolio optimization models. The result showed that the RF-MV optimization model performed the best. More recently, studies by Chen et al. [5], Guijarro [10], Jimbo et al. [17], and Kessaci [19] employed a GA for effective portfolio optimization method.

Gupta, Mehlawat, and Saxena [12] split stocks into three classes: Liquid, high-yield, and less-risky, by utilizing support vector machine (SVM) whose parameters were optimized by the grid-search method. Stocks in the dataset were first filtered for the highly liquid class, then categorized based on high yield, which indicated high returns. The remaining stocks with lower return and liquidity were classified as the residual category. After categorizing each stock into the three different classes, they selected the top-performing stocks in each class by using GA. The final portfolio consisted of all the portfolios created from the three classes.

The previous research studies have several limitations. First of all, liquidity was not considered along with risk and return. Secondly, the allocated number of stocks in the portfolio was known in advance, while in practice, the number of high-performing stocks was not known beforehand. Third of all, portfolios were assembled from each class and then combined to form the final portfolio. This approach may be less effective because the fixed number of stocks chosen for each portfolio might not represent the top-performing stocks in the overall selection. Lastly, the weights of the multi-objective model were not further explored to create several optimal portfolios which were tailored to different investment goals and risk profiles.

The aim of this paper is to demonstrate an extension of the multi-objective model conducted by Gupta et al. [11] by demonstrating how an optimized portfolio is immediately selected from the stocks dataset using real coded genetic algorithm (RCGA), in which it successfully handled economic problems such as portfolio selection and financial budget allocation [4, 23]. The number of stocks desired by an investor for a portfolio is fixed, and the quantity of stocks in each category is adjusted to fit this requirement. One of the classes may contain fewer stocks than the other two in order to create an optimized portfolio that enhanced return and liquidity while reducing risk. Further-

more, we examined different weights to develop various optimal portfolios to accommodate numerous investment objectives of return, liquidity, and risk.

Our work is explained as follows. Section 2 provides the brief introduction to the underlying methods employed in this paper. Section 3 describes the dataset and methodology design to conduct the research. Section 4 presents the results and analysis of the study, including a comparative analysis with the methodology put forward by Gupta et al. [11]. Finally, in Section 5, the conclusions are presented along with suggestions to improve for further research.

2 Overview

2.1 Support vector machines

The SVM method is a machine learning technique based on statistical learning theory, applying structural risk minimization to minimize model errors in generalizing data patterns [1]. SVM divides data into two classes, $+1$ and -1, using a decision boundary known as a hyperplane. Data points closest to the hyperplane that influence its orientation are called support vectors, and the distance between the support vectors is known as the margin. The hyperplane is a line in two-dimensional data, while it being a plane in three-dimensional data.

The primary goal of SVM is to find the optimal hyperplane to avoid misclassification. SVM seeks to maximize the margin so that the model has a large distance between the hyperplane and the support vectors. A model that perfectly separates the data without any data points within the margin and without misclassification errors is called a hard margin SVM. The SVM hyperplane optimization problem is

$$\min_{\mathbf{w},b} \frac{1}{2}||\mathbf{w}||^2$$
$$\text{subject to } y_i(\mathbf{w}\cdot\mathbf{x}_i + b) - 1 \geq 0, i = 1, 2, \ldots, m.$$

This concept is extended to soft margin SVM problems, where data cannot be perfectly separated by a hyperplane. The main objective is to find a

hyperplane that minimizes errors as much as possible [9]. To achieve this objective, slack variables, which allow data points to lie within the margin and account for some classification errors, are referred to as margin errors. The penalty, C, is used to control the model's error in classifying the data.

The SVM method was initially designed for binary classification problems, meaning it can only separate data into two classes. In cases where there are more than two classes, an approach to extend SVM is required. In this study, the multiclass SVM method used is one-versus-one (OVO) SVM, because OVO specifically addresses the differences between each class and has shown good results in previous research [14]. Moreover, its effectiveness has been proven in various economic issues such as stock selection and credit assessment [8, 16]. Suppose that a dataset has k classes. Then one-versus-one support vector machines (OVO SVM) will form $\frac{k}{2}(k-1)$ hyperplanes, where each hyperplane is built based on training data from a pair of classes.

Furthermore, for nonlinear data problems, SVM can utilize kernel functions to transform data into higher dimensions and construct non-linear hyperplanes. Some commonly used kernel functions in SVM are polynomial, radial basis function (RBF), and sigmoid. The kernel used in this study is the RBF kernel, as suggested by Yu, Wang, and Lai [28]. This decision is based on the consideration that the sigmoid kernel behaves similarly to the RBF kernel, while the polynomial kernel requires more time in the SVM training phase. SVM models with RBF kernels have also been proven to perform better than other kernels [11].

2.2 Real coded genetic algorithm

GAs are optimization methods inspired by the principles of natural selection and genetics in biological systems, used to find solutions for problems with infinitely many possible solutions [26]. Essentially, GA represent solutions as chromosomes composed of a series of genes, where each gene represents a variable of the proposed solution. The value of each gene is called an allele. In the search for the optimal solution, GAs operate with a set of chromosomes known as a population. The number of chromosomes in a

population (*popsize*) is a GA parameter that must be defined before running the algorithm. In this study, we utilize the RCGA, where alleles are real numbers.

To optimize the solution to an objective function $f(x)$ with q variables, RCGA first constructs a population consisting of n chromosomes. Each chromosome consists of q genes, with each allele assigned a random value. The quality of a chromosome in solving the given problem is evaluated using a fitness value, which indicates how well the chromosome addresses the problem. The primary goal of RCGA is to find the chromosome with the highest fitness value [21]. The initial population that is randomly generated, will go through a series of genetic operators to create a new, more optimal generation [13].

The genetic operations include elitism, selection to choose the best chromosomes as parent, crossover to produce offspring, and mutation to update genetic information. After going through a series of evolutionary processes, a new generation is formed, the fitness values are recalculated, and the RCGA cycle continues until the predetermined number of generations is reached. The optimal solution to the objective function is the chromosome with the best fitness value.

3 Data collection and problem formulation

3.1 Data collection

Based on research conducted by Gupta, Mehlawat, and Saxena [12], there are three classes that mainly represent the characteristics of each stock. Description of each class is as follows.

1. Liquid Class
   Stocks in this class have high liquidity but tend to have lower returns compared to the high-yield class.

2. High-yield Class
   This class includes stocks with high risk and high returns, thus having a high standard deviation. However, liquidity in this class is low.

3. Less-risky Class
   Stocks in this class do not fall into either the liquid or high-yield categories.

Thus, the information needed to build the OVO SVM model included long-term returns, risk, and liquidity. Long-term returns were measured through the average weekly returns over the past year. The choice of weekly returns, as opposed to daily or monthly, was based on considerations to avoid excessive daily fluctuations and minimized the risk of losing sensitivity to market changes. The risk of each stock was analyzed using the standard deviation of weekly returns over the past year. In the context of risk analysis in stock investment, the standard deviation of returns provided an overview of the distribution of possible returns over time. Next, stock liquidity was measured using the turnover rate to analyze the ability to buy or sell a stock without incurring significant losses [11]. The turnover rate was defined as the ratio between the average number of shares traded and the number of shares held by the public [27].

To train and evaluate the classification model, it was essential to have a large and representative sample of stocks to reflect the investment potential in the Indonesian stock market. This dataset was compiled from stocks listed in the Kompas100 index, along with the top ten performing stocks from each sectoral index in Indonesia, covering the period between January 1 and December 31, 2023. The stocks in the Kompas100 index are chosen for their high liquidity and market capitalization, while sectoral stocks represent diverse investment potentials in Indonesia.

The stocks used in this study were those with at least one year of availability and high quality. In the data processing stage, ELTY, MYRX, BHIT, BNBR, CASA, BLTA, and TAXI stocks were excluded from the analysis due to many rows of data with zero trading volume and returns. The dataset used in this research consisted of 137 selected stocks, consisting of sectoral stock data in Table 1 and stocks from the Kompas100 index. This dataset was then randomly divided into two sets using the train_test_split module from the sklearn.model_selection library in Python. The first set was the training set, which comprised 94 stocks from both the Kompas100 and sectoral indexes.

Table 1: Sectoral stocks excluding stocks in Kompas100 index.

Sectors	**Stock Ticker**
Energy	BYAN, DSSA, ITMG, TCPI
Industry	IMPC, ARNA, HEXA, MLIA, KBLI, SKRN
Technology	DCII, BELI, MTDL, MCAS, NFCX, WIRG
Property	MPRO, LPKR, MKPI, RISE
Health	CARE, PRDA, TSCP, SAME, BMHS
Finance	SMMA, MEGA, DNET
Transportation	TMAS, BIRD, SDMR, GIAA, HATM, ELPI, IMJS
Non-cyclic	PANI, HMSP, CMRY
Cyclic	MSIN, FILM, KPIG, BOGA

The second set, serving as the test set, consisted of the remaining 43 stocks to form a portfolio.

The first dataset was labeled using the third quartile of each criterion to avoid subjectivity in determining the high and low values of an evaluation metric. The choice of using quartiles is adjusted according to the characteristics and distribution of the data. Stocks with liquidity above the third quartile were first included in the liquid class with label 0, whereas the remaining stocks with risk above the third quartile were included in the high-yield class with label 1. Stocks that had not been labeled were included in the less-risky class. Seventy stocks from the first dataset that already had labels were then randomly selected to be training dataset, while the remaining 24 stocks were the validation dataset.

3.2 Multi-objective model

In this study, an optimal multi-criteria portfolio is constructed using a multi-objective model approach. This approach allows investors to consider multiple criteria simultaneously in their investment decision-making process without having to sacrifice one criterion for another [11].

The multi-objective model is based on the assumption that an investor intends to select h stocks from a set of 43 stocks in the testing dataset to construct a portfolio with minimal risk while maximizing short-term and long-term returns and ensuring high liquidity, using the entire available capital without engaging in short-selling. Let $R_{i,j}$ denote the return for the ith week of the jth stock. The short-term return of the jth stock, represented as R_j^{12}, is calculated based on the average of weekly returns over the last 12 weeks, while the long-term return of the jth stock, namely R_j^{52}, is computed based on the average weekly return over the past year. Finally, the liquidity of the jth stock, denoted as L_j, is calculated as the ratio of the average number of shares traded to the number of shares held by the public.

Assume that there are n stocks in the dataset, that h is the number of stocks in a portfolio, that x_j is the proportion of funds invested in stock j relative to the total funds, that z_j is a binary variable indicating whether stock j is selected (1) or not (0), and that l_j and u_j are the minimum and maximum percentage of funds invested in stock j, respectively. With the assumption of inability to short-sell, the multi-objective model is defined as follows:

$$\max\ f_1(x) = \sum_{j=1}^{n} R_j^{12} x_j, \tag{1}$$

$$\max\ f_2(x) = \sum_{j=1}^{n} R_j^{52} x_j, \tag{2}$$

$$\min\ f_3(x) = \sum_{i=1}^{52} \frac{|\sum_{j=1}^{n}(R_{i,j} - R_j^{52})x_j| + \sum_{j=1}^{n}(R_j^{52} - R_{i,j})x_j}{104}, \tag{3}$$

$$\max\ f_4(x) = \sum_{j=1}^{n} L_j x_j, \tag{4}$$

subject to

$$\sum_{j=1}^{n} x_j = 1,\ \sum_{j=1}^{n} z_j = h,\ x_j \leqslant u_j z_j,\ x_j \geqslant l_j z_j,\ u_j \geqslant 0,\ z_j \in \{0,1\},$$

for $j = 1, 2, \ldots, n$.

This multi-objective model is used to construct a portfolio considering four different criteria, resulting in four objective functions that need to be

optimized according to the investor's preferences. The first objective function (1) aims to maximize short-term returns, which in turn provides an overview of the stock's performance in the short term and is used to evaluate opportunities for quick gains. The model also aims to maximize long-term returns through the second objective function (2), which overviews the stock's performance over a longer period and reflects the stability and growth of the investment.

To avoid the risks arising from stock price fluctuations, the third objective function (3), which utilizes mean semi-absolute deviation, is built with the aim of minimizing risk. In this model, risk considers the difference between the actual weekly returns and the expected long-term returns over a one-year evaluation period. The final objective function (4) aims to maximize liquidity. Analyzing liquidity in a portfolio is important for investors to get an idea of the ability to trade stocks without incurring significant losses. The portfolio is constructed by allocating all investment funds into the selected stocks. To ensure sufficient stock diversification and avoiding inefficient investments, h stocks are selected and indicated by binary variables. Furthermore, upper and lower investment limits were set to restrict the proportion of funds allocated to a single stock.

3.3 Constructing multi-criteria portfolio using RCGA

To simulate the RCGA method, we used the Python software. The first step is population initialization, where each chromosome represents a portfolio. Each gene represents a single stock, while the allele on the gene represents the proportion of that stock in the portfolio. Suppose that an investor wants to form a portfolio based on data containing n stocks. In this study, the number of stocks in the portfolio (h) is randomly determined during the initialization stage, and it is assumed that l and u are the minimum and maximum percentage of the funds invested in each stock, respectively.

First, the investor needs to determine the number of chromosomes in the population (*popsize*) as a limit to the solution search space. Each chromosome constructed has n genes, where the first gene of a chromosome represents the

first stock in the data table, the second gene represents the second stock, and so on. Initially, every gene in each chromosome is assigned an allele value of zero. The population initialization stage in this study was the following Algorithm 1.

Algorithm 1 Population initialization

Input: number of stocks in the population (*popsize*), number of stocks in the data (n), lower investment bound (l), and upper investment bound (u).

1: Initialize the population as an empty array to be filled with chromosomes.
2: For each chromosome in *popsize*, do:

- Randomly determine the number of stocks (h) between zero and n.
- Initialize a chromosome as an array of length n with zero values.
- Initialize *selected_gens* as a list of h stocks chosen randomly, ensuring no stock is selected more than once.
- For each selected gene, initialize the allele randomly between l and u.
- Add the new chromosome to the population.

Output: Initial Population.

To evaluate the quality of the formed chromosomes, RCGA enters the evaluation stage. In the context of portfolio formation, the evaluation function is used to assess how well a portfolio meets the established investment criteria. To ensure that the generated chromosomes not exceeding the available capital budget, a penalty function P is applied. This function reduces the fitness value of chromosomes that exceed the budget limit. Chromosomes that produce portfolios with costs exceeding the limit will have lower fitness values. Let x_j be the proportion of stock j. The fifth objective function is

$$f_5(x) = \left| \sum_{j=1}^{n} x_j - 1 \right|,$$

so the penalty function P is defined as

$$P = \begin{cases} 10^4 \cdot f_5(x), & \text{if } f_5(x) \geqslant 10^{-5}, \\ 0, & \text{if } f_5(x) < 10^{-5}. \end{cases}$$

The fitness function for chromosome k, with $k = 1, 2, \ldots, popsize$ with $popsize$ being the number of chromosomes in one population, is defined as

$$fit_k = w_1 \cdot f_1(x) + w_2 \cdot f_2(x) - w_3 \cdot f_3(x) + w_4 \cdot f_4(x) - P, \tag{5}$$

where w_1, w_2, w_3, and w_4 are the weights of the objective functions with $\sum_{i=1}^{4} w_i = 1$ [11]. The weights assigned to the objective functions indicate their importance in building the portfolio. High fitness value of chromosome k indicates how well the solution it provided performs relative to the objective functions. Algorithm 2 simulated the fitness evaluation stage in this study.

Algorithm 2 Fitness evaluation

Input: population, stock risk, long return, short return, liquidity, weights w_1, w_2, w_3, w_4.

1: Initialize $fitness\_results$ as an empty array to be filled with the fitness values of each chromosome.

2: For each chromosome in the population, do:

- Calculate *short_return_value* using equation (1).
- Calculate *return_value* using equation (2).
- Calculate *risk* using equation (3).
- Calculate *liquidity_value* using equation (4).
- Calculate penalty as the absolute value of the difference between the total investment allocation and one.
- If $penalty > 10^{-5}$, then $penalty$ is calculated as $10^4 \times$ penalty, otherwise $penalty = 0$.
- Calculate *fitness_value* using (5).
- Add the fitness value to *fitness_results*.

Output: Fitness evaluation results for each chromosome in the population.

The chromosome with the highest fitness value represents the portfolio with the stock combination that best aligns with the investor's preferences to maximize return and liquidity while minimizing risk. This chromosome is selected as the elite and is directly passed on to the next generation without undergoing evolutionary processes such as crossover and mutation.

Next, the roulette wheel selection operation is performed to select candidate portfolios that best match the investor's preferences as parents. This method works by assigning a selection probability to each chromosome based on its fitness value. Before calculating the selection probabilities, the fitness values of each chromosome are normalized. Normalization ensures that all fitness values fall within a positive range, as fitness values can have both positive and negative ranges. This is done by subtracting the smallest fitness value in the population from the fitness value of each chromosome, then adding a small value such as 10^{-10} to avoid division by zero. The higher the fitness value of a chromosome, the more likely it is to be selected as a parent. The selection stage of RCGA in this study is shown in Algorithm 3.

Algorithm 3 Roulette wheel selection

Input: population, fitness values.

1: initialize $min\_fitness$ with the smallest value from all fitness values in the population.
2: calculate $normalized\_fitness$ by subtracting $min\_fitness$ from the $fitness\_values$ and adding 10^{-10}.
3: calculate $total\_fitness$ by summing all values in $normalized\_fitness$.
4: calculate $probabilities$ for each chromosome by dividing $normalized\_fitness$ by $total\_fitness$.
5: $selected\_index$ is chosen randomly from the population using the previously calculated $probabilities$.

Output: index of the selected chromosome.

The chromosomes selected as parents through the roulette wheel selection operation is paired to create new offspring that combine characteristics from both parents. The offspring are generated using the crossover operation, which swaps portions of genetic information between parent chromosomes.

The next step in this research is Algorithm 4, which denotes the crossover stage of RCGA.

Algorithm 4 *Crossover*

Input: *parent1*, *parent2*, *crossover_rate* (p_c).

1: Initialize r_c randomly with a value between zero and one.

2: If $r_c > p_c$ then $parent1 = child1$ and $parent2 = child2$, if $r_c < p_c$ proceed with:

- Initialize crossover point (C_p) with a value between one and the length of $parent1$.
- $child1$ is created by combining the initial part of $parent1$ with the latter part of $parent2$ after C_p.
- $child2$ is created by combining the initial part of $parent2$ with the latter part of $parent1$ after C_p.

Output: *child1* and *child2*.

The newly formed offspring through the crossover operation will have some of their information updated using the mutation operation, which is performed using Algorithm 5. Next, RCGA will continuously perform the stages of evaluation, selection, crossover, and mutation until the generation limit set by the investor is reached. The chromosome with the best fitness value from the final generation is considered the portfolio that best aligns with the investor's preferences.

4 Result and analysis

The initial phase of developing the OVO SVM model with the RBF kernel was conducted in Python using the SVC module to build the classification model and the OVO classifier module to handle multi-class problems with an OVO approach. Both modules are sourced from the scikit-learn library. The OVO SVM model has two parameters: the penalty parameter C, which controls the classification error, and the parameter γ, which determines the maximum distance of data points that influence the hyperplane. In the initial

Algorithm 5 Mutation

Input: chromosome, *mutation_rate* (p_m), l, u

1: Initialize *mutated_chromosome* as a copy of the original chromosome.
2: Initialize r_m randomly with a value between zero and one.
3: If $r_m < p_m$ then proceed with:

- Determine *num_mutations* randomly with a value between one and the length of the chromosome.
- Initialize *mutation_indices* as a list of chromosome (gene) indices randomly determined up to *num_mutations*.
- For each selected gene, if the allele value is not zero, the allele value is randomly changed between l and u, while if the allele value is zero, the allele value will not be changed.

Output: *mutated_chromosome*

model, where the parameters C and γ were not explicitly set, the accuracy was only 0.54. The model struggled to recognize data belonging to the high-yield class, as indicated by its very low precision and accuracy for that class.

Parameter optimization was performed to find a combination that could enhance the model's performance using grid-search and GA-SVM. The grid-search method revealed that the best parameter combination for the combined data was $C = 10$ and $\gamma = 0.1$. The OVO SVM model produced with these parameters achieved an accuracy of 0.75 and showed significant overall performance improvement, though it still faced challenges with high-yield class data. The results from the GA-SVM method indicated that the best parameter combination for the data used was $C = 403.4645$ and $\gamma = 0.1242$, with an accuracy of 0.88. Based on the balanced evaluation metrics across each class, this model was the most optimal and thus applied to the test stock data to construct the portfolio.

The OVO SVM model developed using the GA-SVM method classified 15 stocks into the liquid and less-risky classes. Additionally, 13 stocks were classified into the high-yield class. Table 2 presents the average values of each criterion for each class. The liquid class consisted of stocks with high liquidity, as reflected by an average liquidity of 3.1147% for this class. This

classification indicated that a significant portion of the stocks in the test dataset has adequate liquidity. The high-yield class, as the name suggested, had the highest average risk of 7.2721% with a long-term average return of 0.8337%, signaling strong investment potential. On the other hand, the less-risky class had an average risk of 3.7069%, which was lower compared to the other two classes, though it did not offer high returns.

Table 2: Average criterion for each class.

Class	**Long return**	**Risk**	**Liquidity**
Liquid	0.180615	5.247010	3.114714
High yield	0.833655	7.272134	1.897383
Less risky	−0.422793	3.706849	1.901513

From the classified testing dataset, the stocks that best align with the investor's preferences are selected to form a portfolio using a multi-objective model solved by RCGA. During the search process, the population size (*pop-size*) is set at 50, and the number of generations is limited to 1,000 to constrain the search space for the optimal solution. To ensure stock diversification, a crossover probability of 0.3 and a mutation probability of 0.2 are specified. Additionally, the minimum and maximum percentage of funds invested in a single stock are set between 0 and 0.2 inclusive. The weights for each objective function are adjusted according to the investor's preferences for short-term return, long-term return, risk, and liquidity criteria.

In this study, nine portfolios were constructed to understand the impact of each criterion on portfolio performance. These nine portfolios were built under several scenarios, such as balancing all four criteria, focusing on a single criterion, and building portfolios formed for extreme cases by ignoring one criterion. Portfolio 1 is constructed under the assumption that the investor considers all criteria equally important, thus the weights for the objective functions in the multi-objective model are set equally at 0.25.

Out of the total 43 stocks in the combined testing data, RCGA allocated investment funds into 11 selected stocks, comprising two stocks from the liquid class, six stocks from the high-yield class, and three stocks from the less-risky class. The distribution of investment funds in Portfolio 1 is shown

in Figure 1. The largest fund allocation is given to the TPIA stock from the high-yield class, while the smallest allocation is given to the HEXA stock from the liquid class. The expected weekly return for this portfolio is relatively balanced for both the long term and short term, at 0.6683% and 0.7867%, respectively. Furthermore, Portfolio 1 has risk level of 67.78% and liquidity level of 1.81%.

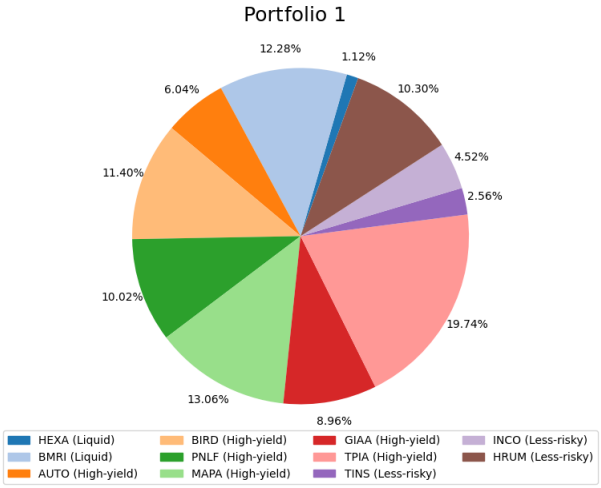


Figure 1: Stocks allocation in Portfolio 1.

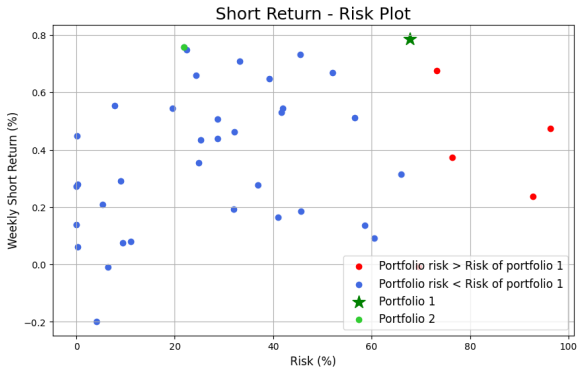


Figure 2: Plot of short return by risk of all optimal portfolios.

To examine the relationship between risk and return, both in the short-term and long-term, steps similar to those in the previous analysis are taken.

Portfolio 2 is built as a modification of equation (5). In this construction, the first objective function f_1 is set to be a constraint, that is, $f_1(x) = r_{sr}$, with $r_{sr} \in \{0.01, 0.02, 0.03, \ldots, 0.80\}$, while the other objective functions still remained the same. Thus, the weights of w_2, w_3, and w_4 are equal to $\frac{1}{3}$. The following is the fitness function of Portfolio 2

$$fit_k = w_2 \cdot f_2(x) - w_3 \cdot f_3(x) + w_4 \cdot f_4(x) - P_1,$$

where $P_1 = P + P_{sr}$, with $P_{sr} = |f_1(x) - r_{sr}| \cdot 10^4$.

Figure 2 illustrates the relationship between risk and short-term return, obtained by first determining the short-term return target and then constructing a portfolio that balances all other criteria. These portfolios are generated by first setting the short-term return target, after which the RCGA is executed to maximize liquidity and long-term return while minimizing risk in a balanced manner.

Portfolio 1 has the highest short-term return with medium risk, similar to Portfolio 2 which is also set with equal weights for all objective functions. This means that high risk does not guarantee high short-term return. The red circles represent portfolios with higher risk than Portfolio 1 but with lower short-term returns. There is one green circle labeled Portfolio 2 with a relatively high short-term return expectation and lower risk compared to Portfolio 1. Figure 3 provides an illustration of the relationship between long-term return and risk. The mapped portfolios do not form a diagonal line because Portfolio 1 has a higher long-term return expectation. Both Figures 2 and 3 show that Portfolio 1 is similarly optimal.

Portfolio 1 was constructed under the assumption that investors balance all four criteria, but in reality, some investors tend to lean more heavily towards one criterion over others. Table 3 summarizes the weights used to build Portfolios 3, 4, and 5. Portfolio 3 was constructed with a primary focus on return, assuming that the investor readily takes on high risk or low liquidity. Portfolio 4 was built with an emphasis on risk, where the investor seeks stability while disregarding return and liquidity. Portfolio 5 was designed to seek out stocks with high liquidity.

Additionally, portfolios are created for extreme scenarios where the investor intentionally ignores one criterion in the portfolio construction process.

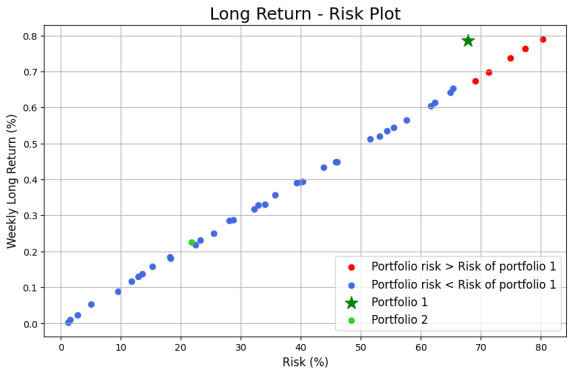


Figure 3: Plot of long return against risk of all optimal portfolios.

Table 3: Weights of Portfolios 3, 4, and 5.

Portfolio	$\mathbf{w_1}$	$\mathbf{w_2}$	$\mathbf{w_3}$	$\mathbf{w_4}$
3	0.3	0.4	0.15	0.15
4	0.15	0.15	0.6	0.1
5	0.1	0.1	0.1	0.7

This approach is taken to understand the impact of a particular criterion on portfolio performance. Portfolio 6 was formed under the assumption that the investor does not consider risk at all. Portfolio 7 is designed for investors who seek quick gains, while disregarding long-term returns. Portfolio 8 is constructed by omitting short-term return considerations entirely. The final extreme case is Portfolio 9, where the investor does not consider liquidity during the portfolio construction process. The weights used to build these three portfolios are summarized in Table 4.

Table 4: Weights of Portfolios 6, 7, 8, and 9.

Portfolio	$\mathbf{w_1}$	$\mathbf{w_2}$	$\mathbf{w_3}$	$\mathbf{w_4}$
6	0.4	0.4	0	0.2
7	0.8	0	0.1	0.1
8	0	0.8	0.1	0.1
9	0.3	0.3	0.4	0

In Figure 4, it is observed that the portfolio with the highest risk is Portfolio 6, which was constructed by ignoring the risk criterion. Although this portfolio does have the highest expected long-term return, there are several other alternative portfolios that offer higher short-term return expectations with lower risk levels. Portfolio 9, which disregards liquidity considerations, is expected to have a negative short-term return and a low long-term return, though it does feature low risk.

For comparison purposes, the portfolio selection method proposed by Gupta et al. [11] was applied. The dataset consisting of 43 testing stocks was categorized into three distinct portfolios by GA-SVM: one with 15 liquid stocks, another with 13 high-yield stocks, and the final one containing 15 less-risky stocks. The RCGA method was used for each portfolio to identify the top-performing stocks, applying weights that prioritized specific criteria, as detailed in Table 5. Portfolio A, which emphasized liquidity, selected 9 out of the 15 liquid stocks. Portfolio B, focused on returns, included 12 of the 13 high-yield stocks. Lastly, Portfolio C, which prioritized risk, chose 10 of the 15 less-risky stocks.

Table 5: Weights of Portfolios A, B, and C.

Portfolio	$\mathbf{w_1}$	$\mathbf{w_2}$	$\mathbf{w_3}$	$\mathbf{w_4}$
A	0.2	0.25	0.2	0.35
B	0.3	0.35	0.2	0.15
C	0.17	0.23	0.45	0.15

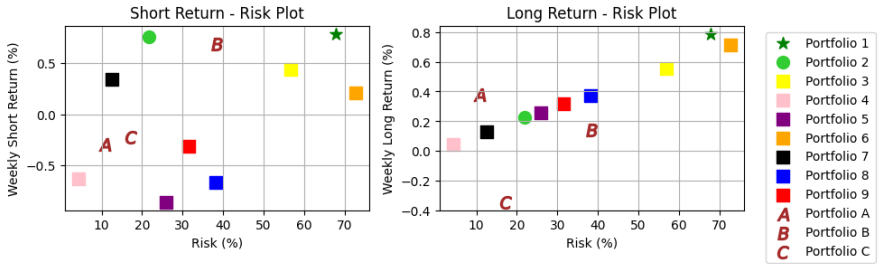


Figure 4: Plot of return compared to risk for all optimal portfolios.

Figure 4 evaluates the performance of all portfolios, including the nine portfolios using our proposed methodology and the three portfolios based on the methodology developed by Gupta et al. [11]. Similar to the previous analysis, Portfolio 1, which is built by balancing all the criteria, has the highest short-term return expectation with risk level above 50%. This portfolio also has a high long-term return expectation, making it a good choice for risk-taker investors. For moderate investors, forming a portfolio by setting a short-term return target and balancing all criteria, as in Portfolio 2, is a good alternative. Furthermore, for conventional investors who avoid high risk, a portfolio could be constructed by ignoring the long-term return criterion, as in Portfolio 7. While Portfolio A has higher long-term return and lower risk compared to Portfolios 1 and 2, its projected short-term return indicates a negative outcome which is lower than Portfolio 7. Portfolio B exhibits a lower long-term return, but a higher short-term return compared to Portfolio 8. Nonetheless, its short-term return is still less than that of Portfolio 2, which also carries a lower risk. Portfolio C is not advisable, as both its short-term and long-term returns are expected to be negative. The other portfolios are considered suboptimal due to the availability of alternative portfolios that offer higher returns for similar risk levels.

5 Conclusions

By classifying Indonesian stocks into three classes, it is concluded that the OVO SVM classification model optimized using the GA-SVM method is better than the model optimized using the grid-search method. The OVO SVM model classifies 13 stocks into the high-yield class, reflecting the growth potential and opportunities for satisfactory returns in the Indonesian stock market. The Indonesian stock market also shows adequate liquidity, with 15 stocks in the liquid class. Additionally, there are 15 stocks with minimal risk in the less-risky class, providing options for investors. Portfolios that balance the four investment criteria have high gains in both short-term and long-term. Investors aiming to boost returns might consider adding more high-yield stocks, whereas those wanting to lower risk could invest in stocks from a less-risky category. It is also observed that by including liquidity cri-

teria, the long-term return performance is enhanced, as shown in Portfolio A. This indicates that adding the liquidity criteria is crucial for optimizing portfolio performance over the long term.

The portfolios in this study were constructed using RCGA to solve the multi-objective model. One characteristic of RCGA is the population initiation stage, which involves randomly generating chromosomes and their allele values, leading to different portfolio compositions each time RCGA is run. This results in inconsistent outcomes, although the fitness values obtained tend to be around the same range. The computational time of RCGA also depends on the number of chromosomes in the population, the number of generations, and the complexity of the fitness function. Future research could explore exact optimization approaches, such as CPLEX suggested by Mutunge and Haugland [22], Jin, Qu, and Atkin [18], and Zhang [29, 30, 31] or data envelopment analysis as suggested by Zhou et al. [32, 33], to solve multi-objective problems and achieve more consistent solutions.

References

[1] Awad, M., and Khanna, R. *Efficient learning machines: Theories, concepts, and applications for engineers and system designers*. Springer Nature, New York, 2015.

[2] Bali, T.G., Peng, L., Shen, Y., and Tang, Y. *Liquidity shocks and stock market reactions,* Rev. Financ. Stud. 27(5) (2014) 1434–1485.

[3] Beardsley, X.W., Field, B., and Xiao, M. *Mean-variance-skewness-kurtosis portfolio optimization with return and liquidity,* Commun. Math. Financ. 1(1) (2012) 13–49.

[4] Chen, J.S., Hou, J.L., Wu, S.M., and Chang-Chien, Y.W. *Constructing investment strategy portfolios by combination genetic algorithms,* Expert Syst. Appl. 36(2) (2009) 3824–3828.

[5] Chen, W., Wang, Y., Gupta, P., and Mehlawat, M.K. *A novel hybrid heuristic algorithm for a new uncertain mean-variance-skewness portfolio selection model with real constraints,* Appl. Intell. 48 (2018) 2996–3018.

[6] Chordia, T., Roll, R., and Subrahmanyam, A. *Liquidity and market efficiency,* Journal of Financial Economics, 87(2) (2008) 249–268.

[7] Fabozzi, F.J., Gupta, F., and Markowitz, H.M. *The legacy of modern portfolio theory,* J. Invest. 11(3) (2002) 7–22.

[8] Fan, A., and Palaniswami, M. *Stock selection using support vector machines,* In International Joint Conference on Neural Networks. Proceedings, (Cat. No. 01CH37222) (2001) 1793–1798.

[9] Flach, P.*Machine learning: The art and science of algorithms that make sense of data*. Cambridge University Press, Cambridge, 2012.

[10] Guijarro, F. *A similarity measure for the cardinality constrained frontier in the mean–variance optimization model,* J. Oper. Res. Soc. 69(6) (2018) 928–945.

[11] Gupta, P., Mehlawat, M.K., Inuiguchi, M., and Chandra, S. *Multicriteria portfolio optimization using support vector machines and genetic algorithms,* Studies in Fuzziness and Soft Computing, vol 316. (pp. 283–309) Springer, Berlin, Heidelberg, 2014.

[12] Gupta, P., Mehlawat, M.K., and Saxena, A. *A hybrid approach to asset allocation with simultaneous consideration of suitability and optimality,* Inf. Sci. 180(11) (2010) 2264–2285.

[13] Herrera, F., Lozano, M., and Verdegay, J.L. *Tackling real-coded genetic algorithms: Operators and tools for behavioural analysis,* Artif. Intell. Rev. 12 (1998) 265–319.

[14] Hsu, C.W., and Lin, C.J. *A comparison of methods for multiclass support vector machines,* IEEE Trans. Neural Networks Learn. Syst. 13(2) (2002) 415–425.

[15] Huang, C.F. *A hybrid stock selection model using genetic algorithms and support vector regression,* Appl. Soft Comput. 12(2) (2012) 807–818.

[16] Huang, Z., Chen, H., Hsu, C.J., Chen, W.H., and Wu, S. *Credit rating analysis with support vector machines and neural networks: A market comparative study,* Decis. Support Syst. 37(4) (2004) 543–558.

[17] Jimbo, H.C., Ngongo, I.S., Andjiga, N.G., Suzuki, T., and Onana, C.A. *Portfolio optimization under cardinality constraints: A comparative study,* Open J. Stat. 7(4) (2017) 731–742.

[18] Jin, Y., Qu, R., and Atkin, J. *Constrained portfolio optimisation: The state-of-the-art markowitz models,* In International Conference on Operations Research and Enterprise Systems (2016) 388–395.

[19] Kessaci, Y. *A multi-objective continuous genetic algorithm for financial portfolio optimization problem,* In Proceedings of the Genetic and Evolutionary Computation Conference Companion (2017) 151–152.

[20] Ma, Y., Han, R., and Wang, W. *Portfolio optimization with return prediction using deep learning and machine learning,* Expert Syst. Appl. 165 (2021) 113973.

[21] Mitchell, M. An introduction to genetic algorithms. MIT Press, Cambridge, 1998.

[22] Mutunge, P., and Haugland, D. *Minimizing the tracking error of cardinality constrained portfolios,* Computers and Operations Research, 90 (2018) 33–41.

[23] Packard, N.H. *A genetic learning algorithm for the analysis of complex data,* Complex Syst. 4(5) (1990) 543–572.

[24] Rather, A.M., Sastry, V.N., and Agarwal, A. *Stock market prediction and portfolio selection models: A survey,* Opsearch, 54 (2017) 558–579.

[25] Santoso, W., Harun, C.A., Hidayat, T., and Wonida, H. *Market liquidity risk as an indicator of financial stability: The case of Indonesia,* Bank of Indonesia, (2010). 1–22.

[26] Sheppard, C. *Genetic algorithms with python.* Smashwords Edition, Austin, TX, USA 2017.

[27] Vogel, H.L. *Entertainment industry economics: A guide for financial analysis.* Cambridge University Press, Cambridge, 2020 .

[28] Yu, L., Wang, S., and Lai, K.K. *Neural network-based mean–variance–skewness model for portfolio selection,* Comput. Oper. Res. 35(1) (2008) 34–46.

[29] Zhang, P. *Multi-period possibilistic mean semivariance portfolio selection with cardinality constraints and its algorithm,* J. Math. Model. Algor. Oper. Res. 14 (2015) 239–253.

[30] Zhang, P. *An interval mean–average absolute deviation model for multiperiod portfolio selection with risk control and cardinality constraints,* Soft Comput. 20 (2016) 1203–1212.

[31] Zhang, P. *Multiperiod mean absolute deviation uncertain portfolio selection,* Ind. Eng. Manag. Syst. 15(1) (2016) 63–76.

[32] Zhou, Z., Jin, Q., Xiao, H., Wu, Q., and Liu, W. *Estimation of cardinality constrained portfolio efficiency via segmented DEA,* Omega, 76 (2018) 28–37.

[33] Zhou, Z., Liu, X., Xiao, H., Wu, S., and Liu, Y. *A DEA-based MOEA/D algorithm for portfolio optimization,* Clust. Comput. 22(Suppl 6) (2019) 14477–14486.

Iranian Journal of Numerical Analysis and Optimization
Vol. 15, No. 2, 2025, pp 625–644
https://doi.org/10.22067/ijnao.2025.89967.1523
https://ijnao.um.ac.ir/

Research Article

An accurate numerical technique for solving a special case of fractional differential equations using the Khalouta transform of two different fractional derivatives

A. Khalouta*

Abstract

The aim of this paper is to present a novel coupling approach of the Khalouta transform method and the homotopy perturbation method in order to obtain an accurate and efficient method for solving a special case of fractional differential equations involving Caputo and Caputo-Fabrizio fractional derivatives. This method is called the fractional Khalouta homotopy perturbation method (FKHHPM). In particular, the FKHHPM is used to obtain a solution to the fractional reaction-diffusion-convection

*Corresponding author

Received 24 September 2024; revised 10 December 2024; accepted 2 March 2025

Ali Khalouta
Laboratory of Fundamental and Numerical Mathematics, Department of Mathematics, Faculty of Sciences, Setif 1 University-Ferhat ABBAS, Algeria.
e-mail: nadjibkh@yahoo.fr ; ali.khalouta@univ-setif.dz

How to cite this article
Khalouta, A., An accurate numerical technique for solving a special case of fractional differential equations using the Khalouta transform of two different fractional derivatives. *Iran. J. Numer. Anal. Optim.*, 2025; 15(2): 625-644.
https://doi.org/10.22067/ijnao.2025.89967.1523

equations. The convergence analysis and a numerical example are presented. To evaluate the effectiveness of the proposed computational strategy, we examine the convergence of the series solution over different fractional values and evaluate the behavior of the solution as the time domain increases. The efficiency and originality of the FKHHPM are demonstrated by calculating the absolute error. This work is supported by two-dimensional and three-dimensional graphical representations made in accordance with MATLAB.

AMS subject classifications (2020): 35R11, 26A33, 34A45, 74G15.

Keywords: Fractional differential equation; Caputo fractional derivative; Caputo-Fabrizio fractional derivative; Khalouta transform; Homotopy perturbation method.

1 Introduction

Fractional calculus is an important mathematical field in the modeling of many natural and physical phenomena. Dating back to Leibniz's letter in the late 17th century, studies on fractional calculus have been constantly improving with some increase in the number in recent decades. The Riemann–Liouville, Grünwald–Letnikov, and Caputo fractional derivative definitions have been widely studied in the literature [20, 24, 26]. More recent definitions, such as the Atangana–Baleanu fractional derivative and the Caputo–Fabrizio fractional derivative, are also popular tools in modeling studies. Katugampola, Hilfer, and many other derivatives have been given in the literature [4, 6, 14, 15] along with conformable fractional derivative, which has become the basis of conformable calculus [1].

Fractional differential equations have been widely used to describe several phenomena in physics and help in their solutions, such as fluid mechanics, viscoelasticity, electrodynamics, quantum mechanics, rheology, damping laws, diffusion processes, economy, biology, geophysics, bioengineering, optimal control, and other areas of science; see [5, 8, 12, 13, 23, 27]. It is very difficult to obtain exact solutions to real-life problems using fractional differential equations, and usually complex mathematical techniques are required.

Recently, many mathematicians and physicists have studied the solutions to fractional differential equations involving different types of fractional derivatives using various analytical and numerical methods. For example, Kumawat et al. [21] treated the solutions of fractional differential equations using both Riemann–Liouville and Caputo fractional derivatives and used the Khalouta transform method. Alazman et al. [2] employed the Laplace decomposition technique to obtain the numerical solutions of the SIR model using the generalized fractional derivative. Yadav et al. [30] applied a precise and analytical method, namely the Shehu transform decomposition method, to examine the solutions of multi-dimensional fractional diffusion equations involving the Caputo–Fabrizio derivative. Dube, Mishraa, and Goswami [9] introduced the local fractional homotopy perturbation Sumudu transform method to solve the oxygen diffusion equation with a local fractional derivative. The author [17] presented two new semi-analytical methods called the Khalouta homotopy perturbation method and the Khalouta variational iteration method to find new approximate analytical solutions of nonlinear fractional partial differential equations via the Atangana–Baleanu fractional derivative. Also, the author [18] used a modification of a new general integral transform to study the properties of solutions of fractional differential equations containing a Caputo generalized fractional derivative. This method is called the ρ-Jafari transform method.

The main objective of the present paper is to propose an accurate numerical technique called the fractional Khalouta homotopy perturbation method (FKHHPM) for solving a special case of fractional differential equations, in particular, fractional reaction-diffusion-convection equations using Caputo and Caputo–Fabrizio derivatives of the forms

$$D_{\upsilon}^{\kappa}\chi = (\Psi(\chi)\chi_{\varrho})_{\varrho} + \Phi(\chi)\chi_{\varrho} + \Pi(\chi), \tag{1}$$

subject to the condition

$$\chi(\varrho, 0) = \chi_0(\varrho), \tag{2}$$

where D_{υ}^{κ} is the fractional derivative of order $0 < \kappa \leq 1$ in the Caputo sense.

Moreover,

$$\mathcal{D}_{\upsilon}^{(\kappa)}\chi = (\Psi(\chi)\chi_{\varrho})_{\varrho} + \Phi(\chi)\chi_{\varrho} + \Pi(\chi), \tag{3}$$

subject to the condition

$$\chi(\varrho, 0) = \chi_0(\varrho), \tag{4}$$

where $\mathcal{D}_{\upsilon}^{(\kappa)}$ is is the fractional derivative of order $0 < \kappa \leq 1$ in the Caputo–Fabrizio sense.

In (1) and (3), $\chi = \{\chi(\varrho, \upsilon), \varrho \in \mathbb{R}, \upsilon \geq 0\}$ is an unknown function, where ϱ represents the space variable and υ represents the time variable, and the arbitrary smooth functions $\Psi(\chi), \Phi(\chi)$, and $\Pi(\chi)$ denote the diffusion term, the convection term, and the reaction term, respectively. The reaction-diffusion-convection problems is a very useful mathematical model in applied sciences such as biology modeling, physics, chemistry, astrophysics, hydrology, medicine, and engineering [3, 7, 10, 29].

The FKHHPM is used to find an approximate solution to the given equations. FKHHPM is a mixture of the Khalouta transform and the homotopy perturbation method. The Khalouta transform is an analytical method that provides exact solutions, while the homotopy perturbation is a semi-analytical method that provides approximate solutions. As a result, coupling these two methods would definitely yield a result extremely similar and highly convergent to the exact solution of the problem.

The advantages of the FKHHPM using the Khalouta transform lies in the capability of combining two powerful methods with the need of only initial conditions for obtaining an exact solution to the fractional reaction-diffusion-convection equations, and also the method is characterized by its rapid convergence and easy finding of the unknown coefficients of the series solution without any perturbation, discretization, and linearization, and this method is fast, simple, and adaptable to solving fractional differential equations and others.

The present paper is organized as follows. In section 2, we start by introducing the basic definitions and properties of fractional calculus and Khalouta transform. In section 3, we present the methodology of the FKHHPM for solving the fractional reaction-diffusion-convection equations (1) and (3). In addition, we establish its convergence analysis, and then provide a numerical example to demonstrate the effectiveness of the proposed method in section 4. A discussion of the results obtained is given in section 5. Finally, Section 6 contains the conclusions of our work.

2 Overview of fractional calculus theory

In this section, we present the necessary definitions of fractional calculus and new results concerning the theory of the Khalouta transform of fractional derivatives.

Definition 1. [20] The fractional integral of order $\kappa > 0$ in the Riemann–Liouville sense is expressed as

$$I_{\upsilon}^{\kappa}\chi(\upsilon) = \frac{1}{\Gamma(\kappa)}\int_0^{\upsilon}(\upsilon-\epsilon)^{\kappa-1}\chi(\epsilon)d\epsilon.$$

Definition 2. [20] The fractional derivative of order κ in the Caputo sense is expressed as

$$D_{\upsilon}^{\kappa}\chi(\upsilon) = \begin{cases} \frac{1}{\Gamma(n-\kappa)}\int_0^{\upsilon}(\upsilon-\epsilon)^{n-\kappa-1}\chi^{(n)}(\epsilon)d\epsilon, & n-1<\kappa<n, \\ \chi^{(n)}(\upsilon), & \kappa = n. \end{cases}$$

Definition 3. [20] The Mittag-Leffler function is expressed as

$$E_{\kappa}(\upsilon) = \sum_{n=0}^{\infty}\frac{\upsilon^n}{\Gamma(n\kappa+1)}, \qquad \kappa > 0. \tag{5}$$

Definition 4. [6] The fractional derivative of order $0 < \kappa < 1$ in the Caputo–Fabrizio sense is expressed as

$$\mathcal{D}_{\upsilon}^{(\kappa)}\chi(\upsilon) = \frac{\mathfrak{N}(\kappa)}{1-\kappa}\int_0^{\upsilon}\chi'(\epsilon)\exp\left(-\frac{\kappa(\upsilon-\epsilon)}{1-\kappa}\right)d\epsilon, \tag{6}$$

where $\mathfrak{N}(\kappa)$ is a normalization function that satisfies $\mathfrak{N}(0) = \mathfrak{N}(1) = 1$ and $\chi(\upsilon) \in H^1(\mathbb{R}^+)$.

Definition 5. [6] The fractional derivative of order $\kappa + n$ with $0 < \kappa \leq 1$ and $n \geq 1$ in the Caputo–Fabrizio sense is expressed as

$$\mathcal{D}_{\upsilon}^{(\kappa+n)}\chi(\upsilon) = \mathcal{D}_{\upsilon}^{(\kappa)}(\mathcal{D}_{\upsilon}^{(n)}\chi(\upsilon)).$$

With the hypothesis that $\mathfrak{N}(\kappa) = 1$ in (6), the second definition of Losada and Nieto reads as follows.

Definition 6. [22] The fractional derivative of order $0 < \kappa \leq 1$ in the Caputo-Fabrizio sense is expressed as

$$\mathcal{D}_{\upsilon}^{(\kappa)}\chi(\upsilon)=\frac{1}{1-\kappa}\int_0^{\upsilon}\chi'(\epsilon)\exp\left(-\frac{\kappa\left(\upsilon-\epsilon\right)}{1-\kappa}\right)d\epsilon.$$

Definition 7. [16] The Khalouta transform is defined over the set of functions

$$\mathcal{S}=\left\{\chi(\upsilon)\ :\text{there exist } K,\vartheta_1,\vartheta_2>0,|\chi(\upsilon)|<K\exp\left(\vartheta_j\left|\upsilon\right|\right),\right.$$
$$\left.\text{if } \upsilon\in(-1)^j\times\mathbb{R}^+\right\},$$

by the following integral

$$\mathbb{KH}\left[\chi(\upsilon)\right]=\mathcal{K}(s,\gamma,\eta)=\frac{s}{\gamma\eta}\int_0^{\infty}\exp\left(-\frac{s\upsilon}{\gamma\eta}\right)\chi(\upsilon)d\upsilon,$$

where $s,\gamma,\eta>0$ are the Khalouta transform variables.

Theorem 1. [16] The basic properties of the Khalouta transform are as follows.

1) If $\chi(\upsilon)$ and $\varpi(\upsilon)$ are defined on the set $\mathcal{S}$, then for all constants a and b, we have

$$\mathbb{KH}\left[a\chi(\upsilon)+b\varpi(\upsilon)\right]=a\mathbb{KH}\left[\chi(\upsilon)\right]+b\mathbb{KH}\left[\varpi(\upsilon)\right].$$

2) If the nth derivative of $\chi(\upsilon)$ is $\chi^{(n)}(\upsilon)$, then its Khalouta transform is given as

$$\mathbb{KH}\left[\chi^{(n)}(\upsilon)\right]=\left(\frac{s}{\gamma\eta}\right)^n\mathcal{K}(s,\gamma,\eta)-\sum_{r=0}^{n-1}\left(\frac{s}{\gamma\eta}\right)^{n-r}\chi^{(r)}(0),\qquad n\geq 1.$$

3) If the Khalouta transform of $\chi(\upsilon)$ and $\varpi(\upsilon)$ are $\mathcal{K}(s,\gamma,\eta)$ and $\mathcal{H}(s,\gamma,\eta)$ respectively, defined on the set $\mathcal{S}$, then

$$\mathbb{KH}\left[\left(\chi*\varpi\right)(\upsilon)\right]=\int_0^{\infty}\chi(\upsilon)\varpi(\upsilon-\epsilon)d\epsilon=\frac{\gamma\eta}{s}\mathcal{K}(s,\gamma,\eta)\mathcal{H}(s,\gamma,\eta),$$

where $\mathbb{KH}\left[\left(\chi*\varpi\right)(\upsilon)\right]$ is the Khalouta convolution of the functions $\chi(\upsilon)$ and $\varpi(\upsilon)$.

4) The Khalouta transforms of some special functions are as follows.

$$\mathbb{KH}\left[1\right]=1,$$
$$\mathbb{KH}\left[\upsilon\right]=\frac{\gamma\eta}{s},$$

$$\mathbb{KH}\left[\frac{\upsilon^n}{n!}\right] = \left(\frac{\gamma\eta}{s}\right)^n, \quad n = 0, 1, 2, \ldots,$$
$$\mathbb{KH}\left[\frac{\upsilon^\kappa}{\Gamma(\kappa+1)}\right] = \left(\frac{\gamma\eta}{s}\right)^\kappa, \quad \kappa > -1,$$
$$\mathbb{KH}\left[\exp(a\upsilon)\right] = \frac{s}{s - a\gamma\eta}, \quad a \in \mathbb{R}^+.$$

Theorem 2. [19] The Khalouta transform of fractional derivative in the Caputo sense is expressed as

$$\mathbb{KH}\left[D_\upsilon^\kappa \chi(\upsilon)\right] = \left(\frac{s}{\gamma\eta}\right)^\kappa \mathbb{KH}\left[\chi(\upsilon)\right] - \sum_{r=0}^{n-1} \left(\frac{s}{\gamma\eta}\right)^{\kappa-r} \chi^{(r)}(0),$$

where $n-1 < \kappa \leq n$ and $n \in \mathbb{N}^*$.

Theorem 3. [19] The Khalouta transform of fractional derivative in the Caputo–Fabrizio sense is expressed as

$$\mathbb{KH}\left[\mathcal{D}_\upsilon^{(\kappa+n)} \chi(\upsilon)\right] = \frac{s}{s - \kappa(s - \gamma\eta)} \left(\left(\frac{s}{\gamma\eta}\right)^n \mathbb{KH}\left[\chi(\upsilon)\right] - \sum_{r=0}^{n} \left(\frac{s}{\gamma\eta}\right)^{n-r} \chi^{(r)}(0) \right),$$

where $0 < \kappa \leq 1$ and $n \geq 1$.

3 FKHHPM and its convergence

In this section, we explain the FKHHPM methodology for the fractional reaction-diffusion-convection equations (1) and (3) and study its convergence.

Theorem 4. Let the fractional reaction-diffusion-convection equations be of the form (1) and (3), respectively. According to FKHHPM, the solution to (1)) and (3) can be expressed as an infinite series as follows:

$$\chi(\varrho, \upsilon) = \lim_{p \to 1} \sum_{n=0}^{\infty} \mathfrak{p}^n \chi_n(\varrho, \upsilon) = \sum_{n=0}^{\infty} \chi_n(\varrho, \upsilon), \tag{7}$$

where $\mathfrak{p} \in [0, 1]$ is the homotopy parameter.

Proof. To prove this result, we define

$$\mathfrak{M}_1 \chi = \left(\Psi(\chi)\chi_\varrho\right)_\varrho,$$
$$\mathfrak{M}_2 \chi = \Phi(\chi)\chi_\varrho,$$

$$\mathfrak{M}_3\chi = \Pi(\chi).$$

Case 1: Equation (1) becomes

$$D_{\upsilon}^{\kappa}\chi = \mathfrak{M}_1\chi + \mathfrak{M}_2\chi + \mathfrak{M}_3\chi. \tag{8}$$

Applying the Khalouta transform on (8) and using Theorem 2, we get

$$\mathbb{KH}\left[\chi\right] = \chi_0(\varrho) + \left(\frac{\gamma\eta}{s}\right)^{\kappa} \mathbb{KH}\left[\mathfrak{M}_1\chi + \mathfrak{M}_2\chi + \mathfrak{M}_3\chi\right]. \tag{9}$$

Operating with the inverse Khalouta transform on (9), we have

$$\chi = \chi_0(\varrho) + \mathbb{KH}^{-1}\left[\left(\frac{\gamma\eta}{s}\right)^{\kappa} \mathbb{KH}\left[\mathfrak{M}_1\chi + \mathfrak{M}_2\chi + \mathfrak{M}_3\chi\right]\right]. \tag{10}$$

Using the homotopy perturbation method (HPM) [11], we assume that the solution can be written as a power series of $\mathfrak{p}$ as follows:

$$\chi(\varrho,\upsilon) = \sum_{n=0}^{\infty} \mathfrak{p}^n \chi_n(\varrho,\upsilon), \tag{11}$$

where $\mathfrak{p} \in [0,1]$.

The decomposition of the nonlinear terms is as follows:

$$\begin{aligned}
\mathfrak{M}_1\chi &= \sum_{n=0}^{\infty} \mathfrak{p}^n \mathfrak{H}_n(\chi),\\
\mathfrak{M}_2\chi &= \sum_{n=0}^{\infty} \mathfrak{p}^n \mathfrak{K}_n(\chi),\\
\mathfrak{M}_3\chi &= \sum_{n=0}^{\infty} \mathfrak{p}^n \mathfrak{J}_n(\chi),
\end{aligned} \tag{12}$$

where $\mathfrak{H}_n(\chi), \mathfrak{K}_n(\chi)$ and $\mathfrak{J}_n(\chi)$ are He's polynomials [25], of $\chi_0, \chi_1, \ldots, \chi_n$ and by using the following formulas they are computed:

$$\begin{aligned}
\mathfrak{H}_n(\chi_0,\chi_1,\ldots,\chi_n) &= \frac{1}{n!}\frac{\partial^n}{\partial\mathfrak{p}^n}\left[\mathfrak{M}_1\left(\sum_{i=0}^{\infty}\mathfrak{p}^i\chi_i\right)\right]_{\mathfrak{p}=0},\\
\mathfrak{K}_n(\chi_0,\chi_1,\ldots,\chi_n) &= \frac{1}{n!}\frac{\partial^n}{\partial\mathfrak{p}^n}\left[\mathfrak{M}_2\left(\sum_{i=0}^{\infty}\mathfrak{p}^i\chi_i\right)\right]_{\mathfrak{p}=0},\\
\mathfrak{J}_n(\chi_0,\chi_1,\ldots,\chi_n) &= \frac{1}{n!}\frac{\partial^n}{\partial\mathfrak{p}^n}\left[\mathfrak{M}_3\left(\sum_{i=0}^{\infty}\mathfrak{p}^i\chi_i\right)\right]_{\mathfrak{p}=0}.
\end{aligned} \tag{13}$$

Perform replacements of (11) and (12) in (10), we have

$$\sum_{n=0}^{\infty}\mathfrak{p}^n\chi_n(\varrho,\upsilon)=\chi_0(\varrho)+\mathbb{KH}^{-1}\left[\left(\frac{\gamma\eta}{s}\right)^{\kappa}\mathbb{KH}\left[\begin{array}{l}\sum\limits_{n=0}^{\infty}\mathfrak{p}^n\mathfrak{H}_n(\chi)\\+\sum\limits_{n=0}^{\infty}\mathfrak{p}^n\mathfrak{K}_n(\chi)\\+\sum\limits_{n=0}^{\infty}\mathfrak{p}^n\mathfrak{J}_n(\chi)\end{array}\right]\right]. \tag{14}$$

Using the coefficient of similar powers of $\mathfrak{p}$ in (14), we have

$$\begin{aligned}
&\mathfrak{p}^0:\chi_0(\varrho,\upsilon)=\chi_0(\varrho),\\
&\mathfrak{p}^1:\chi_1(\varrho,\upsilon)=\mathbb{KH}^{-1}\left[\left(\frac{\gamma\eta}{s}\right)^{\kappa}\mathbb{KH}\left[\mathfrak{H}_0(\chi)+\mathfrak{K}_0(\chi)+\mathfrak{J}_0(\chi)\right]\right],\\
&\mathfrak{p}^2:\chi_2(\varrho,\upsilon)=\mathbb{KH}^{-1}\left[\left(\frac{\gamma\eta}{s}\right)^{\kappa}\mathbb{KH}\left[\mathfrak{H}_1(\chi)+\mathfrak{K}_1(\chi)+\mathfrak{J}_1(\chi)\right]\right],\\
&\mathfrak{p}^3:\chi_3(\varrho,\upsilon)=\mathbb{KH}^{-1}\left[\left(\frac{\gamma\eta}{s}\right)^{\kappa}\mathbb{KH}\left[\mathfrak{H}_2(\chi)+\mathfrak{K}_2(\chi)+\mathfrak{J}_2(\chi)\right]\right],\\
&\quad\vdots
\end{aligned}$$

Therefore, the series solution to (8) is found as $\mathfrak{p}\to 1$

$$\chi(\varrho,\upsilon)=\sum_{n=0}^{\infty}\chi_n(\varrho,\upsilon).$$

Case 2: Equation (3) becomes

$$\mathcal{D}_{\upsilon}^{(\kappa)}\chi=\mathfrak{M}_1\chi+\mathfrak{M}_2\chi+\mathfrak{M}_3\chi. \tag{15}$$

Applying the Khalouta transform on (15) and using Theorem 3, we get

$$\mathbb{KH}\left[\chi\right]=\chi_0(\varrho)+\frac{s-\alpha\left(s-\gamma\eta\right)}{s}\mathbb{KH}\left[\mathfrak{M}_1\chi+\mathfrak{M}_2\chi+\mathfrak{M}_3\chi\right]. \tag{16}$$

Operating with the inverse Khalouta transform on (16), we have

$$\chi=\chi_0(\varrho)+\mathbb{KH}^{-1}\left[\frac{s-\alpha\left(s-\gamma\eta\right)}{s}\mathbb{KH}\left[\mathfrak{M}_1\chi+\mathfrak{M}_2\chi+\mathfrak{M}_3\chi\right]\right]. \tag{17}$$

Using the HPM [11], we assume that the solution can be written as a power series of $\mathfrak{p}$ as follows:

$$\chi(\varrho,\upsilon)=\sum_{n=0}^{\infty}p^n\chi_n(\varrho,\upsilon),$$

and by using the formulas (13), the decomposition of the nonlinear terms are computed.

Perform replacements of equations (11) and (12) in (17), we have

$$\sum_{n=0}^{\infty}\mathfrak{p}^n\chi_n(\varrho,\upsilon)=\chi_0(\varrho)+\mathbb{KH}^{-1}\left[\frac{s-\alpha\,(s-\gamma\eta)}{s}\mathbb{KH}\left[\begin{array}{l}\sum\limits_{n=0}^{\infty}\mathfrak{p}^n\mathfrak{H}_n(\chi)\\+\sum\limits_{n=0}^{\infty}\mathfrak{p}^n\mathfrak{K}_n(\chi)\\+\sum\limits_{n=0}^{\infty}\mathfrak{p}^n\mathfrak{J}_n(\chi)\end{array}\right]\right]. \tag{18}$$

Using the coefficient of similar powers of $\mathfrak{p}$ in (18), we have

$$\begin{aligned}
&\mathfrak{p}^0:\chi_0(\varrho,\upsilon)=\chi_0(\varrho),\\
&\mathfrak{p}^1:\chi_1(\varrho,\upsilon)=\mathbb{KH}^{-1}\left[\frac{s-\alpha\,(s-\gamma\eta)}{s}\mathbb{KH}\left[\mathfrak{H}_0(\chi)+\mathfrak{K}_0(\chi)+\mathfrak{J}_0(\chi)\right]\right],\\
&\mathfrak{p}^2:\chi_2(\varrho,\upsilon)=\mathbb{KH}^{-1}\left[\frac{s-\alpha\,(s-\gamma\eta)}{s}\mathbb{KH}\left[\mathfrak{H}_1(\chi)+\mathfrak{K}_1(\chi)+\mathfrak{J}_1(\chi)\right]\right],\\
&\mathfrak{p}^3:\chi_3(\varrho,\upsilon)=\mathbb{KH}^{-1}\left[\frac{s-\alpha\,(s-\gamma\eta)}{s}\mathbb{KH}\left[\mathfrak{H}_2(\chi)+\mathfrak{K}_2(\chi)+\mathfrak{J}_2(\chi)\right]\right],\\
&\quad\vdots
\end{aligned}$$

Therefore, the series solution of (15) is found as $\mathfrak{p}\to 1$

$$\chi(\varrho,\upsilon)=\sum_{n=0}^{\infty}\chi_n(\varrho,\upsilon).$$

□

Theorem 5. Let $\mathfrak{B}$ be a Banach space. Then the FKHHPM solution is convergent if $\chi_0\in\mathfrak{B}$ is bounded and

$$\|\chi_n\|\le\varepsilon\,\|\chi_{n-1}\|,\qquad\text{for all } n\in\mathbb{N}^*,$$

where $0<\varepsilon<1$.

Proof. Let $\{\mathfrak{S}_n\}_{n\ge 0}$ be the sequence as partial sum of (7) as

$$\mathfrak{S}_n(\varrho,\upsilon)=\chi_0(\varrho,\upsilon)+\chi_1(\varrho,\upsilon)+\chi_2(\varrho,\upsilon)+\cdots+\chi_n(\varrho,\upsilon).$$

To obtain the desired result, we must establish that $\{\mathfrak{S}_n\}_{n\geq 0}$ forms a Cauchy sequence in the Banach space $\mathfrak{B}$. Furthermore, let

$$\|\mathfrak{S}_{n+1}-\mathfrak{S}_n\| \leq \|\chi_{n+1}\| \leq \varepsilon \|\chi_n\| \leq \varepsilon^2 \|\chi_{n-1}\| \leq \cdots \leq \varepsilon^{n+1} \|\chi_0\|.$$

For all $n, m \in \mathbb{N}$, with $n \geq m$, we have

$$\begin{aligned}
\|\mathfrak{S}_n-\mathfrak{S}_m\| &= \|\mathfrak{S}_{m+1}-\mathfrak{S}_m+\mathfrak{S}_{m+2}-\mathfrak{S}_{m+1}+\cdots+\mathfrak{S}_n-\mathfrak{S}_{n-1}\| \\
&\leq \|\mathfrak{S}_{m+1}-\mathfrak{S}_m\|+\|\mathfrak{S}_{m+2}-\mathfrak{S}_{m+1}\|+\cdots+\|\mathfrak{S}_n-\mathfrak{S}_{n-1}\| \\
&\leq \varepsilon^{m+1}\|\chi_0\|+\varepsilon^{m+2}\|\chi_0\|+\cdots+\varepsilon^{n}\|\chi_0\| \\
&= \gamma^{m+1}\left(1+\varepsilon+\cdots+\varepsilon^{n-m-1}\right)\|\chi_0\| \\
&\leq \varepsilon^{m+1}\left(\frac{1-\varepsilon^{n-m}}{1-\varepsilon}\right)\|\chi_0\|.
\end{aligned}$$

Since $0<\varepsilon<1$, we have $1-\varepsilon^{n-m}<1$; then

$$\|\mathfrak{S}_n-\mathfrak{S}_m\| \leq \frac{\varepsilon^{m+1}}{1-\varepsilon}\|\chi_0\|.$$

Since χ_0 is bounded, then $\|\chi_0\|<\infty$, which means

$$\lim_{n\longrightarrow\infty}\|\mathfrak{S}_n-\mathfrak{S}_m\|=0,$$

for any finite value of m.

Consequently, $\{\mathfrak{S}_n\}_{n\geq 0}$ is a Cauchy sequence in $\mathfrak{B}$. It follows that the series solution of (1) and (3) as (7) is convergent. □

Theorem 6. Suppose that $\sum_{k=0}^{m}\chi_k(\varrho,\upsilon)<\infty$ and that $\chi(\varrho,\upsilon)$ is its truncated solution. Then the maximum absolute error is presented as

$$\left\|\chi(\varrho,\upsilon)-\sum_{k=0}^{m}\chi_k(\varrho,\upsilon)\right\| \leq \frac{\varepsilon^{m+1}}{1-\varepsilon}\|\chi_0(\varrho,\upsilon)\|.$$

Proof. From Theorem 5, we have

$$\|\mathfrak{S}_n-\mathfrak{S}_m\| \leq \frac{\varepsilon^{m+1}}{1-\varepsilon}\|\chi_0(\varrho,\upsilon)\|. \tag{19}$$

Indeed $\mathfrak{S}_n=\sum_{k=0}^{n}\chi_k(\varrho,\upsilon)$ and considering $n\to\infty$, then $\mathfrak{S}_n\to\chi(\varrho,\upsilon)$; therefore (19) becomes

$$\|\chi(\varrho,\upsilon)-\mathfrak{S}_m\|=\left\|\chi(\varrho,\upsilon)-\sum_{k=0}^{m}\chi_k(\varrho,\upsilon)\right\|\leq\frac{\varepsilon^{m+1}}{1-\varepsilon}\left\|\chi_0(\varrho,\upsilon)\right\|.$$

□

4 Numerical application

Consider the fractional reaction-diffusion-convection equation under Caputo fractional derivative

$$D_{\upsilon}^{\kappa}\chi=\left(\chi\chi_{\varrho}\right)_{\varrho}+3\chi\chi_{\varrho}+2\left(\chi-\chi^{2}\right),0<\kappa\leq1, \tag{20}$$

subject to the condition

$$\chi(\varrho,0)=2\sqrt{e^{\varrho}-e^{-4\varrho}}. \tag{21}$$

Here $\chi=\left\{\chi(\varrho,\upsilon):(\varrho,\upsilon)\in\mathbb{R}^{+}\times\mathbb{R}^{+}\right\}$.

Using the methodology of the FKHHPM presented in Section 3, we get the following result:

$$\begin{aligned}
&\mathfrak{p}^{0}:\chi_0(\varrho,\upsilon)=2\sqrt{e^{\varrho}-e^{-4\varrho}},\\
&\mathfrak{p}^{1}:\chi_1(\varrho,\upsilon)=\frac{4\upsilon^{\kappa}}{\Gamma(\kappa+1)}\sqrt{e^{\varrho}-e^{-4\varrho}},\\
&\mathfrak{p}^{2}:\chi_2(\varrho,\upsilon)=\frac{8\upsilon^{2\kappa}}{\Gamma(2\kappa+1)}\sqrt{e^{\varrho}-e^{-4\varrho}},\\
&\mathfrak{p}^{3}:\chi_3(\varrho,\upsilon)=\frac{16\upsilon^{3\kappa}}{\Gamma(3\kappa+1)}\sqrt{e^{\varrho}-e^{-4\varrho}},\\
&\quad\vdots
\end{aligned}$$

Therefore, the series solution to (20) and (21) is

$$\begin{aligned}
\chi(\varrho,\upsilon)&=\lim_{p\to1}\sum_{n=0}^{\infty}\mathfrak{p}^{n}\chi_n(\varrho,\upsilon)=\sum_{n=0}^{\infty}\chi_n(\varrho,\upsilon)\\
&=\chi_0(\varrho,\upsilon)+\chi_1(\varrho,\upsilon)+\chi_2(\varrho,\upsilon)+\chi_3(\varrho,\upsilon)+\cdots\\
&=2\sqrt{e^{\varrho}-e^{-4\varrho}}\left(1+\frac{2\upsilon^{\kappa}}{\Gamma(\kappa+1)}+\frac{2^{2}\upsilon^{2\kappa}}{\Gamma(2\kappa+1)}+\frac{2^{3}\upsilon^{3\kappa}}{\Gamma(3\kappa+1)}+\cdots\right)\\
&=2\sqrt{e^{\varrho}-e^{-4\varrho}}\sum_{n=0}^{\infty}\frac{\left(2\upsilon^{\kappa}\right)^{n}}{\Gamma(n\kappa+1)}
\end{aligned}$$

$$= 2\sqrt{e^{x} - e^{-4x}} E_{\alpha}\left(2\upsilon^{\kappa}\right),$$

where $E_{\alpha}\left(2\upsilon^{\kappa}\right)$ is the Mittag-Leffler function defined by (5).

At $\kappa \longrightarrow 1$, the exact solution obtained is

$$\chi(\varrho, \upsilon) = 2e^{2\upsilon}\sqrt{e^{\varrho} - e^{-4\varrho}}.$$

which is the same result, when we use the homotopy analysis method (HAM) [28].

Now, we consider the fractional reaction-diffusion-convection equation under Caputo–Fabrizio fractional derivative

$$\mathcal{D}_{\upsilon}^{(\kappa)}\chi = \left(\chi\chi_{\varrho}\right)_{\varrho} + 3\chi\chi_{\varrho} + 2\left(\chi - \chi^{2}\right), 0 < \kappa \leq 1, \tag{22}$$

subject to the condition

$$\chi(\varrho, 0) = 2\sqrt{e^{\varrho} - e^{-4\varrho}}. \tag{23}$$

Here $\chi = \left\{\chi(\varrho, \upsilon) : (\varrho, \upsilon) \in \mathbb{R}^{+} \times \mathbb{R}^{+}\right\}$.

Using the methodology of the FKHHPM presented in Section 3, we get the following result:

$$\begin{aligned}
&\mathfrak{p}^{0} : \chi_{0}(\varrho, \upsilon) = 2\sqrt{e^{\varrho} - e^{-4\varrho}},\\
&\mathfrak{p}^{1} : \chi_{1}(\varrho, \upsilon) = 4\left(1 - \kappa + \kappa\upsilon\right)\sqrt{e^{\varrho} - e^{-4\varrho}},\\
&\mathfrak{p}^{2} : \chi_{2}(\varrho, \upsilon) = 8\left(\left(1-\kappa\right)^{2} + 2\kappa(1-\kappa)\upsilon + \kappa^{2}\frac{\upsilon^{2}}{2!}\right)\sqrt{e^{\varrho} - e^{-4\varrho}},\\
&\mathfrak{p}^{3} : \chi_{3}(\varrho, \upsilon) = 16\left(\left(1-\kappa\right)^{3} + 3\kappa(1-\kappa)^{2}\upsilon + 3\kappa^{2}(1-\kappa)\frac{\upsilon^{2}}{2!} + \kappa^{3}\frac{\upsilon^{3}}{3!}\right)\sqrt{e^{\varrho} - e^{-4\varrho}},\\
&\quad\vdots
\end{aligned}$$

Therefore, the approximate series solution to equations (22) and (23) is

$$\begin{aligned}
\chi(\varrho, \upsilon) &= \lim_{p \to 1}\sum_{n=0}^{\infty}\mathfrak{p}^{n}\chi_{n}(\varrho, \upsilon) = \sum_{n=0}^{\infty}\chi_{n}(\varrho, \upsilon)\\
&= \chi_{0}(\varrho, \upsilon) + \chi_{1}(\varrho, \upsilon) + \chi_{2}(\varrho, \upsilon) + \chi_{3}(\varrho, \upsilon) + \cdots.
\end{aligned}$$

At $\kappa \longrightarrow 1$, the exact solution obtained is

$$\chi(\varrho,\upsilon) = 2\sqrt{e^{\varrho}-e^{-4\varrho}}\left(1+2\upsilon+\frac{(2\upsilon)^2}{2!}+\frac{(3\upsilon)^3}{3!}+\cdots\right)$$
$$= 2e^{2\upsilon}\sqrt{e^{\varrho}-e^{-4\varrho}}.$$

which is the same result when we use the HAM [28].

5 Results and discussion

In this section, we examine the figures and tables. The results are calculated graphically and numerically using MATLAB R2016a to demonstrate the approximate analytical solution compared to the exact solution. The simulation results reveal that the results obtained by the proposed method closely approximate the exact solution to (20) and (22). Figure 1 represents the surface graph of the 4-term approximate solutions by Caputo-FKHHPM at $\kappa = 0.8, 0.9, 1$ and the exact solution for (20). Figure 2 represents the surface graph of the 4-term approximate solutions by Caputo–Fabrizio-FKHHPM at $\kappa = 0.8, 0.9, 1$ and the exact solution for (22). Figure 3 represents the behavior of the 4-approximate solutions at $\varrho = 1$ for different values of κ and the exact solution for equations (20) and (22), respectively. According to these figures, we can say that when the order of the fractional derivative tends to 1, the approximate solutions obtained by FKHHPM tend continuously to the exact solutions. Tables 1 and 2 show the numerical values of the 4-term approximate solutions and the exact solution for different values of κ and υ at $\varrho = 0.1$ for equations (20) and (22), respectively. From the above, numerical results confirm that the approximate solutions are in good agreement with exact solutions at $\kappa = 1$ for (20) and (22), respectively, and the high precision of the proposed method. Also, it is to be noted that the accuracy can be improved by computing more terms of approximated solutions. In addition, it can be observed that an increase in υ does not significantly affect the accuracy of FKHHPM and that our method still approximates the exact solution sufficiently due to the fractional parameter.

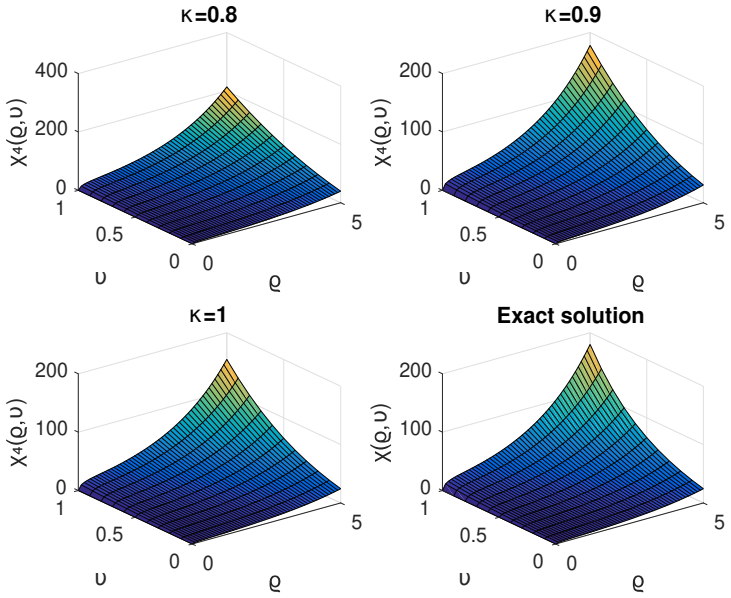


Figure 1: Surface graphs of the Caputo-FKHHPM solution and exact solution for (20).

Table 1: The values of the exact and numerical solutions of (20) for different values of κ and υ at $\varrho = 0.1$

υ	$\kappa = 0.7$	$\kappa = 0.8$	$\kappa = 0.9$	$\kappa = 1$	χ_{exact}	$\lvert\chi_{exact} - \chi_{FKHHPM-C}\rvert$
0.01	1.4415	1.3924	1.3631	1.3455	1.3455	8.8277×10^{-9}
0.03	1.6026	1.5045	1.4416	1.4004	1.4004	7.2082×10^{-7}
0.05	1.7484	1.6099	1.5190	1.4576	1.4576	5.6070×10^{-6}
0.07	1.8908	1.7147	1.5975	1.5170	1.5171	2.1716×10^{-5}
0.09	2.0334	1.8207	1.6780	1.5789	1.5790	5.9828×10^{-5}

6 Conclusion

In this study, we proposed a novel combination process known as the FKHHPM to solve the fractional reaction-diffusion-convection equations with Caputo and Caputo–Fabrizio fractional derivatives. A numerical application is solved to illustrate that the current method is efficient, accurate, and converges rapidly to the exact solution. Therefore, we can say that the proposed

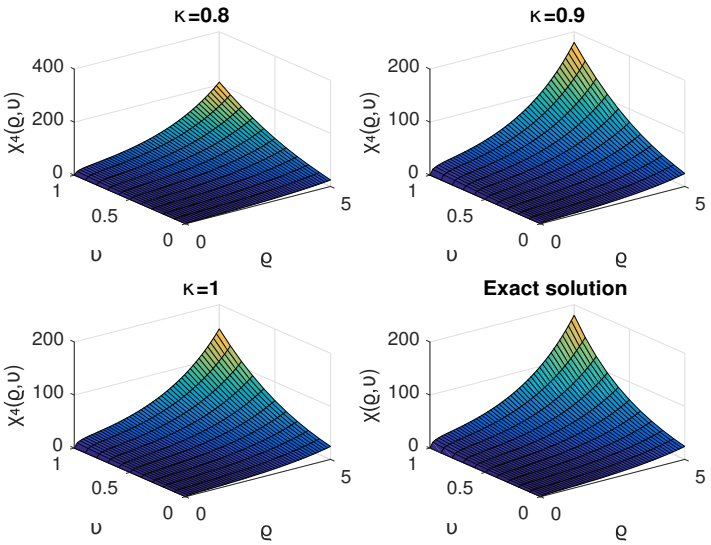


Figure 2: Surface graphs of the Caputo–Fabrizio-FKHHPM solution and exact solution for (22).

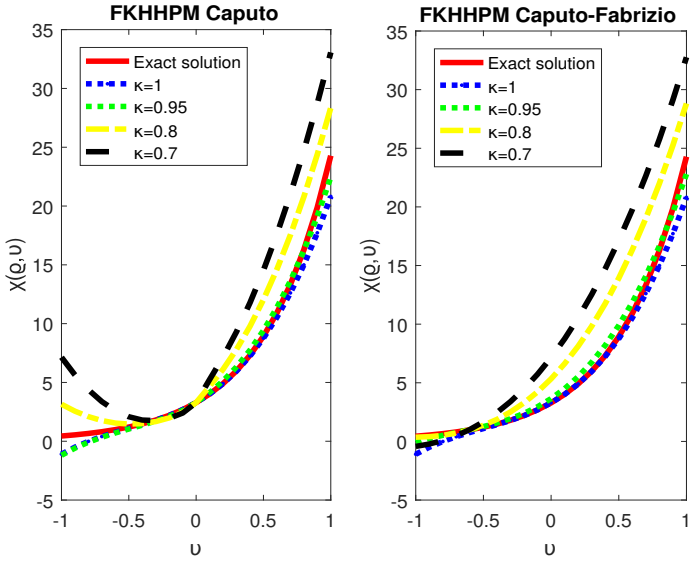


Figure 3: 2D plots of the numerical and exact solutions for (20) and (22) for different values of κ at $\varrho = 1$.

Table 2: The values of the exact and numerical solutions of (22) for different values of κ and υ at $\varrho = 0.1$

υ	$\kappa = 0.7$	$\kappa = 0.8$	$\kappa = 0.9$	$\kappa = 1$	χ_{exact}	$\lvert\chi_{exact} - \chi_{FKHHPM-CF}\rvert$
0.01	2.9308	2.1903	1.6824	1.3455	1.3455	8.8277×10^{-9}
0.03	3.0548	2.2895	1.7573	1.4004	1.4004	7.2082×10^{-7}
0.05	3.1818	2.4971	1.8351	1.4576	1.4576	5.6070×10^{-6}
0.07	3.3117	2.5510	1.9157	1.5170	1.5171	2.1716×10^{-5}
0.09	3.4447	2.6056	1.9993	1.5789	1.5790	5.9828×10^{-5}

method is a reliable and robust method for solving fractional differential equations, which has wide applications in natural sciences and engineering.

Acknowledgements

The author thanks the anonymous reviewers and the editor for their constructive comments.

References

[1] Abdeljawad, T. *On conformable fractional calculus*, Int. J. Comput. Appl. Math. 279 (2015) 57–66.

[2] Alazman, I., Mishra, M.N., Alkahtani, B.S. and Dubey, R.S *Analysis of infection and diffusion coefficient in an SIR model by using generalized fractional derivative*, Fractal Fract. 2024(8) (2024) 1–21.

[3] Allaire, G. and Raphael, A.L. *Homogenization of a convection-diffusion model with reaction in a porous medium*, C. R. Acad. Sci. Ser. I. 344 (2007) 523–528.

[4] Atangana, A. and Baleanu, D. *New fractional derivatives with non-local and non-singular kernel theory and application to heat transfer model*, Therm. Sci. 20(2) (2016), 763–769.

[5] Baskonus, H.M., Mekkaoui, T., Hammouch, Z. and Bulut, H. *Active control of a chaotic fractional order economic system*, Entropy. 17(8) (2015) 5771–5783.

[6] Caputo, M. and Fabrizio, M. *A new definition of fractional derivative without singular kernel*, Prog. Fract. Differ. Appl. 1(2) (2015) 73–85.

[7] Crauste, F., Lhassan, M. and Kacha, A. *A delay reaction–diffusion model of the dynamics of botulinum in fish*, Math. Biosci. 216 (2008) 17–29.

[8] Dastjerdi, R.H. and Ahmadi, G. *Designing the sinc neural networks to solve the fractional optimal control problem*, Iran. J. Numer. Anal. Optim. 14(4) (2024) 1016–1036.

[9] Dubeya, R.S., Mishraa, M.N. and Goswami, P. *Systematic analysis of oxygen diffusion problem having local fractional derivative*, Applied Innovative Research, 3(1-4) (2022) 113–120.

[10] Ferreira, S., Martins, M. and Vilela, M. *Reaction-diffusion model for the growth of avascular tumor*, Phys. Rev. 65(2) (2002) 1–12.

[11] He, J.H. *Application of homotopy perturbation method to nonlinear wave equations*, Chaos Soliton. Fract. 26 (2005) 695–700.

[12] Hilton, H.H. *Generalized fractional derivative anisotropic viscoelastic characterization*, Mater. 5 (2012) 169–191.

[13] Iaffaldano, G., Caputo, M. and Martino, S. *Experimental and theoretical memory diffusion of water in sand*, ; Hydrol. Earth Syst. Sci. 10(1) (2006) 93–100.

[14] Kamocki, R. *A new representation formula for the Hilfer fractional derivative and its application*, Int. J. Comput. Appl. Math. 308 (2016) 39–45

[15] Katugampola, U.N. *New approach to generalized fractional integral*, Appl. Math. Comput. 218 (2011) 860–865.

[16] Khalouta, A. *A new exponential type kernel integral transform: Khalouta transform and its applications*, Math. Montisnigri. 57 (2023) 5–23.

[17] Khalouta, A. *The study of nonlinear fractional partial differential equations via the Khalouta-Atangana-Baleanu operator*, J. Appl. Anal. Comput. 14(6) (2024) 3175–3196.

[18] Khalouta, A. *New results of the ρ-Jafari transform and their application to linear and nonlinear generalized fractional differential equations*, Rev. Colomb. Mat. 58(1) (2024), 25–46.

[19] Khalouta, A. *Khalouta transform via different fractional derivative operators,* Vestn. Samar. Gos. Tekhn. Univ., Ser. Fiz.-Mat. Nauki. 28(3) (2024), 407–425.

[20] Kilbas, A., Srivastava, H.M. and Trujillo, J.J. *Theory and application of fractional differential equations*, Elsevier, Amsterdam, 2006.

[21] Kumawat, N., Shukla, A., Mishra, M.N., Sharma, R. and Dubey, R.S. *Khalouta transform and applications to Caputo-fractional differential equations*, Front. Appl. Math. Stat. 10 (2024) 1351526.

[22] Losada, J. and Nieto, J.J. *Properties of a new fractional derivative without singular kernel*. Prog. Fract. Differ. Appl. 1(2) (2015) 87–92.

[23] Magin, R.L. *Fractional calculus in bioengineering*, Begell House Inc. Publishers, 2006.

[24] Miller, K.S. and Ross, B. *An introduction to the fractional calculus and fractional differential equations*, Wiley, New York, 1993.

[25] Mohyud-Din, S.T. Noor, M.A. and Noor, K.I. *Traveling wave solutions of seventh-order generalized KdV equation using He's polynomials*, Int. J. Nonlinear Sci. Numer. Simul. 10 (2009) 227–233.

[26] Podlubny, I. *Fractional differential equations*, Academic Press, New York, 1999.

[27] Sandev, T., Metzler, R. and Tomovski, Z. *Fractional diffusion equation with a generalized Riemann–Liouville time fractional derivative*, J. Phys. A: Math. Theor. 44 (2011) 1–21.

[28] Shidfar, A., Babaei, A. Molabahrami, A. and Alinejadmofradi, M. *Approximate analytical solutions of nonlinear reaction-diffusion-convection problem*, Math. Comput. Model. 53(1-2) (2011) 261–268.

[29] Siddheshwar, P.G. and Manjunath, S. *Unsteady convective-diffusion with heterogeneous chemical reaction in a plane-Poseuille flow of a micropolar fluid*, Internat. J. Engrg. Sci. 38 (2000) 765–783.

[30] Yadav, S.K., Purohit, M., Gour, M.M., Yadav, L.K. and Mishra, M.N., *Hybrid technique for multi-dimensional fractional diffusion problems involving Caputo–Fabrizio derivative*, Int. J. Math. Ind. 16(01) (2024) 2450020.

Iranian Journal of Numerical Analysis and Optimization
Vol. 15, No. 2, 2025, pp 645–654
https://doi.org/10.22067/ijnao.2024.89575.1500
https://ijnao.um.ac.ir/

Research Article

Numerical study of the Sturm–Liouville problem

S.D. Algazin, A.A. Sinitsyn*,

Abstract

The article discusses the general Sturm–Liouville problem. To solve it numerically, a new algorithm is proposed, which is based on the variational principle and does not use saturation. The problem of constructing numerical methods for solving eigenvalue problems can be divided into two stages. First, we need to reduce the infinite-dimensional problem into a finite-dimensional one, and then find a method for solving this finite-dimensional algebraic eigenvalue problem. In this paper, we only consider the first stage, and solve the resulting algebraic problem using the QR algorithm. A comparison with the results of other authors is also carried out. Methodical calculations confirm the correctness of the new approach.

*Corresponding author

Received 26 August 2024; revised 22 September 2024; accepted 11 December 2024

Sergey Algazin
Doctor of Physical and Mathematical Sciences, Leading Researcher, Laboratory of Mechanics and Optimization of Structures, Ishlinsky Institute for Problems of Mechanics of the Russian Academy of Sciences, e-mail: algazinsd@mail.ru

Artemiy Sinitsyn
Student of the MSU Faculty of Mechanics and Mathematics, Department of Elasticity Theory, e-mail: art@sinitsyn.info

How to cite this article
Algazin, S.D. and Sinitsyn, A.A., Numerical study of the Sturm–Liouville problem. *Iran. J. Numer. Anal. Optim.*, 2025; 15(2): 645-654. https://doi.org/10.22067/ijnao.2024.89575.1500

AMS subject classifications (2020): Primary 34B24; Secondary 65Y20, 74S99.

Keywords: Sturm–Liouville problem, Numerical algorithm without saturation, Computational experiment.

1 Introduction

There are a number of competing methods for the numerical solution of the eigenvalue problem. These are, first of all, projection methods: the Ritz method, the Bubnov–Galerkin method, and so on. We know quite a lot about the accuracy provided by these methods. For example, the approximations for the eigenvalues of self-adjoint problems given by the Ritz method lie on top of the exact values. A number of convergence results are known, and in some special cases, error estimates of projection methods are obtained [10]. Along with projection methods, difference methods have also become widespread [9]. However, when designing these numerical methods, a number of important circumstances are not taken into account, which significantly reduces their effectiveness. Usually, when solving an eigenvalue problem, we have colossal a priori information. Most often, the solutions sought are infinitely differentiable or even analytical. Therefore, they are elements of functional compacts, quite simply arranged. As a rule, the asymptotics of their diameters are well known for such compacts. On the other hand, any projection method is based on the choice of a certain set of finite-dimensional subspaces and thereby some way of approximating the desired solution (and this method, as a rule, is not consistent with the optimal methods mentioned above). This naturally leads to the fact that the numerical algorithm based on such a projection method is far from optimal in its properties. At the same time, by basing the numerical algorithm on a rational way of approximating the desired element, we obtain an algorithm close to the optimal one. This approach will be developed below, and it is based on the ideas of the work [4]. The different methods have significant disadvantages [4] and, in particular, the fact that they are methods with saturation (quite a lot of works have been devoted to the accuracy of these methods, and of them we will point only to [9, 5]). Therefore, the difference method of solving the eigenvalue problem

again ignores a priori information about the smoothness of the solution, and taking into account the loss of smoothness inherent in difference methods, we obtain algorithms that are far from optimal. The problem of constructing numerical methods for solving the eigenvalue problem is divided into two. First of all, you need to reduce an infinite-dimensional problem to a finite-dimensional problem, and then specify a method for solving the resulting algebraic eigenvalue problem. In this paper, only the first stage is considered, and the resulting algebraic problem is solved by the QR method.

Abstract theorems on error estimation in eigenvalue problems are published in [2, 7]. Note that in [7] only compact operators are considered, and in [2] arbitrary closed operators are considered.

A special case of the Sturm–Liouville problem was considered earlier in [3].

2 Problem statement and variational principle

Consider the eigenvalue problem (Sturm–Liouville problems):

$$-\frac{d}{dx}\left(p(x)\frac{du}{dx}\right)+r(x)u(x)-\lambda\rho(x)u(x)=0, \qquad a\le x\le b, \tag{1}$$

$$\alpha u'(a)-\beta u(a)=0, \quad \gamma u'(b)+\delta u(b)=0, \tag{2}$$

with $\alpha>0$, $\gamma>0$, $\beta\ge 0$, $\delta\ge 0$, and at least one of the coefficients β and δ is different from zero, $p(x)\ge 0, r(x)\ge 0, A=\int_a^b \frac{dx}{p(x)}<\infty$, $\rho_0\le\rho(x)\le\rho_1$, $\rho_0,\rho_1>0$. Then the problem (1)–(2) has a discrete spectrum [8].

Variational principle. Denote $Au=-\frac{d}{dx}\left(p(x)\frac{du}{dx}\right)+r(x)u(x)$.

Differential operator is defined on functions satisfying boundary conditions (2). Then

$$|u|^2=(Au,u)=\int_a^b (pu'^2+ru^2)dx+\frac{\beta}{\alpha}p(a)u^2(a)+\frac{\delta}{\gamma}p(b)u^2(b)\equiv J(u).$$

Let $\bar{u}=u+\varepsilon\eta(x)$. Then

$$J(\bar{u})=\int_a^b (p(u'+\varepsilon\eta')^2+r(u+\varepsilon\eta)^2)dx$$

$$+\frac{\beta}{\alpha}p(a)(u(a)+\varepsilon\eta(a))^2+\frac{\delta}{\gamma}p(b)(u(b)+\varepsilon\eta(b))^2$$
$$\equiv\Phi(\varepsilon).$$

Moreover,

$$\Phi'(0)=\int_a^b(p(x)u'(x)(u'+\eta')+ru(u+\eta))dx+\frac{\beta}{\alpha}p(a)u(a)(u(a)+\eta(a))$$
$$+\frac{\delta}{\gamma}p(b)u(b)(u(b)+\eta(b))$$
$$=\int_a^b p(x)(u'(x))^2dx+\int_a^b pu'\eta'dx+\int_a^b ru^2dx$$
$$+\int_a^b ru\eta dx+\frac{\beta}{\alpha}p(a)u(a)(u(a)+\eta(a))+\frac{\delta}{\gamma}p(b)u(b)(u(b)+\eta(b)).$$

Note that

$$\int_a^b pu'\eta'dx=\int_a^b pu'd\eta=pu'\eta|_a^b-\int_a^b(pu')'\eta dx,$$
$$\int_a^b pu'^2dx=\int_a^b pu'du=pu'u|_a^b-\int_a^b(pu')'udx,$$
$$\int_a^b[-(pu')'+ru]udx=0.$$

By virtue of the equation, it remains without η:

$$pu'u|_a^b+\frac{\beta}{\alpha}p(a)u^2(a)+\frac{\delta}{\gamma}p(b)u^2(b)$$
$$=p(u'(b)+\frac{\delta}{\gamma}u(b))u(b)+p(-u'(a)+\frac{\beta}{\alpha}u(a))u(a)=0.$$

Due to the boundary conditions, it remains with η:

$$\int_a^b\eta[-(pu')'+ru]dx+p(b)[u'(b)+\frac{\delta}{\gamma}u(b)]\eta(b)+p(a)[-u'(a)+\frac{\beta}{\alpha}u(a)]\eta(a)=0.$$

Because the function η is arbitrary, then we get from the condition $J(u)\to\min$, equations and boundary conditions (1)–(2).

3 Interpolation formula

Let $u=u(y), y\in[a,b]$, and replacement $y=\frac{b-2}{2}x+\frac{a+b}{2},\ x\in[-1,+1]$.

The interpolation formula by x reads as follows:

$$\begin{aligned} P_{n+2}(x;u) =& \sum_{j=0}^{n+1} \frac{(x^2-1)T_n(x)}{[\ldots]'_{x=x_j}(x-x_j)} u_j,\ u_j = u(y_j), \\ u_a =& u(a),\ u_b = u(b). \\ y_j =& \frac{b-2}{2} x_j + \frac{a+b}{2},\ x_j = \cos\theta_j,\ \theta_j = \frac{2j-1}{2n}\pi, \\ x_0 =& -1, \quad x_{n+1} = +1, \quad T_n(x) = \cos(n \arccos x). \end{aligned}$$

From here we get that because
$[\ldots]'_{x=x_j} = 2xT_n(x) + (x^2-1)T'_n(x)\big|_{x=x_j} = (x_j^2-1)T'_n(x_j)$, then

$$\frac{T_n(x)}{T'_n(x_j)(x-x_j)} = {\sum_{k=0}^{n-1}}' a_k^{(n)} T_k(x),\ a_k^{(n)} = \frac{2}{n} T_k(x_j) = \frac{2}{n}\cos k\theta_j.$$

where symbol "$'$" of the sign of the sum indicates that the summand for $k=0$ is taken with the coefficient $\frac{1}{2}$.

Quadrature formula. Let us define the coefficients of the quadrature formula of the integral. They are the sum of the weights on the function values in the nodes as follows:

$$\begin{aligned} \int_a^b f(y)dy \qquad & (y = \frac{b-a}{2}x + \frac{a+b}{2}, \quad dy = \frac{b-a}{2}dx) \\ =& \frac{b-a}{3}\int_{-1}^{+1} f(x)dx \\ =& \frac{b-a}{2} - \frac{u_a}{2T_n(-1)} I_n^a + -\frac{u_b}{2T_n(1)} I_n^b + \int_{-1}^{+1}(\ldots)dx, \\ T_n(1) =& 1, \qquad T_n(-1) = (-1)^n. \end{aligned}$$

Here $T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)$ which implies

$$\int_{-1}^{+1} T_k(x)dx = \frac{(-1)+(-1)^{k+1}}{k^2-1} = \begin{cases} 0, & k\text{-odd}, \\ \frac{-2}{k^2-1}, & k\text{ -even}, \end{cases}$$

$$I_n^a = \int_{-1}^{+1}(x-1)T_n(x)dx = \begin{cases} \frac{2}{n^2-1}, & k\text{-even}, \\ \frac{-2}{n^2-4}, & k\text{ -odd}, \end{cases}$$

$$I_n^b = \int_{-1}^{+1} (x+1)T_n(x)dx = \begin{cases} \frac{-2}{n^2-1}, & k\text{-even}, \\ \frac{1}{n^2-4}, & k \text{ -odd}, \end{cases}$$

$$I_k = \int_{-1}^{+1} (x^2-1)T_k(x)dx = \begin{cases} \frac{1}{k^2-1} - \frac{1}{2}\{\frac{1}{(k+1)(k+3)} + \frac{1}{(k-1)(k-3)}\}, & k\text{-even}, \\ 0, & k \text{ -odd}, \end{cases}$$

$$c_a = \frac{b-a}{2}\{\frac{(-1)^{n+1}}{2} I_n^a\},$$

$$c_b = \frac{b-a}{2}\{\frac{1}{2} I_n^b\},$$

$$c_j = \frac{b-a}{2}\{\frac{-2}{n} \sum_{k=0(2)}^{n-1} \frac{\cos k\theta_j}{\sin^2 \theta_j} I_k\},$$

$$I_0 = -\frac{4}{3},$$

$$\int_a^b f(y)dy = \frac{b-a}{2} \int_{-1}^{+1} f\left(\frac{b-a}{2}x + \frac{a+b}{2}\right) dx$$
$$= c_a f_a + c_b f_b + \sum_{j=1}^{n} c_j f_j, \qquad f_j = \cos\theta_j, \ j = 1, 2, \ldots, n.$$

Numerical differentiation. Let us differentiate the interpolation formula by x as follows:

$$P_{n+2}(x,u) = \frac{1}{2}(x+1)T_n(x)u_b - \frac{(-1)^n}{2}(x-1)T_n(x)u_a + \frac{2}{n}\sum_{j=1}^{n} \frac{1-x^2}{\sin^2\theta_j}\{\sum_{k=0}^{n-1} \cos k\theta_j T_k(x)\}u_j,$$

$$f(y) = f\{\frac{b-a}{2}x + \frac{a+b}{2}\} f'_x = f'_y \frac{b-a}{2} \Rightarrow f'_y = \frac{2}{b-a} f'_x,$$

$$P'_{n+2}(x,u) = \frac{u_b}{2}\{(x+1)T_n{}'(x) + T_n(x)\} - \frac{(-1)^n}{2}\{(x-1)T_n(x) + T_n(x)\}u_a + \frac{2}{n}\sum_{j=1}^{n}\{\frac{-1}{\sin^2\theta_j}\{\sum_{k=0}^{n-1}\{2xT_k(x) + (x^2-1)T'_k(x)\}\cos k\theta_j\}u_j,$$

Let $x = \pm 1$, or let $x = \cos\theta_i,\ i = 1, 2, \ldots, n$. Then

$$P'_{n+2}(-1,u) = \frac{u_b}{2}(-1)^n - \frac{1+2n^2}{2}u_a + \frac{2}{n}\sum_{j=1}^{n}\{\frac{1}{\sin^2\theta_j}\{\sum_{k=0}^{n-1}\{2(-1)^k\}\cos k\theta_j\}\}hu_j,$$

$$P'_{n+2}(+1,u) = \frac{1+2n^2}{2}u_b - \frac{(-1)^n}{2}u_a + \frac{2}{n}\sum_{j=1}^{n}\{\frac{-1}{\sin^2\theta_j}\{\sum_{k=0}^{n-1}2\cos k\theta_j\}\}u_j,$$

$$P'_{n+2}(x_i,u) = \frac{u_b}{2}\{n(x_i+1)\frac{\sin n\theta_i}{\sin\theta_i}\} - \frac{(-1)^n}{2}\{n(x_i-1)\frac{\sin n\theta_i}{\sin\theta_i}\}u_a + \frac{2}{n}\sum_{j=1}^{n}\{\frac{-1}{\sin^2\theta_j}\{\sum_{k=0}^{n-1}\{2x_i\cos k\theta_i - k\sin\theta_i\sin k\theta_i\}\cos k\theta_j\})\}u_j.$$

4 Discretization

The problem of discretizing a quadratic functional can be reduced to the problem of finding the minimum of a quadratic function as follows:

$$\int_a^b (pu'^2 + ru^2)dx + \frac{\beta}{\alpha}p(a)u^2(a) + \frac{\delta}{\gamma}p(b)u^2(b) \equiv J(u) \to \min.$$

Calculating the integral by the quadrature formula, we reduce the problem of the minimum of the quadratic functional to the problem of the minimum of the quadratic form:

$$J(u) \to J(u_b, u_1, u_2, \ldots, u_n, u_a),$$
$$\frac{\partial J(u_b, u_1, u_2, \ldots, u_n, u_a)}{\partial u_l} = 0, \quad l = b, 1, 2, \ldots, n, a.$$

Calculation results. The calculations use $Q = 0, N_2 = 1026$. Then the calculations give $\lambda_0 = -0.878$; $k = 1$, $\lambda_1 = 1.000002$; $k = 72$, $\lambda_{72} = 5184.001$.

Table 1: $p = 1,\ q = 20\cos 2y + 100\sin^2 2y\,, u'(0) = 0,\ u'(\pi) = 0$

i	λ_i	λ_i	λ_i	λ_i
N2	34	64	134	[6]
1	$2.0 * 10^-5$	$5.2 * 10^-8$	$5.1 * 10^-8$	0.0000000
2	37.7733909214894	37.8059004212607	37.8059002738378	37.7596285
3	41.3283618161036	45.6727315410478	46.6101985290947	37.8059002
4	69.8880664248824	69.8212185270856	69.8010881122681	37.8525995
5	70.7581077855011	70.6037307664520	70.5599120331492	70.5475097
6	71.5270508988527	71.4361363622447	71.4119790617174	92.6538177
7	96.2500910443853	96.2105986913062	96.2068706709564	96.2058159
8	110.785329891999	110.714479891137	110.695004912174	102.254347

Table 2: $p = 1, q = \exp(y), u'(0) = 0,\ u'(\pi) = 0$

i	λ_i	λ_i	λ_i	λ_i
N2	34	66	1026	[6]
1	4.9013	4.8978	4.8966739510209	4.89571
2	5.1951	5.2204	5.22886153287197	-
3	10.053	10.047	10.0451980276767	9.99955
4	16.033	16.022	16.0192812195823	15.4685
5	23.292	23.292	23.2662960864049	21.0371
6	32.306	32.274	32.2637489027225	28.1893
7	43.284	43.236	43.2200825906882	37.7907
8	56.271	56.204	56.1816819519642	49.6137
9	71.272	71.182	71.1531142297211	63.5205
10	88.285	88.170	88.1322684247181	79.4646
11	107.30	107.16	107.116861697407	97.4279
12	128.33	128.16	128.105246952299	117.402
13	151.37	151.16	151.096313344083	139.384
14	176.41	176.17	176.089314130927	163.370
15	203.46	203.17	203.083739890564	189.359
16	232.51	232.18	232.079236175438	217.351
17	263.65	263.19	263.075551299445	247.344
18	296.87	296.21	296.072503036626	279.338
19	334.82	331.22	331.069957054295	313.334
20	373.85	368.24	368.067812678323	349.330
21	437.56	407.26	407.065993320228	387.326

Table 3: $p = 1, q = \frac{e\cos(2\pi y)}{1+e\cos(2\pi y)}, u'(0) = 0, u' = 0$

N2	34	64	[1]
e = 0.1	9.83337323117731	9.81926933696682	$\lambda_1^s = 9.81382$
e = 0.2	9.76485503817489	9.75081106707683	$\lambda_1^s = 9.74579$

5 Discussion of the results

Methodical calculations at $q = 0$ demonstrate the correctness of the methodology. Comparison with the results of [1] (see Table 3) confirms its correctness. However a comparison with the results of [6] demonstrates a discrepancy. In Table 2, in the cited work, the second eigenvalue is omitted. The remaining eigenvalues match satisfactorily only for small numbers of eigenvalues. In Table 1, only individual eigenvalues match satisfactorily.

Acknowledgements

Authors are grateful to there anonymous referees and editor for their constructive comments. The work was carried out on the topic of state assignment no. 124012500437-9.

References

[1] Akulenko, L.D. and Nesterov. S.V. *High-precision methods in eigenvalue problems and their applications*, Chapman and Hall/CRC, 2004.

[2] Algazin, S.D. *Localization of eigenvalues of closed linear operators*, Sib. math. J. 24 (2) (1983) 155–159.

[3] Algazin, S.D., Babenko, K.I. and Kosorukoye, A.L. *The numerical solution of the eigenvalue problem*, IPMatem Akad. Nauk S.S.S.R. Moscow, 1975.

[4] Babenko, K.I. *Fundamentals of numerical analysis*, Moscow, Nauka, 744, 1986 [in Russian].

[5] Hubbard, B.E. *Bounds for eigenvalues of the Sturm-Liouville problem by finite difference methods*. Arch. Ration. Mech. Anal. 10 (2) (1962) 171–179.

[6] Ghelardoni, P. *Approximations of Sturm-Liouville eigenvalues using Boundary Value Methods*, Journal of Computational and Applied Mathematics, Volume 23, Issue 3, May 1997, Pages 311-325.

[7] Mercier, B., Osborn, J., Rappaz, J. and Raviart, P.A. *Eigenvalues approximation by mixed and hybrid methods*, Math. Comp. 36 (154) (1981) 427–453.

[8] Mikhlin, S.G. *Variational methods in mathematical physics*, Pergamon Press, Oxford, England, 1964.

[9] Prikazchikov, V.G. *High-accuracy homogeneous difference schemes for the Sturm-Liouville problem*, JUSSR Comput. Math. Math. Phys. 9 (2) (1964) 76–106.

[10] Vainiko, G.M. *Asymptotic evaluations of the error of projection methods for the eigenvalue problem*, USSR Comput. Math. Math. Phys. 4 (3) (1964) 9–36.

Iranian Journal of Numerical Analysis and Optimization
Vol. 15, No. 2, 2025, pp 655–675
https://doi.org/10.22067/ijnao.2025.90483.1543
https://ijnao.um.ac.ir/

Research Article

Explicit collocation algorithm for the nonlinear fractional Duffing equation via third-kind Chebyshev polynomials

Y.H. Youssri*, A.G. Atta, M.O. Moustafa and Z.Y. Abu Waar

Abstract

Herein, we propose an accurate algorithm to approximate the solution of the nonlinear fractional-order Duffing equation (NFDE). The algorithm is

*Corresponding author

Received 27 October 2024; revised 3 Janurary 2025; accepted 4 Janurary 2025

Youssri Hassan Youssri
Department of Mathematics, Faculty of Science, Cairo University, Giza 12613, Egypt. e-mail: youssri@cu.edu.eg

Ahmed Gamal Atta
Department of Mathematics, Faculty of Education, Ain Shams University, Roxy, Cairo 11341, Egypt. e-mail: ahmed_gamal@edu.asu.edu.eg

Mohamed Orabi Moustafa
Department of Physics, School of Science and Engineering, The American University in Cairo (AUC), New Cairo 11835, Egypt. e-mail: mohamed.orabi@aucegypt.edu

Ziad Yousef Abu Waar
Department of Physics, College of Science, The University of Jordan, Amman, 11942, Jordan. e-mail: ziad.abuwaar@ju.edu.jo

How to cite this article
Youssri, Y.H., Atta, A.G., Moustafa, M.O. and Abu Waar, Z.Y., Explicit collocation algorithm for the nonlinear fractional Duffing equation via third-kind Chebyshev polynomials. *Iran. J. Numer. Anal. Optim.*, 2025; 15(2): 655-675. https://doi.org/10.22067/ijnao.2025.90483.1543

based on using shifted Chebyshev polynomials of the third-kind as basis functions and the spectral collocation method as a solver. We study the error analysis of the method in-depth, and we exhibit some numerical test problems to check the applicability of the method. Also, we compare it with other existing techniques to show the superiority of our proposed numerical scheme. Our results show that the method employed provides a useful tool to simulate the solution of the NFDE. The main advantages of the proposed method are that it does not require a huge number of retained modes, simply a few terms, and does not exhaust the machine used to render the codes.

AMS subject classifications (2020): 80M22, 33D50, 40A05, 34A08.

Keywords: Chebyshev polynomials; Spectral methods; Nonlinear fractional-order Duffing equation; Convergence analysis.

1 Introduction

Within the fields of mathematics, physics, and engineering, joint efforts have produced a new instrument for characterizing difficult problems: the area of fractional calculus. This field of study allows scientists to accurately handle real-world issues in a variety of domains. Fractional-order differential equations can characterize application issues in physics, biology, mechanics, medicine, astronomy, engineering, and chemistry. Due to the importance of finding solutions to these equations, a number of numerical techniques have been used, including the Laplace transform approach, Chebyshev collocation, domain decomposition technique, differential transformation technique, spectral Legendre method, Tau procedure, variational iteration method, and the wavelet method, which are all numerical techniques used in solving differential equations and related problems. When it comes to breaking down fractional-order problems into systems of algebraic equations, the spectral Tau technique and the operational matrix (OM) method stand out for their simplicity, speed of convergence, and computational ease. Selecting appropriate basis functions is essential in order to use spectral approaches to achieve approximate solutions. Because of their useful features and trigonometric

representation, the Chebyshev polynomials of the first kind are one class of functions that have shown to be efficient in this respect. Those who are interested in learning more about Chebyshev polynomials might consult [12]. It is known that Chebyshev polynomials have been used more often in spectral techniques to solve different kinds of differential problems; recent works like [2, 5, 23, 14, 20] attest to this.

A specific class of problems in nonlinear dynamics is the one-dimensional Duffing equation, capturing the behavior of a simple oscillator with cubic nonlinearity. In the context of fractional calculus, the nonlinear fractional-order Duffing equation (NFDE) takes the form:

$$D_t^{\mu}\chi(t) + a\, D_t^{\beta}\chi(t) + b\,\chi(t) + c\,\chi^2(t) + d\,\chi^5(t) + e\,\chi^7(t) = f(t), \qquad (1)$$

subject to the simple initial conditions $\chi(0) = \chi_0, \quad \chi'(0) = \chi_1$, where μ is in the range $(1, 2]$ and β in the range $(0, 1]$.

Spectral techniques have expanded to solve fractional-order differential equations in recent years [9, 10, 13, 27, 30, 31, 4, 3, 1], providing a nuanced description of physical phenomena through non-integer derivatives. The inherent characteristics of Chebyshev polynomials provide spectral approaches for such equations, improving numerical efficiency and enabling the reliable approximation of fractional derivatives. Spectral collocation is one of these techniques that stands out for its excellent accuracy and convergence. A very accurate and convergent computational approach for solving nonlinear Volterra integrodifferential equations of high order was presented by the authors in [24]. The spectral collocation approach was used by Thirumalai et al. [21] to solve nonlinear high-order pantograph equations.

As a fundamental spectral approach for solving partial differential equations (PDEs), the Galerkin method is emphasized in [26, 28, 32]. With this approach, the unknown function is expressed as a PDE using a series expansion, and the derivatives of the PDE are then approximated by using the collocation technique [29]. When working with linear or nonlinear PDEs, particularly those that have beginning or boundary conditions, this kind of method is quite helpful.

The NFDE, a well-established nonlinear equation, serves as a potent tool for addressing significant practical phenomena in applied science [19]. This

equation gained prominence in the mid-twentieth century, particularly in the study of electronics [18]. Functioning as the simplest oscillator, it exhibits catastrophic increases in amplitude and phase when the frequency of the forcing term varies gradually. The Duffing equation finds wide application in diverse areas, including brain modeling [33], passive islanding detection in inverter-based distributed generation units using Duffing oscillators [22], propagation of electromagnetic pulses in nonlinear media, radar systems, digital communication [7], nonlinear electrical circuits [17], and other fields. Numerous attempts have been made to numerically solve the Duffing equation [15, 11].

The suggested approach presents a novel use of shifted Chebyshev polynomials of the third-kind (SCP3K) in collocation methods, providing an accurate, efficient and fast numerical scheme for spectrally solving the NFDE. It addresses important difficulties in computational complexity and stability, employing few computer resources and greatly improving solution accuracy when compared to existing approaches. To the best of our knowledge, the main contribution and novelty of this paper can be listed in the following points:

- A new theoretical background to the SCP3K is presented.
- Establishing a new OM of fractional derivatives for SCP3K. This matrix is considered an important tool for treating NFDE.
- The study of the error bound is new.

The advantages of the presented approach are as follows:

- By choosing SCP3K as basis functions, a few terms of the retained modes make it possible to produce approximations with excellent precision.
- Less calculation is required to obtain the desired approximate solution.

The following is how this job is organized: In Section 2, the definitions and mathematical equations of SCP3K are presented; OM of fractional order is constructed using SCP3K along with a brief introduction to the tools

of fractional calculus. The problem formulation and the proposed numerical solution utilizing derived OM are given in Section 3. For the proposed Chebyshev expansion, Section 4 provides error estimation in L_∞ space. Test cases in Section 5 illustrate the precision and effectiveness of the suggested approach. In conclusion, Section 6 provides some final thoughts.

2 Preliminaries and essential relations

2.1 The fractional derivative in the Caputo sense

Definition 1. [16] The Caputo fractional derivative of order s is defined as

$$D^s u(t) = \frac{1}{\Gamma(m-s)} \frac{d}{dt} \int_0^t (t-y)^{m-s-1} u^{(m)}(y)\, dy, \quad s > 0, \quad t > 0, \tag{2}$$

where $m-1 \leqslant s < m, \quad m \in \mathbb{N}$.

The following properties are satisfied by the operator D^s for $m-1 \leqslant s < m, \quad m \in \mathbb{N}$,

$$D^s c = 0, \ (c\text{is a constant}) \tag{3}$$

$$D^s\, t^r = \begin{cases} 0, & \text{if } r \in \mathbb{N}_0 \quad \textit{and} \quad r < \lceil s \rceil, \\ \frac{r!}{\Gamma(r-s+1)}\, t^{r-s}, & \text{if } r \in \mathbb{N}_0 \quad \textit{and} \quad r \geq \lceil s \rceil, \end{cases} \tag{4}$$

where $\mathbb{N} = \{1, 2, 3, \ldots\}$, $\mathbb{N}_0 = \{0\} \cup \mathbb{N}$, and the notation $\lceil s \rceil$ denotes the ceiling function.

2.2 A account on the SCP3K

The SCP3K $\mathcal{V}_{S,\ell}(t)$ are special ones of the Jacobi polynomials that can be defined as [25]

$$\mathcal{V}_{S,\ell}(t) = \frac{2^{2\,\ell}}{\binom{2\,\ell}{\ell}} P_\ell^{\left(-\frac{1}{2},\frac{1}{2}\right)}(2\,t-1). \tag{5}$$

The orthogonality relation of $\mathcal{V}_{S,\ell}(t)$ is given by

$$\int_0^1 \mathcal{V}_{S,\ell}(t)\,\mathcal{V}_{S,j}(t)\,\omega(t)dt = \frac{\pi}{2}\,\delta_{\ell,j}, \tag{6}$$

where $\omega(t) = \left(\frac{t}{1-t}\right)^{\frac{1}{2}}$ and $\delta_{\ell,j}$ is the Kronecker delta.

The analytic representation of $\mathcal{V}_{S,\ell}(t)$ can be represented as

$$\mathcal{V}_{S,m}(t) = \sum_{\ell=0}^{m} \frac{(-1)^{\ell}\,(2\,m+1)\,2^{2\,m-2\,\ell}(-\ell+2m)!}{\ell!\,(-2\,\ell+2\,m+1)!}\,t^{m-\ell}, \tag{7}$$

or

$$\mathcal{V}_{S,m}(t) = \sum_{\ell=0}^{m} \frac{2^{2\,\ell}\,(2\,m+1)\,(-1)^{m-\ell}\,(\ell+m)!}{(2\,\ell+1)!\,(-\ell+m)!}\,t^{\ell}. \tag{8}$$

Moreover, the inversion formula of $\mathcal{V}_{S,\ell}(t)$ is

$$t^{r} = \frac{\Gamma(2\,r+2)}{2^{2\,r}} \sum_{\ell=0}^{r} \frac{1}{(-\ell+r)!\,(\ell+r+1)!}\,\mathcal{V}_{S,\ell}(t). \tag{9}$$

Theorem 1. For $0 < \beta < 1$, we have

$$D^{\beta}\mathcal{V}_{S,\ell}(t) = \boldsymbol{\mathcal{A}}\,\boldsymbol{\mathcal{V}}, \tag{10}$$

where

$\boldsymbol{\mathcal{V}} = [\mathcal{V}_{S,0}(t), \mathcal{V}_{S,1}(t), \ldots, \mathcal{V}_{S,M}(t)]^{T}$. Also, $\boldsymbol{\mathcal{A}} = (\mathcal{A}_{\ell,k})$ is an OM of ordinary derivatives of dimension $(M+1)\times(M+1)$, and the entries of this matrix can be written in the form

$$\mathcal{A}_{\ell,k} = \sum_{n=1}^{k} \frac{4^{n}\,(2\,\ell+1)\,(2\,k+1)(-1)^{-n+\ell+k}\,(n+k)!\,\Gamma(n+1)\,\Gamma\left(n-\beta+\frac{3}{2}\right)}{\Gamma(2\,n+2)\,\Gamma(1-n+k)\,\Gamma\left(n-\beta+1\right)} \times {}_3\tilde{F}_2\left(\left.\begin{matrix}-\ell,\ell+1,-\beta+n+\frac{3}{2}\\ \frac{3}{2},-\beta+n+2\end{matrix}\right|1\right). \tag{11}$$

Proof. Using Definition 1 jointly with the analytic power form of $\mathcal{V}_{S,n}(t)$, we have

$$D^{\beta}\mathcal{V}_{S,\kappa}(t) = \sum_{n=1}^{\kappa} \frac{4^{n}\,(2\,\kappa+1)\,(-1)^{\kappa-n}\,(\kappa+n)!\,\Gamma(n+1))}{(2\,n+1)!\,(-n+\kappa)!\,\Gamma\left(n-\beta+1\right)}\,t^{n-\beta}. \tag{12}$$

Now, $t^{n-\beta}$ can be approximated as

$$t^{n-\beta} = \sum_{\ell=0}^{M} c_{\ell}\,\mathcal{V}_{S,\ell}(t), \tag{13}$$

where

$$c_\ell = \frac{2}{\pi} \int_0^1 t^{n-\beta}\, \mathcal{V}_{S,\ell}(t)\, \omega(t)\, dt. \tag{14}$$

Integrating the RHS of the last equation, we have

$$c_\ell = \sum_{s=0}^{\ell} \frac{2\,(2\,\ell+1)\,2^{2\,s}\,(-1)^{\ell-s}\,(\ell+s)!\,\Gamma\left(n+s-\beta+\frac{3}{2}\right)}{\sqrt{\pi}\,(2\,s+1)!\,(\ell-s)!\,(i+s-\beta+1)!}. \tag{15}$$

Equation (15) can be written as

$$c_\ell = (-1)^\ell\,(2\,\ell+1)\,\Gamma\left(n-\beta+\frac{3}{2}\right)\,{}_3\tilde{F}_2\left(\begin{matrix}-\ell, \ell+1, -\beta+n+\frac{3}{2}\\ \frac{3}{2}, -\beta+n+2\end{matrix}\,\middle|\, 1\right). \tag{16}$$

Inserting (16) into (13) yields

$$\begin{aligned} t^{n-\beta} = &\sum_{\ell=0}^{M} \left((-1)^\ell\,(2\,\ell+1)\,\Gamma\left(n-\beta+\frac{3}{2}\right)\right.\\ &\left.\times\,{}_3\tilde{F}_2\left(\begin{matrix}-\ell, \ell+1, -\beta+n+\frac{3}{2}\\ \frac{3}{2}, -\beta+n+2\end{matrix}\,\middle|\, 1\right)\right) \mathcal{V}_{S,\ell}(t). \end{aligned} \tag{17}$$

Plugging (17) into (12), we obtain

$$D^\beta \mathcal{V}_{S,\kappa}(t) = \sum_{\ell=0}^{M} \mathcal{A}_{\ell,\kappa}\, \mathcal{V}_{S,\ell}(t) = \boldsymbol{\mathcal{A}}\,\boldsymbol{\mathcal{V}},$$

where

$$\begin{aligned} \mathcal{A}_{\ell,\kappa} = &\sum_{n=1}^{\kappa} \frac{4^n\,(2\,\ell+1)\,(2\,\kappa+1)(-1)^{-n+\ell+\kappa}\,(n+\kappa)!\,\Gamma(n+1)\,\Gamma\left(n-\beta+\frac{3}{2}\right)}{\Gamma(2\,n+2)\,\Gamma(1-n+\kappa)\,\Gamma\,(n-\beta+1)}\\ &\times\,{}_3\tilde{F}_2\left(\begin{matrix}-\ell, \ell+1, -\beta+n+\frac{3}{2}\\ \frac{3}{2}, -\beta+n+2\end{matrix}\,\middle|\, 1\right). \end{aligned} \tag{18}$$

Therefore, we conclude that

$$D^\beta \mathcal{V}_{S,\kappa}(t) = \boldsymbol{\mathcal{A}}\,\boldsymbol{\mathcal{V}},$$

where $\boldsymbol{\mathcal{A}} = (\mathcal{A}_{\ell,\kappa})$ is an OM of derivatives of order $(M+1)\times(M+1)$. □

Theorem 2. For $1 < \mu < 2$, we have

$$D^\mu \mathcal{V}_{S,\ell}(t) = \boldsymbol{\mathcal{B}}\,\boldsymbol{\mathcal{V}}, \tag{19}$$

where $\boldsymbol{\mathcal{V}} = [\mathcal{V}_{S,0}(t), \mathcal{V}_{S,1}(t), \ldots, \mathcal{V}_{S,M}(t)]^T$. Also, $\boldsymbol{\mathcal{B}} = (\mathcal{B}_{\ell,k})$ is the OM of derivatives with dimension $(M+1)\times(M+1)$, and the entries of this matrix can be written in the form

$$\mathcal{B}_{\ell,k} = \sum_{n=2}^{k} \frac{4^n\,(2\,\ell+1)\,(2\,k+1)(-1)^{-n+\ell+k}\,(n+k)!\,\Gamma(n+1)\,\Gamma\left(n-\mu+\frac{3}{2}\right)}{\Gamma(2\,n+2)\,\Gamma(1-n+k)\,\Gamma\left(n-\mu+1\right)} \times {}_3\tilde{F}_2\left(\begin{matrix}-\ell, \ell+1, -\mu+n+\frac{3}{2} \\ \frac{3}{2}, -\mu+n+2\end{matrix}\middle| 1\right). \tag{20}$$

Proof. The proof of this theorem is similar to the proof of Theorem 1. □

3 Collocation algorithm for the NFDE

In this section, we consider the following NFDE [8]:

$$D_t^{\mu}\chi(t) + a\,D_t^{\beta}\chi(t) + b\,\chi(t) + c\,\chi^2(t) + d\,\chi^5(t) + e\,\chi^7(t) = f(t), \tag{21}$$

$$\mu \in]1,2], \quad \beta \in]0,1], \quad t \in [0,1],$$

subject to the conditions

$$\chi(0) = g_1, \quad \chi'(0) = g_2. \tag{22}$$

The following set $\{\mathcal{V}_{S,i}(t), i : 0,1,2,\ldots\}$ forms an orthogonal basis of $L^2_{\omega(t)}(0,1)$. This means that for any given function $\chi(t) \in L^2_{\omega(t)}(0,1)$, one has

$$\chi(t) = \sum_{i=0}^{\infty} \lambda_i\,\mathcal{V}_{S,i}(t), \tag{23}$$

and approximated as

$$\chi(t) \approx \chi^M(t) = \sum_{i=0}^{M} \lambda_i\,\mathcal{V}_{S,i}(t) = \boldsymbol{\lambda}\,\boldsymbol{\mathcal{V}}, \tag{24}$$

where

$$\boldsymbol{\lambda} = [\lambda_0, \lambda_1, \ldots, \lambda_M], \tag{25}$$

and

$$\boldsymbol{\mathcal{V}} = [\mathcal{V}_{S,0}(t), \mathcal{V}_{S,1}(t), \ldots, \mathcal{V}_{S,M}(t)]^T. \tag{26}$$

By virtue of Theorems 1 and 2, the residual $\mathbf{Res}(t)$ of (21) is given by

$$\begin{aligned}\mathbf{Res}(t) =& D_t^{\mu}\chi^M(t) + a\, D_t^{\beta}\chi^M(t) + b\,\chi^M(t) + c\,\left(\chi^M(t)\right)^2 + d\,\left(\chi^M(t)\right)^5 \\ &+ e\,\left(\chi^M(t)\right)^7 - f(t) \\ =& \boldsymbol{\lambda}\boldsymbol{\mathcal{B}}\boldsymbol{\mathcal{V}} + a\,\boldsymbol{\lambda}\boldsymbol{\mathcal{A}}\boldsymbol{\mathcal{V}} + b\,\boldsymbol{\lambda}\boldsymbol{\mathcal{V}} + c\,(\boldsymbol{\lambda}\boldsymbol{\mathcal{V}})^2 + d\,(\boldsymbol{\lambda}\boldsymbol{\mathcal{V}})^5 + e\,(\boldsymbol{\lambda}\boldsymbol{\mathcal{V}})^7 - f(t).\end{aligned} \tag{27}$$

Now, the application of the standard collocation method yields to the following $(M+1)$ algebraic system of equations in the unknown expansion coefficients b_i

$$\begin{aligned}&\mathbf{Res}(t_i) = 0, \quad i = 1, 2, ..., M-1, \\ &\boldsymbol{\lambda}\bar{\boldsymbol{\mathcal{V}}} = g_1, \quad \boldsymbol{\lambda}\hat{\boldsymbol{\mathcal{V}}} = g_2,\end{aligned} \tag{28}$$

where

$$\begin{aligned}\bar{\boldsymbol{\mathcal{V}}} &= [\mathcal{V}_{S,0}(0), \mathcal{V}_{S,1}(0), ..., \mathcal{V}_{S,M}(0)]^T, \\ \hat{\boldsymbol{\mathcal{V}}} &= [\mathcal{V}'_{S,0}(0), \mathcal{V}'_{S,1}(0), ..., \mathcal{V}'_{S,M}(0)]^T,\end{aligned} \tag{29}$$

and $\{t_i : i = 1, 2, \ldots, M\}$ are the first M distinct roots of $\mathcal{V}_{S,M}(t)$. Therefore, the system in (28) can be solved to get λ_i with the aid of the well-known Newton's iterative method.

4 Error analysis

In this section, we study the error analysis of the numerical solution $\chi^M(t)$ to the exact solution $\chi(t)$ of (21) by imitating similar steps as in [29, 6] in L_∞- norm

$$L_\infty[0,1] = \{\chi : \|\chi\|_\infty = \max_{t\in[0,1]} |\chi| < \infty\}.$$

Consider the following space function:

$$L^M[0,1] = \text{span}\{\mathcal{V}_{S,i}(t) : i = 0, 1, \ldots, M\}. \tag{30}$$

Assume that $\chi^M(t) \in L^M[0,1]$ is the best approximation of $\chi(t)$; then, by the definition of the best approximation, we have

$$||\chi(t) - \chi^M(t)||_\infty \leq ||\chi(t) - \mathfrak{u}^M(t)||_\infty, \quad \text{for all } \mathfrak{u}^M(t) \in L^M[0,1]. \tag{31}$$

It turns out that the previous inequality is also true if $\mathfrak{u}^M(t)$ denotes the interpolating polynomial for $\chi(t)$ at points t_i, where t_i are the roots of $\mathcal{V}_{S,i}(t)$. Now,

$$\chi(t) - \mathfrak{u}^M(t) = \frac{d^{M+1}\,\chi(s)}{d\,t^{M+1}\,(M+1)!}\,\prod_{i=0}^{M}(t-t_i), \tag{32}$$

where $s \in [0,1]$, and hence, one has

$$||\chi(t) - \mathfrak{u}^M(t)||_\infty \leq \max_{t\in[0,1]}\left|\frac{d^{M+1}\,\chi(s)}{d\,t^{M+1}}\right|\,\frac{||\prod_{i=0}^{M}(t-t_i)||_\infty}{(M+1)!}. \tag{33}$$

Since $\chi(t)$ is a smooth function on $[0,1]$, then there exist a constant n such that

$$\max_{t\in[0,1]}\left|\frac{d^{M+1}\,\chi(s)}{d\,t^{M+1}}\right| \leq n. \tag{34}$$

Now, our main aim is to minimize the factor $||\prod_{i=0}^{M}(t-\iota_i)||_\infty$.

Assuming the one-to-one mapping $t = \frac{1}{2}\,(t+1)$ between the intervals $[-1,1]$ and $[0,1]$ to deduce that

$$\begin{aligned}\min_{t_i\in[0,1]}\max_{t\in[0,1]}\left|\prod_{i=0}^{M}(t-t_i)\right| &= \min_{t_i\in[-1,1]}\max_{t\in[-1,1]}\left|\prod_{i=0}^{M}\frac{1}{2}\,(t-t_i)\right| \\ &= \left(\frac{1}{2}\right)^{M+1}\min_{t_i\in[-1,1]}\max_{t\in[-1,1]}\left|\prod_{i=0}^{M}(t-t_i)\right| \\ &= \left(\frac{1}{2}\right)^{M+1}\min_{t_i\in[-1,1]}\max_{t\in[-1,1]}\left|\frac{\mathcal{V}_{M+1}(t)}{\eta^M}\right|,\end{aligned} \tag{35}$$

where $\eta^M = 2^{M+1}$ is the leading coefficient of Chebyshev polynomials of the third kind $\mathcal{V}_{M+1}(t)$ and t_i are the roots of $\mathcal{V}_{M+1}(t)$.

It is known that

$$\max_{t\in[-1,1]}|\mathcal{V}_{M+1}(t)| = \mathcal{V}_{M+1}(1) = 1. \tag{36}$$

In virtue of inequality (34), equations (35), and (36), we get

$$||\chi(t) - \chi^M(t)||_\infty \leq \frac{n}{2^{2\,(M+1)}\,(M+1)!}. \tag{37}$$

The previous inequality gives us an upper bound estimation of the absolute error (AE). This completes the deducing of our error estimation.

5 Illustrative examples

To demonstrate the effectiveness and validity of the proposed algorithm, the procedure will be illustrated using several numerical examples. Let us define the following L_∞-error:

$$L_\infty = \max_{t\in[0,1]} \left|\chi(t) - \chi^M(t)\right|.$$

Test Problem 1. [8] Consider the following NFDE:

$$D_t^\mu \chi(t) + a\, D_t^\beta \chi(t) + b\,\chi(t) + c\,\chi^2(t) + d\,\chi^5(t) + e\,\chi^7(t) = \frac{1}{8} e^{-3\,t}\left(3 - 20\, e^{2\,t}\right), \tag{38}$$

subject to the conditions

$$\chi(0) = \frac{1}{2}, \quad \chi'(0) = \frac{-1}{2}, \tag{39}$$

where the exact solution is $\chi(t) = \frac{e^{-t}}{2}$ at $\mu = 2$, $\beta = 1$, $a = 4$, $b = -2$, $c = 3$, $d = 0$, $e = 0$.

Table 1 presents a comparison between our method at $M = 13$ and the method in [8] at $M = 12$, demonstrating the accuracy of our method. Also, Table 2 shows the L_∞-error at $\mu = 2$, $\beta = 1$. Figure 1 shows the AE when $\mu = 2$, $\beta = 1$ at different values of M. Figure 2 illustrates that the approximate solutions have smaller variations for values of μ and β near $\mu = 2$, $\beta = 1$ when $M = 12$. Figure 3 confirms that the method remains stable at $\mu = 2$, $\beta = 1$ for higher values of M, with any indication of numerical instability or divergence in error.

Table 1: Comparison of AE of Example 1 at $\mu = 2$, $\beta = 1$.

t	Method in [8] at $M = 12$	Our method at $M = 13$	Our CPU time
0.1	5.5511×10^{-17}	0	
0.2	0	0	
0.3	5.5511×10^{-17}	0	
0.4	1.1102×10^{-16}	5.55112×10^{-17}	
0.5	0	0	0.952
0.6	5.5511×10^{-17}	$5.55112\times^{-17}$	
0.7	2.7756×10^{-17}	0	
0.8	$8.3267\times^{-17}$	0	
0.9	2.7756×10^{-17}	0	

Table 2: The L_∞-error of Example 1 at $\mu = 2$, $\beta = 1$.

M	3	6	9	12
Error	2.54449×10^{-4}	3.85844×10^{-8}	1.06212×10^{-12}	5.55112×10^{-17}

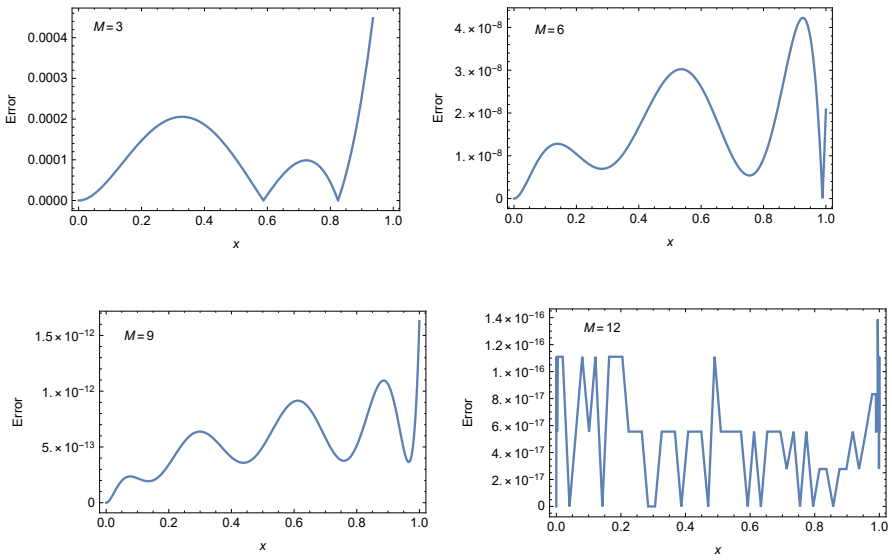


Figure 1: The AE of Example 1 at $\mu = 2$, $\beta = 1$.

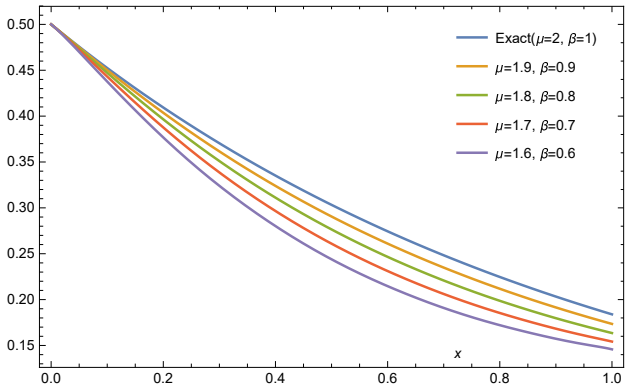


Figure 2: Different solutions of Example 1 at $M = 12$ and different values of μ, β.

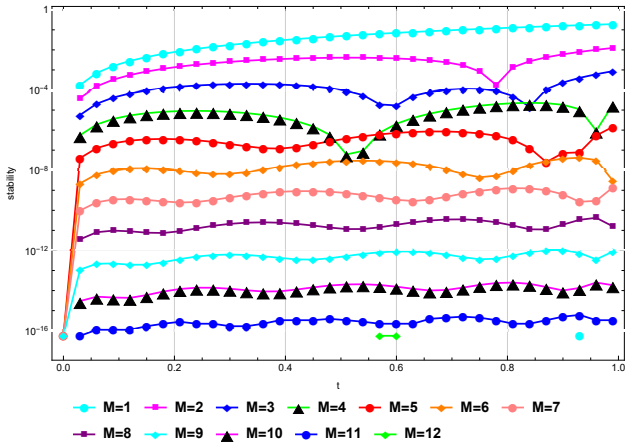


Figure 3: Stability $|\chi^{M+1}(t) - \chi^{M}(t)|$ at $\mu = 2$, $\beta = 1$ for Example 1.

Test Problem 2. [8] Consider the following NFDE:

$$D_t^{\mu}\chi(t)+a\,D_t^{\beta}\chi(t)+b\,\chi(t)+c\,\chi^2(t)+d\,\chi^5(t)+e\,\chi^7(t) = -2\,\sin(t)+\cos^5(t)+8\,\cos^3(t), \tag{40}$$

subject to the conditions

$$\chi(0) = 1, \quad \chi'(0) = 0, \tag{41}$$

where the exact solution is $\chi(t) = \cos(t)$, at $\mu = 2$, $\beta = 1$, $a = 2$, $b = 1$, $c = 8$, $d = 1$, $e = 0$.

Table 3 presents a comparison between our method at $M = 13$ and the method in [8] at $M = 11$, showing the accuracy of our method. Figure 4 shows AE when $\mu = 2$, $\beta = 1$ at different values of M. Figure 5 illustrates that the approximate solutions have smaller variations for values of μ and β near $\mu = 2$, $\beta = 1$ when $M = 12$. Figure 6 confirms that the method remains stable at $\mu = 2$, $\beta = 1$ for higher values of M, with no indication of numerical instability or divergence in error.

Table 3: Comparison of AE of Example 2 at $\mu = 2$, $\beta = 1$.

t	Method in [8] at $M = 11$	Our method at $M = 13$	Our CPU time
0.1	5.9730×10^{-14}	2.22045×10^{-16}	
0.2	1.0669×10^{-13}	0	
0.3	1.1813×10^{-13}	1.11022×10^{-16}	
0.4	9.9032×10^{-14}	1.11022×10^{-16}	
0.5	6.1506×10^{-14}	0	0.891
0.6	1.8208×10^{-14}	0	
0.7	2.0650×10^{-14}	0	
0.8	$4.9738\times^{-14}$	1.11022×10^{-16}	
0.9	6.7835×10^{-14}	1.11022×10^{-16}	

Test Problem 3. Consider the following NFDE:

$$D_t^{\mu}\chi(t) + a\, D_t^{\beta}\chi(t) + b\,\chi(t) + c\,\chi^2(t) + d\,\chi^5(t) + e\,\chi^7(t) = f(t), \tag{42}$$

subject to the conditions

$$\chi(0) = \chi'(0) = 0, \tag{43}$$

where $f(t)$ is chosen such that the exact solution is $\chi(t) = t^{\beta+\mu}$ at $a = 1$, $b = 1$, $c = 1$, $d = 1$, $e = 1$.

Table 4 presents the AE at different values of μ, β when $M = 16$. Figure 7 shows AE when $\mu = 1.8$, $\beta = 0.7$ at different values of M.

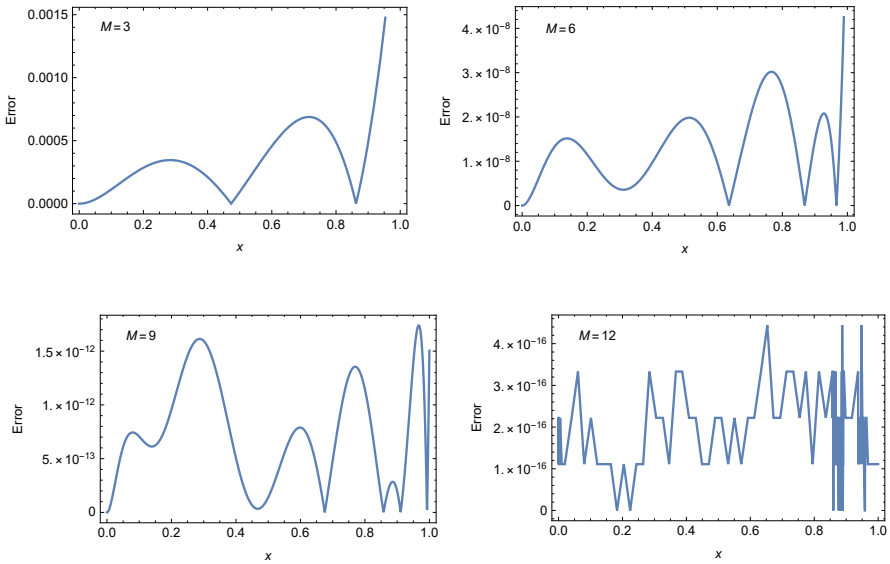


Figure 4: The AE of Example 2 at $\mu = 2$, $\beta = 1$.

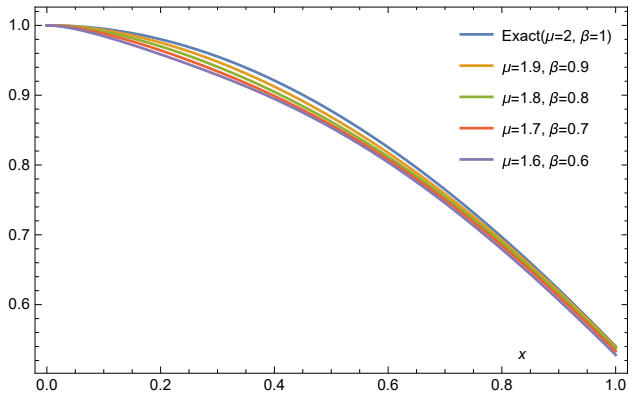


Figure 5: Different solutions of Example 2 at $M = 12$ and different values of μ, β.

6 Concluding Remarks

In this study, we used SCP3K and the spectral collocation method to solve the NFDE. We evaluated the approach with numerical test examples and carried out a thorough error analysis, contrasting it with other approaches that were already in use. Our findings demonstrated that the suggested

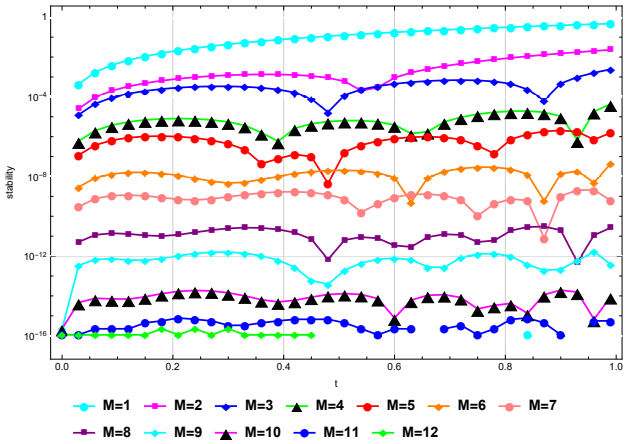


Figure 6: Stability $|\chi^{M+1}(t) - \chi^{M}(t)|$ at $\mu = 2$, $\beta = 1$ for Example 2.

Table 4: The AE of Example 3 at $M = 16$.

t	$\mu = 1.3$, $\beta = 0.3$	CPU time	$\mu = 1.6$, $\beta = 0.6$	CPU time	$\mu = 1.8$, $\beta = 0.7$	CPU time
0.1	1.37101×10^{-4}		9.05091×10^{-6}		5.27826×10^{-6}	
0.2	1.41098×10^{-4}		1.152×10^{-5}		8.17987×10^{-6}	
0.3	7.39605×10^{-5}		1.26987×10^{-5}		1.04948×10^{-5}	
0.4	1.48116×10^{-4}		1.50502×10^{-5}		1.2503×10^{-5}	
0.5	5.9031×10^{-5}	2	1.42085×10^{-5}	1.969	1.35739×10^{-5}	1.609
0.6	1.25943×10^{-4}		1.5650×10^{-5}		1.47554×10^{-5}	
0.7	4.4265×10^{-5}		1.39166×10^{-5}		1.50037×10^{-5}	
0.8	5.48668×10^{-5}		1.40047×10^{-5}		1.53348×10^{-5}	
0.9	1.11037×10^{-4}		1.2731×10^{-5}		1.47999×10^{-5}	

approach, which only needs a small number of terms and little computer power, can successfully simulate the answer.

Conflict of interest

The authors declare that they have no conflict of interest.

Funding

The authors received no financial support for the research.

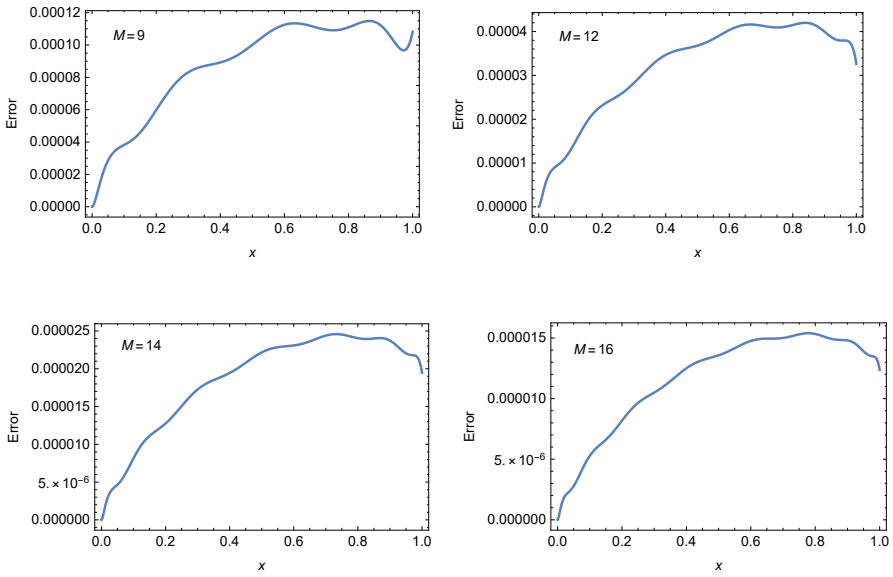


Figure 7: The AE of Example 3 at $\mu = 1.8$, $\beta = 0.7$.

Data availibility

No data is associated with this research.

Author Contributions Statement

YHY conducted the mathematical analysis, developed the methodology, verified the results, wrote the initial draft, and reviewed the final version. AGA contributed to the original manuscript, software development, and methodology. MM and ZYA reviewed and edited the final version of the manuscript.

References

[1] Abd-Elhameed, W.M., Youssri, Y.H., and Atta, A.G. *Tau algorithm for fractional delay differential equations utilizing seventh-kind Chebyshev polynomials,* J. Math. Model., (2024), 277–299.

[2] Amodio, P., Brugnano, L., and Iavernaro, F. *(Spectral) Chebyshev col-*

location methods for solving differential equations, Numer. Algorithms, (2023), 1–26.

[3] Atta, A.G., Abd-Elhameed, W.M., Moatimid, G.M., and Youssri, Y.H. *Modal shifted fifth-kind Chebyshev tau integral approach for solving heat conduction equation,* Fract. Fract., 6(11) (2022), 619.

[4] Atta, A.G., Abd-Elhameed, W.M., Moatimid, G.M., and Youssri, Y.H. *Novel spectral schemes to fractional problems with nonsmooth solutions,* Math. Meth. Appl. Sci., 46(13) (2023), 14745–14764.

[5] Atta, A.G. and Youssri, Y.H. *Advanced shifted first-kind Chebyshev collocation approach for solving the nonlinear time-fractional partial integro-differential equation with a weakly singular kernel,* Comput. Appl. Math., 41(8) (2022), 381.

[6] Bhrawy, A.H. and Zaky, M.A. *A method based on the Jacobi tau approximation for solving multi-term time-space fractional partial differential equations,* J. Comput. Phys., 281 (2015), 876–895.

[7] Deng, X.Y., Liu, B., and Long, T. *A new complex Duffing oscillator used in complex signal detection,* Chin. Sci. Bull., 57 (2012), 2185–2191.

[8] El-Sayed, A.A.E. *Pell-Lucas polynomials for numerical treatment of the nonlinear fractional-order Duffing equation,* Demonstr. Math., 56(1) (2023), 2022.

[9] Habibirad, A., Hesameddini, E., Azin, H., and Heydari, M.H. *The direct meshless local Petrov-Galerkin technique with its error estimate for distributed-order time fractional cable equation,* Eng. Anal. Bound. Elem., 150 (2023), 342–352.

[10] Hafez, R.M., Youssri, Y.H., and Atta, A.G. *Jacobi rational operational approach for time-fractional sub-diffusion equation on a semi-infinite domain,* Contemp. Math., (2023), 853–876.

[11] Maldon, B.J., Lamichhane, B.P., and Thamwattana, N. *Numerical solutions for nonlinear partial differential equations arising from modelling dye-sensitized solar cells,* ANZIAM J., 60 (2018), C231–C246.

[12] Mason, J.C. and Handscomb, D.C. *Chebyshev polynomials,* Chapman & Hall/CRC, 2002.

[13] Mortensen, M. *A generic and strictly banded spectral Petrov–Galerkin method for differential equations with polynomial coefficients,* SIAM J. Sci. Comput., 45(1) (2023), A123–A146.

[14] Moustafa, M., Youssri, Y.H., and Atta, A.G. *Explicit Chebyshev Petrov–Galerkin scheme for time-fractional fourth-order uniform Euler–Bernoulli pinned–pinned beam equation,* Nonlinear Eng., 12(1) (2023), 20220308.

[15] Novin, R. and Dastjerd, Z.S. *Solving Duffing equation using an improved semi-analytical method,* Commun. Adv. Comput. Sci. Appl., 2 (2015), 54–58.

[16] Podlubny, I. *Fractional differential equations: an introduction to fractional derivatives, fractional differential equations, to methods of their solution and some of their applications,* Elsevier, 1998.

[17] Salas, A.H. and Castillo, J.E. *Exact solutions to cubic Duffing equation for a nonlinear electrical circuit,* Visión Electrónica, 8(1) (2014), 1–8.

[18] Stokes, J. *Nonlinear vibrations,* Interscience, New York, 1950.

[19] Tabatabaei, K. and Gunerhan, E. *Numerical solution of Duffing equation by the differential transform method,* Appl. Math. Inf. Sci. Lett., 2(1) (2014), 1–6.

[20] Taema, M.A. and Youssri, Y.H. *Third-kind Chebyshev spectral collocation method for solving models of two interacting biological species,* Contemp. Math., (2024) 6189–6207.

[21] Thirumalai, S., Seshadri, R., and Yuzbasi, S. *Spectral collocation method based on special functions for solving nonlinear high-order pantograph equations,* Comput. Meth. Differ. Equ., 11(3) (2023), 589–604.

[22] Vahedi, H., Gharehpetian, G.B., and Karrari, M. *Application of Duffing oscillators for passive islanding detection of inverter-based distributed generation units,* IEEE Trans. Power Deliv., 27(4) (2012), 1973–1983.

[23] Wang, H. *Analysis of error localization of Chebyshev spectral approximations,* SIAM J. Numer. Anal., 61(2) (2023), 952–972.

[24] Yang, Y., Tohidi, E., and Deng, G. *A high accurate and convergent numerical framework for solving high-order nonlinear Volterra integro-differential equations,* J. Comput. Appl. Math., 421 (2023), 114852.

[25] Youssri, Y.H., Abd-Elhameed, W.M., and Ahmed, H.M. *New fractional derivative expression of the shifted third-kind Chebyshev polynomials: Application to a type of nonlinear fractional pantograph differential equations,* J. Funct. Spaces, 2022 (2022) 3966135.

[26] Youssri, Y.H., Abd-Elhameed, W.M., and Atta, A.G. *Spectral Galerkin treatment of linear one-dimensional telegraph type problem via the generalized Lucas polynomials,* Arab. J. Math., 11(3) (2022), 601–615.

[27] Youssri, Y.H. and Atta, A.G. *Modal spectral Tchebyshev Petrov-Galerkin stratagem for the time-fractional nonlinear Burgers' equation,* Iran. J. Numer. Anal. Optim., 14(1) (2023) 172–199.

[28] Youssri, Y.H. and Atta, A.G. *Petrov-Galerkin Lucas polynomials procedure for the time-fractional diffusion equation,* Contemp. Math. (2023) 230–248.

[29] Youssri, Y.H. and Atta, A.G. *Spectral collocation approach via normalized shifted Jacobi polynomials for the nonlinear Lane–Emden equation with fractal-fractional derivative,* Fract. Fract., 7(2) (2023), 133.

[30] Youssri, Y.H. and Atta, A.G. *Radical Petrov–Galerkin approach for the time-fractional KdV–Burgers' equation,* Math. Comput. Appl., 29(6) (2024), 107.

[31] Youssri, Y.H., Atta, A.G., Abu Waar, Z.Y., and Moustafa, M.O. *Petrov–Galerkin method for small deflections in fourth-order beam equations in civil engineering,* Nonlinear Eng., 13(1) (2024), 20240022.

[32] Youssri, Y.H., Hafez, R.M., and Atta, A.G. *An innovative pseudo-spectral Galerkin algorithm for the time-fractional Tricomi-type equation,* Phys. Scr., 99(10) (2024), 105238.

[33] Zeeman, E.C. *Duffing's equation in brain modelling*, Bull. Inst. Math. Appl., 12 (1975), 207–214.

Iranian Journal of Numerical Analysis and Optimization
Vol. 15, No. 2, 2025, pp 676–703
https://doi.org/10.22067/ijnao.2025.90229.1534
https://ijnao.um.ac.ir/

Research Article

Sequential approximate optimality conditions for a constrained convex vector minimization problem and application to multiobjective fractional programming problem

A. Ed-dahdah*, M. Laghdir, M. Echchaabaoui and M. Mabrouk

*Corresponding author

Received 12 October 2024; revised 8 Janurary 2025; accepted 11 Janurary 2025

Ahmed Ed-dahdah
Department of Mathematics, Faculty of Sciences, University of Chouaib Doukkali, El jadida, Morocco. e-mail: ahmededdahdah@gmail.com

Mohamed Laghdir
Department of Mathematics, Faculty of Sciences, University of Chouaib Doukkali, El jadida, Morocco. e-mail: laghdirmm@gmail.com

El Mahjoub Echchaabaoui
Department of Mathematics, Faculty of Sciences, University of Chouaib Doukkali, El jadida, Morocco. e-mail: mahjoubmath@gmail.com

M'hamed Mabrouk
Department of Mathematics, Faculty of Sciences, University of Chouaib Doukkali, El jadida, Morocco. e-mail: mabrouk.mhamed@ucd.ac.ma

How to cite this article
Ed-dahdah, A., Laghdir, M., Echchaabaoui, M. and Mabrouk, M., Sequential approximate optimality conditions for a constrained convex vector minimization problem and application to multiobjective fractional programming problem. *Iran. J. Numer. Anal. Optim.*, 2025; 15(2): 676-703. https://doi.org/10.22067/ijnao.2025.90229.1534

Abstract

The aim of this paper is to establish sequential necessary and sufficient approximate optimality conditions for a constrained convex vector minimization problem without any constraint qualifications, characterizing the approximate proper and weak efficient solutions. The constraints are described by mappings taking values in different preorder vector spaces. Our approach is based essentially on the sequential approximate subdifferential calculus rule for the sums of a finite family of cone convex mappings. To illustrate our main result, an application to multiobjective fractional programming problem is given. Finally, we present an important subclass of such problems showing the applicability of the obtained conditions.

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

Keywords: Vector optimization; Approximate Pareto subdifferential of convex mappings; Sequential approximate optimality conditions.

1 Introduction

Many multiobjective fractional programming problems arose in different fields of modern research because their algorithmic aspects and abundance of applications, for example, economics, management, medicine, information theory, engineering, and optimal control (see [17, 2, 15, 16, 20] and the references therein).

Generally, for establishing the approximate optimality conditions for convex optimization problems, certain types of constraint qualifications must be imposed. However, we know that the constraint qualifications do not always hold even for finite-dimensional optimization problems and frequently fail for infinite-dimensional optimization problems. In order to eliminate these drawbacks, many authors have paid their attention to characterize optimality conditions for convex optimization problems without any constraint qualifications (see [19, 3, 8, 9, 1]). Thibault [19] derived sequential optimality conditions via the subdifferential calculus for convex functions with cone convex constraints. Boţ, Csetnek, and Wanka [3] obtained sequential characteri-

zations of an optimal solution for composed convex optimization problems. Recently, works have been done in this direction for multiobjective fractional programming problems (see [14, 13, 10, 11, 18]). Moustaid et al. [14, 13] established sequential optimality conditions for a constrained fractional programming problem without any constraint qualifications, characterizing the approximate weak efficient and efficient solution. Kim, Kim, and Lee [10] and Kohli [11] developed sequential optimality conditions for multiobjective fractional programming problems via scalarization and by using the concept of the epigraphs of conjugate functions in terms of approximate subdifferentials at an approximate weak efficient solution.

In this paper, we consider the following constrained vector minimization problem with inequality constraints, which are described by mappings taking values in different preorder vector spaces

$$(P_1) \qquad \begin{cases} \min f(x) \\ x \in C, \\ h_1(x) \in -Z_1^+, \\ \vdots \\ h_p(x) \in -Z_p^+, \end{cases}$$

where $(X, \|\cdot\|_X)$ and $(Z_i, \|\cdot\|_{Z_i})$ $(i = 1, \ldots, p)$ are real reflexive Banach spaces and Y is a real Hausdorff topological vector space, C, $Y_+ \subset Y$, and $Z_i^+ \subset Z_i$ $(i = 1, \ldots, p)$ are convex cones and $f : X \longrightarrow Y \cup \{+\infty_Y\}$ is a proper and Y_+-convex and $h_i : X \longrightarrow Z_i \cup \{+\infty_{Z_i}\}$ $(i = 1, \ldots, p)$ are proper and Z_i^+-convex mappings.

The goal of our paper is to establish, in the absence of constraint qualifications, sequential approximate optimality conditions characterizing properly and weakly approximate efficient solutions for the problem (P_1) in terms of the approximate subdifferentials of the associated functions.

This paper is organized as follows. In Section 2, we present the main notions and give some preliminary results used in what follows. In Section 3, we establish a sequential formula for the approximate subdifferential for the sums of two vector mapping with a sum of p $(p \geq 1)$ vector composites. In Section 4, by the results in Section 3, we study sequential necessary and

sufficient approximate optimality conditions for (P_1). Section 5 is devoted to calculus sequential approximate optimality conditions for multiobjective fractional programming problems without any constraint qualification. Finally, we present an example illustrating the main result of this work.

2 Notations, definitions, and preliminaries

In this section, we give some basic definitions and results. Let X, Y, and Z be real Hausdorff topological vector spaces with continuous dual spaces X^*, Y^*, and Z^* and duality pairing also denoted by $\langle\cdot,\cdot\rangle$. Let Y_+ be a convex cone of Y with $\text{int}Y_+ \neq \emptyset$. The subset $l(Y_+) := Y_+ \cap -Y_+$ is the linearity of Y_+, when it is null; thus the cone Y_+ is said to be pointed. Throughout the paper, we denote by $L(X,Y)$ the set of all continuous linear operators from X into Y. In what follows, the convex cone Y_+ is not supposed to be a linear subspace so it cannot coincide with its linearity. For any $v_1, v_2 \in Y$, the cone Y_+ induces the following ordering relations:

$$\begin{aligned} v_1 \leq_{Y_+} v_2 &\Longleftrightarrow v_2 - v_1 \in Y_+, \\ v_1 <_{Y_+} v_2 &\Longleftrightarrow v_2 - v_1 \in \text{int}Y_+, \\ v_1 \lneq_{Y_+} v_2 &\Longleftrightarrow v_2 - v_1 \in Y_+ \backslash l\,(Y_+)\,. \end{aligned}$$

To space Y, we attach an abstract maximal element with respect to "$\leq_{Y_+}$", denoted by $+\infty_Y$ such that $v \leq_{Y_+} +\infty_Y$, for all $v \in Y$ and $v + (+\infty_Y) := (+\infty_Y) + v := +\infty_Y$ for all $v \in Y \cup \{+\infty_Y\}$.

The polar cone Y_+^* and the strict polar cone $(Y_+^*)^\circ$ of Y_+ are defined, respectively, as

$$Y_+^* := \{y^* \in Y^* \ : \ y^*(Y_+) \subseteq \mathbb{R}_+\},$$

and

$$(Y_+^*)^\circ := \{y^* \in Y^* \ : \ y^*(Y_+ \backslash l(Y_+)) \subseteq \mathbb{R}_+ \backslash \{0\}\}.$$

Now, let us recall the following definitions.

Definition 1. A mapping $F : X \longrightarrow Y \cup \{+\infty_Y\}$ is said to be

- Y_+-convex if for any $\beta \in [0,1]$ and any $u_1, u_2 \in X$

$$F\left(\beta u_1+(1-\beta)u_2\right) \leq_{Y_+} \beta F\left(u_1\right)+(1-\beta)F\left(u_2\right).$$

- proper if its effective domain satisfies

$$\mathrm{dom}F := \{x \in X : F(x) \in Y\} \neq \emptyset.$$

- Y_+-epi-closed if its epigraph

$$\mathrm{epi}F := \{(x,y) \in X \times Y : F(x) \leq_{Y_+} y\} \text{ is closed.}$$

- star Y_+-lower semicontinuous if $y^* \circ F$ is lower semicontinuous for any $y^* \in Y_+^*$.

Let "$\leq_{Z_+}$" be a partial order on Z induced by a nonempty convex cone $Z_+ \subset Z$. We say that a mapping $g : Z \longrightarrow Y \cup \{+\infty_Y\}$ is (Z_+, Y_+)-nondecreasing if for any $w_1, w_2 \in Z$ we have

$$w_1 \leq_{Z_+} w_2 \Longrightarrow g(w_1) \leq_{Y_+} g(w_2).$$

If $h : X \longrightarrow Z \cup \{+\infty_Z\}$, then the composed mapping $g \circ h : X \longrightarrow Y \cup \{+\infty_Y\}$ is defined by

$$(g \circ h)(x) := \begin{cases} g(h(x)) & \text{if } x \in \text{ dom } h, \\ +\infty_Y & \text{otherwise.} \end{cases}$$

It is easy to see that if g is (Z_+, Y_+)-nondecreasing and Y_+-convex and h is Z_+-convex, then $g \circ h$ is Y_+-convex.

Given $F : X \supseteq S \longrightarrow Y \cup \{+\infty\}$ and $\epsilon \in Y_+$, we consider the following vector minimization problem:

$$(P) \quad \min_{x \in S} F(x).$$

We recall some ε-efficiency solutions of (P).

Definition 2. A point $\bar{x} \in S \cap \mathrm{dom}F$ is said to be

- a weakly ϵ-efficient solution if there does not exist $x \in S$ such that $F(x) <_{Y_+} F(\bar{x}) - \epsilon$,

- a properly ϵ-efficient solution if there exists a convex cone $\hat{Y}_+ \subsetneq Y$ that satisfies $Y_+ \backslash l(Y_+) \subseteq \operatorname{int} \hat{Y}_+$, there does not exist $x \in X, F(x) \lesssim_{\hat{Y}_+} F(\bar{x}) - \epsilon$.

The sets of weakly and properly ϵ-efficient points will be denoted by $E_\varepsilon^w(F, S, Y_+)$ and $E_\varepsilon^p(F, S, Y_+)$, respectively. Let us note that if $E_\varepsilon^\sigma(F, S, Y_+) \neq \emptyset$, then we can see easily $\varepsilon \in D^\sigma$, where

$$D^\sigma = \begin{cases} Y \backslash -\operatorname{int} Y_+ & \text{if } \sigma = w, \\ \\ Y \backslash (-Y_+ \backslash l(Y_+)) & \text{if } \sigma = p. \end{cases}$$

The ε-subdifferential of F at $\bar{x} \in \operatorname{dom} F$, may be defined regarding the different concepts of Pareto ε-solutions efficient in the above with respect to $\sigma \in \{w, p\}$ as follows:

$$\partial_\varepsilon^\sigma F(\bar{x}) := \{A \in L(X, Y) : \bar{x} \in E_\varepsilon^\sigma(F - A, S, Y_+)\},$$

that is,

$$\begin{aligned} \bullet \partial_\varepsilon^w F(\bar{x}) = \{&A \in L(X, Y) : \text{ there does not exist } x \in X, \\ &F(x) - F(\bar{x}) <_{Y_+} A(x - \bar{x}) - \varepsilon\}, \\ \bullet \partial_\varepsilon^p F(\bar{x}) = \{&A \in L(X, Y) : \text{ there exists } \hat{Y}_+ \subsetneq Y \text{ a convex cone that satisfies} \\ &Y_+ \backslash l(Y_+) \subseteq \operatorname{int} \hat{Y}_+, \\ &\text{there does not exist } x \in X, F(x) - F(\bar{x}) \lesssim_{\hat{Y}_+} A(x - \bar{x}) - \varepsilon\}. \end{aligned}$$

The conjugate function of any function $F : X \longrightarrow \mathbb{R} \cup \{+\infty\}$ is denoted by $F^* : X^* \to \overline{\mathbb{R}}$ and is defined as

$$F^*(x^*) := \sup_{x \in X} \{\langle x^*, x \rangle - F(x)\}.$$

For $\eta \geq 0$, the η-subdifferential of F at a point $\bar{x} \in \operatorname{dom} F$ is defined by

$$\partial_\eta F(\bar{x}) := \{x^* \in X^* : F(x) - F(\bar{x}) \geq \langle x^*, x - \bar{x} \rangle - \eta, \text{ for all } x \in X\}.$$

It is easy to check that

$$X^* \in \partial_\eta F(\bar{x}) \Longleftrightarrow F^*(x^*) + F(\bar{x}) - \langle x^*, \bar{x} \rangle \leq \eta.$$

By the definition of F^*, the so-called Young–Fenchel inequality is

$$F^*(x^*) + F(x) \geq \langle x^*, x\rangle, \quad \text{for all } (x, x^*) \in X \times X^*.$$

If $\varepsilon = 0_Y$, then the set $\partial_0^\sigma F(\bar{x}) := \partial^\sigma F(\bar{x})$ is the Pareto subdifferential (see [7]), for any $\sigma \in \{s, w, p\}$. For simplicity, we consider the following notation:

$$Y_+^\sigma := \begin{cases} Y_+^* \backslash \{0_Y\} \text{ if } \sigma = w, \\ \\ (Y_+^*)^\circ \qquad \text{if } \sigma = p, \end{cases}$$

For any subset $C \subset X$, the vector indicator mapping $\delta_C^v : X \longrightarrow Y \cup \{+\infty_Y\}$ of C is defined by

$$\delta_C^v(x) := \begin{cases} 0, \qquad \text{if } x \in C, \\ +\infty_Y, \text{ otherwise.} \end{cases}$$

When $Y = \mathbb{R}$, the scalar indicator function is denoted by δ_C. The vector indicator mapping δ_C^v appears to possess properties like the scalar indicator function δ_C. Moreover, we point out the relation between δ_C^v and δ_C

$$y^* \circ \delta_C^v = \delta_C, \text{ for all } y^* \in Y_+^* \backslash \{0_Y\}.$$

For any $\eta \geq 0$, the η-normal set to C at $\overline{x} \in C$ is defined as the η-subdifferential of the indicator function δ_C at $\overline{x} \in C$, that is,

$$N_\eta(\bar{x}, C) := \partial_\eta \delta_C(\overline{x}) = \{x^* \in X^* : \langle x^*, x - \bar{x}\rangle \leq \eta, \text{ for all } x \in C\}.$$

In what follows, we will need the following theorems. The first one characterizes scalarly the approximate σ-subdifferential for $\sigma \in \{w, p\}$ and the second one gives a general formula on the sequential approximate subdifferential of the sum of p ($p \geq 2$) proper lower semicontinuous convex functions defined in a reflexive Banach space.

Theorem 1 ([5])**.** Let $K : X \to Y \cup \{+\infty_Y\}$ and let $\bar{x} \in \text{dom}K$. Then, for $\sigma \in \{p, w\}$,

$$\partial_\varepsilon^\sigma K(\bar{x}) \supseteq \bigcup_{y^* \in Y_+^\sigma} \{A \in L(X, Y) : y^* \circ A \in \partial_{\langle y^*, \varepsilon\rangle}(y^* \circ K)(\bar{x})\}, \quad \text{for all } \varepsilon \in D^\sigma,$$

with equality if K is Y_+-convex and Y_+ pointed as $\sigma = p$.

Let $(X, \|\cdot\|_X)$ be a reflexive Banach space and let $(X^*, \|\cdot\|_{X^*})$ be its topological dual space. Let $(c_n^*)_{n\in\mathbb{N}}$ be a sequence in X^* (resp., $(c_n)_{n\in\mathbb{N}}$ be a sequence in X) and $c^* \in X^*$ (resp., $c \in X$). We write $c_n^* \overset{\|\cdot\|_{X^*}}{\underset{n\longmapsto+\infty}{\longrightarrow}} c^*$ (resp., $c_n \overset{\|\cdot\|_X}{\underset{n\longmapsto+\infty}{\longrightarrow}} c$) if $\|c_n^* - c^*\|_{X^*} \to 0$ when $n \to +\infty$ (resp., $\|c_n - c\|_X \to 0$ when $n \to +\infty$).

Theorem 2 ([14]). Let X be a reflexive Banach space, let $\eta \geq 0$, and let $k_1, \ldots, k_p : X \longrightarrow \mathbb{R} \cup \{+\infty\}$ be p proper, convex, and lower semicontinuous functions. For any $\bar{e} \in \bigcap_{i=1}^p \mathrm{dom} k_i$, we have $x^* \in \partial_\eta \left(\sum_{i=1}^p k_i\right)(\bar{e})$ if and only if there exist $\eta_i \geq 0$ and $(e_{i,n}, e_{i,n}^*) \in \mathrm{dom}\, k_i \times X^*$ satisfying

$$\begin{cases} \sum_{i=1}^p \eta_i = \eta, \quad e_{i,n} \overset{\|\cdot\|_X}{\underset{n\longmapsto+\infty}{\longrightarrow}} \bar{e}, \\ e_{i,n}^* \in \partial_{\eta_i} k_i\left(e_{i,n}\right), \\ \sum_{i=1}^p e_{i,n}^* \overset{\|\cdot\|_{X^*}}{\underset{n\longmapsto+\infty}{\longrightarrow}} x^*, \\ k_i\left(e_{i,n}\right) - \left\langle e_{i,n}^*, e_{i,n} - \bar{e}\right\rangle \underset{n\mapsto+\infty}{\longrightarrow} k_i(\bar{e}). \end{cases}$$

3 Sequential formula for approximate Pareto subdifferential of the convex vector mapping $f_1 + f_2 + \sum_{i=1}^p g_i \circ h_i$ $(p \geq 1)$

In the following, $(X, \|\cdot\|_X)$ and $(Z_1, \|\cdot\|_{Z_1}), \ldots, (Z_p, \|\cdot\|_{Z_p})$ are real reflexive Banach spaces, Y is a real Hausdorff topological vector space with continuous dual spaces $(X^*, \|\cdot\|_{X^*})$, $(Z_1^*, \|\cdot\|_{Z_1^*})$,..., $(Z_p^*, \|\cdot\|_{Z_p^*})$ and Y^*, duality pairing is also denoted by $\langle\cdot,\cdot\rangle$. We suppose two spaces Y and Z_i are partially ordered by nonempty convex cones Y_+ and Z_i^+, respectively. Moreover, we use the following norm on $X \times \prod_{k=1}^p Z_k$:

$$\|(x, z_1, \ldots, z_p)\|_{X\times Z_1\times\cdots\times Z_p} = \sqrt{\|x\|_X^2 + \sum_{i=1}^p \|z_i\|_{Z_i}^2}.$$

Similarly, we define the norm on $X^* \times \prod_{k=1}^p Z_k^*$. Our aim in this section, is to develop the sequential calculus rules for the approximate (weak and proper) Pareto subdifferential of the convex vector mapping $f_1 + f_2 + \sum_{i=1}^p g_i \circ h_i : X \longrightarrow Y \cup \{+\infty_Y\}$, where

- $f_1, f_2 : X \longrightarrow Y \cup \{+\infty_Y\}$ are proper, Y_+-convex and star Y_+-lower semicontinuous mappings,
- $h_i : X \longrightarrow Z_i \cup \{+\infty_{Z_i}\}$, $i = 1, \ldots, p$, are proper, Z_i^+-convex and Z_i^+-epi-closed mappings,
- $g_i : Z_i \longrightarrow Y \cup \{+\infty_Y\}$, $i = 1, \ldots, p$, are proper, Y_+-convex, (Z_i^+, Y_+)-nondecreasing and star Y_+-lower semicontinuous mappings,
- $(\bigcap_{i=1}^p h_i^{-1}(\mathrm{dom} g_i) \cap \mathrm{dom} h_i) \cap \mathrm{dom} f_1 \cap \mathrm{dom} f_2 \neq \emptyset$ and $g_i(+\infty_{Z_i}) = +\infty_Y$, $i = 1, \ldots, p$.

The approach that will be used is to reduce the (weak and proper) Pareto ε-subdifferential of $f_1 + f_2 + \sum_{i=1}^p g_i \circ h_i$ to that of the sums of a finite family of proper convex star Y_+-lower semicontinuous mappings. For this, let us consider the following auxiliary mappings:

$$\begin{aligned} F_1 : X \times \textstyle\prod_{k=1}^p Z_k &\longrightarrow Y \cup \{+\infty_Y\} \\ (x, z_1, \ldots, z_p) &\longrightarrow F_1(x, z_1, \ldots, z_p) := f_1(x), \end{aligned}$$

$$\begin{aligned} F_2 : X \times \textstyle\prod_{k=1}^p Z_k &\longrightarrow Y \cup \{+\infty_Y\} \\ (x, z_1, \ldots, z_p) &\longrightarrow F_2(x, z_1, \ldots, z_p) := f_2(x), \end{aligned}$$

$$\begin{aligned} G_i : X \times \textstyle\prod_{k=1}^p Z_k &\longrightarrow Y \cup \{+\infty_Y\} \qquad (i = 1, \ldots, p) \\ (x, z_1, \ldots, z_p) &\longrightarrow G_i(x, z_1, \ldots, z_p) := g_i(z_i), \end{aligned}$$

$$\begin{aligned} H_i : X \times \textstyle\prod_{k=1}^p Z_k &\longrightarrow Y \cup \{+\infty_Y\} \qquad (i = 1, \ldots, p) \\ (x, z_1, \ldots, z_p) &\longrightarrow H_i(x, z_1, \ldots, z_p) := \delta^v_{\mathrm{epi} h_i}(x, z_i). \end{aligned}$$

The following lemmas will be very helpful in what follows.

Lemma 1. Let $\bar{x} \in (\bigcap_{i=1}^p h_i^{-1}(\mathrm{dom} g_i) \cap \mathrm{dom} h_i) \cap \mathrm{dom} f_1 \cap \mathrm{dom} f_2$, let $\sigma \in \{p, w\}$, and let $\varepsilon \in D^\sigma$ with Y_+ be pointed as $\sigma = p$. Let $A \in L(X, Y)$ and let $T \in L(X \times \prod_{k=1}^p Z_k, Y)$ be defined by $T(x, z_1, \ldots, z_p) := A(x)$, for all $(x, z_1, \ldots, z_p) \in X \times \prod_{k=1}^p Z_k$. Then

$$A \in \partial^\sigma_\varepsilon (f_1 + f_2 + \sum_{i=1}^p g_i \circ h_i)(\bar{x})$$

if and only if

$$T \in \partial_\varepsilon^\sigma (F_1 + F_2 + \sum_{i=1}^{p} G_i + H_i)(\bar{x}, h_1(\bar{x}), \ldots, h_p(\bar{x})).$$

Proof. Let us prove the direct implication for the first case $\sigma = w$. Let $A \in \partial_\varepsilon^w (f_1+f_2+\sum_{i=1}^p g_i \circ h_i)(\bar{x})$. We proceed by contradiction, if $T \notin \partial_\varepsilon^w (F_1+F_2+\sum_{i=1}^p G_i+H_i)(\bar{x}, h_1(\bar{x}), \ldots, h_p(\bar{x}))$, then there exists $(x_0, z_{1,0}, \ldots, z_{p,0}) \in X \times \prod_{k=1}^p Z_k$ such that

$$\begin{aligned}&(F_1 + F_2 + \sum_{i=1}^{p} G_i + H_i)(x_0, z_{1,0}, \ldots, z_{p,0})\\ &-(F_1 + F_2 + \textstyle\sum_{i=1}^{p} G_i + H_i)(\bar{x}, h_1(\bar{x}), \ldots, h_p(\bar{x})) - A(x_0 - \bar{x}) + \varepsilon \in -\operatorname{int} Y_+,\end{aligned}$$

which implies that

$$x_0 \in \operatorname{dom} f_1 \cap \operatorname{dom} f_2, z_{i,0} \in \operatorname{dom} g_i, (x_0, z_{i,0}) \in \operatorname{epi} h_i, \qquad i = 1, \ldots, p, \quad (1)$$

and

$$(f_1(x_0)+f_2(x_0)+\sum_{i=1}^{p} g_i(z_{i,0}))-(f_1+f_2+\sum_{i=1}^{p} g_i \circ h_i)(\bar{x})-A(x_0-\bar{x})+\varepsilon \in -\text{ int } Y_+. \quad (2)$$

From relations (1), (2) and by using the monotonicity of g_i, we obtain

$$(f_1+f_2+\sum_{i=1}^{p} g_i \circ h_i)(x_0)-(f_1+f_2+\sum_{i=1}^{p} g_i \circ h_i)(\bar{x})-A(x_0-\bar{x})+\varepsilon \in -\text{ int } Y_+-Y_+,$$

and by using the fact that $-\text{ int } Y_+ - Y_+ \subseteq -\text{ int } Y_+$, we would obtain

$$(f_1+f_2+\sum_{i=1}^{p} g_i \circ h_i)(x_0)-(f_1+f_2+\sum_{i=1}^{p} g_i \circ h_i)(\bar{x})-A(x_0-\bar{x})+\varepsilon \in -\text{ int } Y_+,$$

which would contradict $A \in \partial_\varepsilon^w (f_1 + f_2 + \sum_{i=1}^p g_i \circ h_i)(\bar{x})$. To show the case $\sigma = p$, let us consider $A \in \partial_\varepsilon^p (f_1 + f_2 + \sum_{i=1}^p g_i \circ h_i)(\bar{x})$ and $T \notin \partial_\varepsilon^p (F_1+F_2+\sum_{i=1}^p G_i+H_i)(\bar{x}, h_1(\bar{x}), \ldots, h_p(\bar{x}))$. Hence, there exists a convex cone $\hat{Y}_+ \subsetneq Y$ such that $Y_+ \setminus \{0_Y\} \subseteq \operatorname{int} \hat{Y}_+$ and $(x_0, z_{i,0}) \in ((\operatorname{dom} f_1 \cap \operatorname{dom} f_2) \times \operatorname{dom} g_i) \cap \operatorname{epi} h_i$ $(i = 1, \ldots, p)$ satisfying

$$(F_1 + F_2 + \sum_{i=1}^{p} G_i + H_i)(x_0, z_{1,0}, \ldots, z_{p,0})$$

$$-(F_1+F_2+\sum_{i=1}^{p}G_i+H_i)(\bar{x},h_1(\bar{x}),\ldots,h_p(\bar{x}))-A(x_0-\bar{x})+\varepsilon\in-\hat{Y}_+\backslash l(\hat{Y}_+),$$

Reasoning in the same way as above for $\sigma=w$, we obtain

$$(f_1+f_2+\sum_{i=1}^{p}g_i\circ h_i)(x_0)-(f_1+f_2+\sum_{i=1}^{p}g_i\circ h_i)(\bar{x})-A(x_0-\bar{x})+\varepsilon\in-\hat{Y}_+\backslash l(\hat{Y}_+)-Y_+.$$

Observe that $-\hat{Y}_+\backslash l(\hat{Y}_+)-Y_+\subseteq-\hat{Y}_+\backslash l(\hat{Y}_+)$. Indeed, let $v:=v_1+v_2$ with $v_1\in-\hat{Y}_+\backslash l(\hat{Y}_+)$ and $v_2\in-Y_+$. If $v_2:=0_Y$, then $v\in-\hat{Y}_+\backslash l(\hat{Y}_+)$. Otherwise, $v_2\in-Y_+\backslash\{0_Y\}$ and since $-Y_+\backslash\{0_Y\}\subseteq-\operatorname{int}\hat{Y}_+\subseteq-\hat{Y}_+\backslash l(\hat{Y}_+)$, we get $v\in-\hat{Y}_+\backslash l(\hat{Y}_+)-\hat{Y}_+\backslash l(\hat{Y}_+)\subseteq-\hat{Y}_+\backslash l(\hat{Y}_+)$, and hence we have

$$(f_1+f_2+\sum_{i=1}^{p}g_i\circ h_i)(x_0)-(f_1+f_2+\sum_{i=1}^{p}g_i\circ h_i)(\bar{x})-A(x_0-\bar{x})+\varepsilon\in-\hat{Y}_+\backslash l(\hat{Y}_+),$$

obtaining again a contradiction with $A\in\partial_\varepsilon^p(f_1+f_2+\sum_{i=1}^{p}g_i\circ h_i)(\bar{x})$.

Conversely, it is obvious by contradiction, too. □

Lemma 2. Let $(\bar{x},\bar{z}_1,\ldots,\bar{z}_p)\in\operatorname{dom}F_1\cap\operatorname{dom}F_2\cap\operatorname{dom}G_i\cap\operatorname{dom}H_i$ and $\alpha_1,\alpha_2,\eta_i,\beta_i\geq0$ $(i=1,\ldots,p)$. Then for all $y^*\in Y_+^\sigma$, we have

1)
$$(x^*,z_1^*,\ldots,z_p^*)\in\partial_{\beta_i}(y^*\circ H_i)(\bar{x},\bar{z}_1,\ldots,\bar{z}_p)$$
$$\Longleftrightarrow\begin{cases}\text{there exist}\beta_{i,1}\geq0,\beta_{i,2}\geq0\text{ with }\beta_{i,1}+\beta_{i,2}=\beta_i,\\ x^*\in\partial_{\beta_{i,1}}(-z_i^*\circ h_i)(\bar{x}),\\ -z_i^*\in(Z_i^+)^*,\ \langle z_i^*,h_i(\bar{x})-\bar{z}_i\rangle\leq\beta_{i,2},\\ z_j^*=0,\ j\in\{1,\ldots,p\}\backslash\{i\}.\end{cases}$$

2) $(x^*,z_1^*,\ldots,z_p^*)\in\partial_{\alpha_1}(y^*\circ F_1)(\bar{x},\bar{z}_1,\ldots,\bar{z}_p)\Longleftrightarrow\begin{cases}x^*\in\partial_{\alpha_1}(y^*\circ f_1)(\bar{x}),\\ z_j^*=0,\ j\in\{1,\ldots,p\}.\end{cases}$

$$3)\ (x^*, z_1^*, \ldots, z_p^*) \in \partial_{\alpha_2}(y^* \circ F_2)(\bar{x}, \bar{z_1}, \ldots, \bar{z_p}) \Longleftrightarrow \begin{cases} x^* \in \partial_{\alpha_2}(y^* \circ f_2)(\bar{x}), \\ \\ z_j^* = 0,\ j \in \{1, \ldots, p\}. \end{cases}$$

$$4)\ (x^*, z_1^*, \ldots, z_p^*) \in \partial_{\eta_i}(y^* \circ G_i)(\bar{x}, \bar{z_1}, \ldots, \bar{z_p}) \Longleftrightarrow \begin{cases} z_i^* \in \partial_{\eta_i}(y^* \circ g_i)(\bar{z_i}), \\ \\ x^* = 0,\ z_j^* = 0, \\ j \in \{1, \ldots, p\} \backslash \{i\}. \end{cases}$$

Proof. 1) Let $(x, z_1, \ldots, z_p) \in \mathrm{dom} H_i$, let $\beta_i \geq 0$, $i = 1, \ldots, p$, and let $\left(x^*, z_1^*, \ldots, z_p^*\right) \in X^* \times \prod_{k=1}^p Z_k^*$. We can see that

$$H_i^*(x^*, z_1^*, \ldots, z_p^*) = \delta_{Z_i^*}^*(z_i^*) + (-z_i^* \circ h_i)^*(x^*) + \sum_{\substack{k=1 \\ k \neq i}}^{p} \delta_{\{0\}}(z_k^*).$$

Thus,

$$\left(x^*, z_1^*, \ldots, z_p^*\right) \in \partial_{\beta_i} H_i(x, z_1, \ldots, z_p)$$

if and only if

$$H_i^*\left(x^*, z_1^*, \ldots, z_p^*\right) + H_i(x, , z_1, \ldots, z_p) - \langle x^*, x \rangle - \sum_{\substack{k=1 \\ k \neq i}}^{p} \langle z_k^*, z_k \rangle - \langle z_i^*, z_i \rangle \leq \beta_i,$$

that is,

$$z_k^* = 0, \qquad k \in \{1, \ldots, p\} \backslash \{i\},$$

and

$$(-z_i^* \circ h_i)^*(x^*) + \delta_{Z_i^+}^*(z_i^*) + \delta_{\mathrm{epi} h_i}(\bar{x}, \bar{z_i}) - \langle x^*, \bar{x} \rangle - \langle z_i^*, \bar{z_i} \rangle \leq \beta_i, \quad (3)$$

by taking $w_i := \bar{z_i} - h_i(\bar{x}) \in Z_i^+$, we may rewrite (3) as follows:

$$[(-z_i^* \circ h_i)^*(x^*) + (-z_i^* \circ h_i)(\bar{x}) - \langle x^*, \bar{x} \rangle] + [\delta_{Z_i^+}^*(z_i^*) + \delta_{Z_i^+}(w_i) - \langle z_i^*, w_i \rangle] \leq \beta_i.$$

By using the Young–Fenchel inequality, it follows that

$$\begin{cases} (-z_i^* \circ h_i)^*(x^*) + (-z_i^* \circ h_i)(\bar{x}) - \langle x^*, \bar{x} \rangle \geq 0, \\ \\ \delta_{Z_i^+}^*(z_i^*) + \delta_{Z_i^+}(w_i) - \langle z_i^*, w_i \rangle \geq 0, \end{cases}$$

and hence, there exist some $\beta_{i,1} \geq 0$ and $\beta_{i,2} \geq 0$ satisfying $\beta_i = \beta_{i,1} + \beta_{i,2}$ and

$$\begin{cases} (-z_i^* \circ h_i)^*(x^*) + (-z_i^* \circ h_i)(\bar{x}) - \langle x^*, \bar{x} \rangle \leq \beta_{i,1}, \\ \\ \delta^*_{Z_i^+}(z_i^*) + \delta_{Z_i^+}(w_i) - \langle z_i^*, w_i \rangle \leq \beta_{i,2}, \end{cases}$$

that is,

$$\begin{cases} x^* \in \partial_{\beta_{i,1}}(-z_i^* \circ h_i)(\bar{x}), \\ \\ z_i^* \in \partial_{\beta_{i,2}} \delta_{Z_i^+}(w_i) = N_{\beta_{i,2}}(\bar{z}_i - h_i(\bar{x}), Z_i^+). \end{cases}$$

Since $Z_1^+, \ldots, Z_p^+$ are all convex cones, we get easily

$$z_i^* \in N_{\beta_{i,2}}(\bar{z}_i - h_i(\bar{x}), Z_i^+) \Longleftrightarrow \begin{cases} -z_i^* \in (Z_i^+)^*, \\ \\ \langle z_i^*, h_i(\bar{x}) - \bar{z}_i \rangle \leq \beta_{i,2}. \end{cases}$$

The proof of (2), (3), and (4) is similar to (1). □

Now, we present our main result in this section.

Theorem 3. Let $f_1, f_2 : X \longrightarrow Y \cup \{+\infty_Y\}$ be two proper, Y_+-convex and star Y_+-lower semicontinuous mappings, let $h_i : X \longrightarrow Z_i \cup \{+\infty_{Z_i}\}$, $i = 1, \ldots, p$, be p proper, Z_i^+-convex and Z_i^+-epi-closed mappings, and let $g_i : Z_i \longrightarrow Y \cup \{+\infty_Y\}$, $i = 1, \ldots, p$, be p proper, Y_+-convex, (Z_i^+, Y_+)-nondecreasing and star Y_+-lower semicontinuous mappings. Let $\bar{x} \in (\bigcap_{i=1}^p h_i^{-1}(\mathrm{dom} g_i) \cap \mathrm{dom} h_i) \cap \mathrm{dom} f_1 \cap \mathrm{dom} f_2$, $\varepsilon \in D^\sigma$ and $\sigma \in \{p, w\}$. Suppose that Y_+ is pointed as $\sigma = p$. Then, $A \in \partial_\varepsilon^\sigma (f_1 + f_2 + \sum_{i=1}^p g_i \circ h_i)(\bar{x})$ if and only if there exist $y^* \in Y_+^\sigma$, $x_n \in \mathrm{dom} f_1$, $r_n \in \mathrm{dom} f_2$, $(x_{i,n}, z_{i,n}) \in \mathrm{epi} h_i$, $v_{i,n} \in \mathrm{dom} g_i$, $x_n^*, r_n^*, x_{i,n}^* \in X^*$, $z_{i,n}^*, v_{i,n}^* \in Z_i^*$, and $\alpha_1, \alpha_2, \lambda_i, \gamma_i, \eta_i \geq 0$ satisfying

$$\begin{cases} \alpha_1 + \alpha_2 + \sum_{i=1}^p \lambda_i + \gamma_i + \eta_i = \langle y^*, \varepsilon \rangle, \\ \\ \begin{cases} x_n \overset{\|\cdot\|_X}{\underset{n \longmapsto +\infty}{\longrightarrow}} \bar{x}, \quad r_n \overset{\|\cdot\|_X}{\underset{n \longmapsto +\infty}{\longrightarrow}} \bar{x}, \quad x_{i,n} \overset{\|\cdot\|_X}{\underset{n \longmapsto +\infty}{\longrightarrow}} \bar{x}, \\ z_{i,n} \overset{\|\cdot\|_{Z_i}}{\underset{n \longmapsto +\infty}{\longrightarrow}} h_i(\bar{x}), \quad v_{i,n} \overset{\|\cdot\|_{Z_i}}{\underset{n \longmapsto +\infty}{\longrightarrow}} h_i(\bar{x}), \end{cases} \\ \\ x_n^* \in \partial_{\alpha_1}(y^* \circ f_1)(x_n), r_n^* \in \partial_{\alpha_2}(y^* \circ f_2)(x_n), v_{i,n}^* \in \partial_{\eta_i}(y^* \circ g_i)(v_{i,n}), \\ \\ x_{i,n}^* \in \partial_{\lambda_i}(-z_{i,n}^* \circ h_i)(x_{i,n}), -z_{i,n}^* \in (Z_i^+)^*, \langle z_{i,n}^*, h_i(x_{i,n}) - z_{i,n} \rangle \leq \gamma_i, \end{cases}$$

and

$$\begin{cases} x_n^* + r_n^* + \sum_{i=1}^{p} x_{i,n}^* \overset{\|\cdot\|_{x^*}}{\underset{n\mapsto+\infty}{\longrightarrow}} y^* \circ A \text{ and } z_{i,n}^* + v_{i,n}^* \overset{\|\cdot\|_{Z_i^*}}{\underset{n\mapsto+\infty}{\longrightarrow}} 0, \\ (y^* \circ f_1)(x_n) - \langle x_n^*, x_n - \bar{x}\rangle \underset{n\mapsto+\infty}{\longrightarrow} (y^* \circ f_1)(\bar{x}), \\ (y^* \circ f_2)(r_n) - \langle r_n^*, r_n - \bar{x}\rangle \underset{n\mapsto+\infty}{\longrightarrow} (y^* \circ f_2)(\bar{x}), \\ \langle x_{i,n}^*, x_{i,n} - \bar{x}\rangle + \langle z_{i,n}^*, z_{i,n} - h_i(\bar{x})\rangle \underset{n\mapsto+\infty}{\longrightarrow} 0, \\ (y^* \circ g_i)(v_{i,n}) - \langle v_{i,n}^*, v_{i,n} - h_i(\bar{x})\rangle \underset{n\mapsto+\infty}{\longrightarrow} (y^* \circ g_i)(h_i(\bar{x})). \end{cases}$$

Proof. Let $A \in \partial_\varepsilon^\sigma(f_1 + f_2 + \sum_{i=1}^p g_i \circ h_i)(\bar{x})$. By applying Lemma 1, we have

$$\begin{aligned} & A \in \partial_\varepsilon^\sigma(f_1 + f_2 + \sum_{i=1}^{p} g_i \circ h_i)(\bar{x}) \\ \Longleftrightarrow\ & T \in \partial_\varepsilon^\sigma(F_1 + F_2 + \sum_{i=1}^{p} G_i + H_i)(\bar{x}, h_1(\bar{x}), \ldots, h_p(\bar{x})). \end{aligned}$$

By virtue of scalarization Theorem 1, there exists $y^* \in Y_+^\sigma$ such that

$$y^* \circ T \in \partial_{\langle y^*,\varepsilon\rangle}(y^* \circ F_1 + y^* \circ F_2 + \sum_{i=1}^{p} y^* \circ G_i + y^* \circ H_i)(\bar{x}, h_1(\bar{x}), \ldots, h_p(\bar{x})),$$

which is equivalent to

$$(y^* \circ A, 0, 0, \ldots, 0) \in \partial_{\langle y^*,\varepsilon\rangle}(y^* \circ F_1 + y^* \circ F_2 + \sum_{i=1}^{p} y^* \circ G_i + y^* \circ H_i)(\bar{x}, h_1(\bar{x}), \ldots, h_p(\bar{x})).$$

The functions $y^* \circ G_i$, $y^* \circ H_i$ $(i = 1, \ldots, p,)$ $y^* \circ F_1$ and $y^* \circ F_2$ are all lower semicontinuous, convex, and proper on $X \times \prod_{k=1}^p Z_k$.

The condition $\bar{x} \in (\bigcap_{i=1}^p h_i^{-1}(\mathrm{dom} g_i) \cap \mathrm{dom} h_i) \cap \mathrm{dom} f_1 \cap \mathrm{dom} f_2$ is equivalent to $(\bar{x}, h_1(\bar{x}), \ldots, h_p(\bar{x})) \in (\bigcap_{i=1}^p \mathrm{dom} G_i \cap \mathrm{dom} H_i) \cap \mathrm{dom} F_1 \cap \mathrm{dom} F_2 = (\bigcap_{i=1}^p \mathrm{dom}(y^* \circ G_i) \cap \mathrm{dom}(y^* \circ H_i)) \cap \mathrm{dom}(y^* \circ F_1) \cap \mathrm{dom}(y^* \circ F_2)$. Hence, $y^* \circ G_i$, $y^* \circ H_i$, $i = 1, \ldots, p$, $y^* \circ F_1$ and $y^* \circ F_2$ satisfy all the hypotheses of Theorem 2, and therefore there exist $(x_n, w_{1,n}, \ldots, w_{p,n}) \in \mathrm{dom}(y^* \circ F_1)$,

$(r_n, t_{1,n}, \ldots, t_{p,n}) \in \mathrm{dom}(y^* \circ F_2)$, $(x_{i,n}, w_{1,i,n}, \ldots, w_{p,i,n}) \in \mathrm{dom}(y^* \circ H_i)$, $(c_{i,n}, u_{1,i,n}, \ldots, u_{p,i,n}) \in \mathrm{dom}(y^* \circ G_i)$, $(x_n^*, w_{1,n}^*, \ldots, w_{p,n}^*)$, $(r_n^*, t_{1,n}^*, \ldots, t_{p,n}^*)$ $(x_{i,n}^*, w_{1,i,n}^*, \ldots, w_{p,i,n}^*)$, $(c_{i,n}^*, u_{1,i,n}^*, \ldots, u_{p,i,n}^*) \in X^* \times \prod_{k=1}^{p} Z_k^*$ and α_1,α_2,β_i,$\eta_i \geq 0$, $i = 1, \ldots, p$, satisfying

$$\alpha_1 + \alpha_2 + \sum_{i=1}^{p} \beta_i + \eta_i = \langle y^*, \varepsilon \rangle, \tag{4}$$

$$\begin{cases} (x_n, w_{1,n}, \ldots, w_{p,n}) \overset{\|\cdot\|_{X \times Z_1 \times \cdots \times Z_p}}{\underset{n \longmapsto +\infty}{\longrightarrow}} (\bar{x}, h_1(\bar{x}), \ldots, h_p(\bar{x})), \\ (r_n, t_{1,n}, \ldots, t_{p,n}) \overset{\|\cdot\|_{X \times Z_1 \times \cdots \times Z_p}}{\underset{n \longmapsto +\infty}{\longrightarrow}} (\bar{x}, h_1(\bar{x}), \ldots, h_p(\bar{x})), \\ (c_{i,n}, u_{1,i,n}, \ldots, u_{p,i,n}) \overset{\|\cdot\|_{X \times Z_1 \times \cdots \times Z_p}}{\underset{n \longmapsto +\infty}{\longrightarrow}} (\bar{x}, h_1(\bar{x}), \ldots, h_p(\bar{x})), \\ (x_{i,n}, w_{1,i,n}, \ldots, w_{p,i,n}) \overset{\|\cdot\|_{X \times Z_1 \times \cdots \times Z_p}}{\underset{n \longmapsto +\infty}{\longrightarrow}} (\bar{x}, h_1(\bar{x}), \ldots, h_p(\bar{x})), \end{cases} \tag{5}$$

$$\begin{cases} (x_n^*, w_{1,n}^*, \ldots, w_{p,n}^*) \in \partial_{\alpha_1}(y^* \circ F_1)(x_n, w_{1,n}, \ldots, w_{p,n}), \\ (r_n^*, t_{1,n}^*, \ldots, t_{p,n}^*) \in \partial_{\alpha_2}(y^* \circ F_2)(r_n, t_{1,n}, \ldots, t_{p,n}), \\ (c_{i,n}^*, u_{1,i,n}^*, \ldots, u_{p,i,n}^*) \in \partial_{\eta_i}(y^* \circ G_i)(c_{i,n}, u_{1,i,n}, \ldots, u_{p,i,n}), \\ (x_{i,n}^*, w_{1,i,n}^*, \ldots, w_{p,i,n}^*) \in \partial_{\beta_i}(y^* \circ H_i)(x_{i,n}, w_{1,i,n}, \ldots, w_{p,i,n}), \end{cases} \tag{6}$$

$$\begin{cases} (y^* \circ F_1)(x_n, w_{1,n}, \ldots, w_{p,n}) - \langle (x_n^*, w_{1,n}^*, \ldots, w_{p,n}^*), (x_n, w_{1,n}, \ldots, w_{p,n}) \\ \qquad -(\bar{x}, h_1(\bar{x}), \ldots, h_p(\bar{x})) \rangle \underset{n \mapsto +\infty}{\longrightarrow} (y^* \circ F_1)(\bar{x}, h_1(\bar{x}), \ldots, h_p(\bar{x})), \\ (y^* \circ F_2)(r_n, t_{1,n}, \ldots, t_{p,n}) - \langle (r_n^*, t_{1,n}^*, \ldots, t_{p,n}^*), (r_n, t_{1,n}, \ldots, t_{p,n}) \\ \qquad -(\bar{x}, h_1(\bar{x}), \ldots, h_p(\bar{x})) \rangle \underset{n \mapsto +\infty}{\longrightarrow} (y^* \circ F_2)(\bar{x}, h_1(\bar{x}), \ldots, h_p(\bar{x})), \\ (y^* \circ G_i)(c_{i,n}, u_{1,i,n}, \ldots, u_{p,i,n}) - \langle (c_{i,n}^*, u_{1,i,n}^*, \ldots, u_{p,i,n}^*), (c_{i,n}, u_{1,i,n}, \ldots, u_{p,i,n}) \\ \qquad -(\bar{x}, h_1(\bar{x}), \ldots, h_p(\bar{x})) \rangle \underset{n \mapsto +\infty}{\longrightarrow} (y^* \circ G_i)(\bar{x}, h_1(\bar{x}), \ldots, h_p(\bar{x})), \\ (y^* \circ H_i)(x_{i,n}, w_{1,i,n}, \ldots, w_{p,i,n}) - \langle (x_{i,n}^*, w_{1,i,n}^*, \ldots, w_{p,i,n}^*), (x_{i,n}, w_{1,i,n}, \ldots, w_{p,i,n}) \\ \qquad -(\bar{x}, h_1(\bar{x}), \ldots, h_p(\bar{x})) \rangle \underset{n \mapsto +\infty}{\longrightarrow} (y^* \circ H_i)(\bar{x}, h_1(\bar{x}), \ldots, h_p(\bar{x})), \end{cases} \tag{7}$$

$$(x_n^* + r_n^* + \sum_{i=1}^{p} x_{i,n}^* + c_{i,n}^*, w_{1,n}^* + t_{1,n}^* + \sum_{i=1}^{p} w_{1,i,n}^* + u_{1,i,n}^*, \ldots, w_{p,n}^* + t_{p,n}^* + \sum_{i=1}^{p} w_{p,i,n}^* + u_{p,i,n}^*) \overset{\|\cdot\|_{x^* \times z_1^* \times \cdots \times z_p^*}}{\underset{n \longmapsto +\infty}{\longrightarrow}} (y^* \circ A, 0, \ldots, 0). \tag{8}$$

Note that (5) becomes equivalent to

$$\begin{cases} x_n \overset{\|\cdot\|_X}{\underset{n \longmapsto +\infty}{\longrightarrow}} \bar{x}, \quad r_n \overset{\|\cdot\|_X}{\underset{n \longmapsto +\infty}{\longrightarrow}} \bar{x}, \quad x_{i,n} \overset{\|\cdot\|_X}{\underset{n \longmapsto +\infty}{\longrightarrow}} \bar{x}, \quad c_{i,n} \overset{\|\cdot\|_X}{\underset{n \longmapsto +\infty}{\longrightarrow}} \bar{x}, \\ w_{j,n} \overset{\|\cdot\|_{Z_j}}{\underset{n \longmapsto +\infty}{\longrightarrow}} h_j(\bar{x}), \quad t_{j,n} \overset{\|\cdot\|_{Z_j}}{\underset{n \longmapsto +\infty}{\longrightarrow}} h_j(\bar{x}) \qquad (j = 1, \ldots, p), \\ w_{j,i,n} \overset{\|\cdot\|_{Z_j}}{\underset{n \longmapsto +\infty}{\longrightarrow}} h_j(\bar{x}), \ u_{j,i,n} \overset{\|\cdot\|_X}{\underset{n \longmapsto +\infty}{\longrightarrow}} h_j(\bar{x}) \qquad (j = 1, \ldots, p). \end{cases}$$

According to Lemma 2, (6) becomes equivalent to

$$\begin{cases} (x_n^*, w_{1,n}^*, \ldots, w_{p,n}^*) \in \partial_{\alpha_1}(y^* \circ F_1)(x_n, w_{1,n}, \ldots, w_{p,n}) \\ \qquad \Longleftrightarrow \begin{cases} x_n^* \in \partial_{\alpha_1}(y^* \circ f_1)(x_n), \\ w_{j,n}^* = 0, \ j \in \{1, \ldots, p\}. \end{cases} \\ (r_n^*, t_{1,n}^*, \ldots, t_{p,n}^*) \in \partial_{\alpha_2}(y^* \circ F_2)(r_n, t_{1,n}, \ldots, t_{p,n}) \\ \qquad \Longleftrightarrow \begin{cases} r_n^* \in \partial_{\alpha_2}(y^* \circ f_2)(r_n), \\ t_{j,n}^* = 0, \ j \in \{1, \ldots, p\}. \end{cases} \\ (c_{i,n}^*, u_{1,i,n}^*, \ldots, u_{p,i,n}^*) \in \partial_{\eta_i}(y^* \circ G_i)(c_{i,n}, u_{1,i,n}, \ldots, u_{p,i,n}) \\ \qquad \Longleftrightarrow \begin{cases} u_{i,i,n}^* \in \partial_{\eta_i}(y^* \circ g_i)(u_{i,i,n}), \\ c_{i,n}^* = 0, \ u_{j,i,n} = 0, \ j \in \{1, \ldots, p\} \backslash \{i\}. \end{cases} \\ (x_{i,n}^*, w_{1,i,n}^*, \ldots, w_{p,i,n}^*) \in \partial_{\beta_i}(y^* \circ H_i)(x_{i,n}, w_{1,i,n}, \ldots, w_{p,i,n}) \\ \qquad \Longleftrightarrow \begin{cases} \text{there exist} \beta_{i,1} \geq 0, \beta_{i,2} \geq 0 \text{ with } \beta_{i,1} + \beta_{i,2} = \beta_i, \\ x_{i,n}^* \in \partial_{\beta_{i,1}}(-w_{i,i,n}^* \circ h_i)(x_{i,n}), \\ -w_{i,i,n}^* \in (Z_i^+)^*, \langle w_{i,i,n}^*, h_i(x_{i,n}) - w_{i,i,n} \rangle \leq \beta_{i,2}, \\ w_{j,i,n}^* = 0, \ j \in \{1, \ldots, p\} \backslash \{i\}. \end{cases} \end{cases}$$

By putting $\lambda_i := \beta_{i,1}$, $\gamma_i := \beta_{i,2}$, $i = 1, \ldots, p$, and using (4), we get $\alpha_1 + \alpha_2 + \sum_{i=1}^{p} \lambda_i + \gamma_i + \eta_i = \langle y^*, \varepsilon \rangle$. As $(x_{i,n}, w_{i,i,n}) \in \mathrm{epi} h_i$ and $(\bar{x}, h_i(\bar{x})) \in \mathrm{epi} h_i$, then obviously the relation (7) is equivalent to

$$\begin{cases} (y^* \circ f_1)(x_n) - \langle x_n^*, x_n - \bar{x} \rangle \underset{n \mapsto +\infty}{\longrightarrow} (y^* \circ f_1)(\bar{x}), \\ (y^* \circ f_2)(x_n) - \langle x_n^*, x_n - \bar{x} \rangle \underset{n \mapsto +\infty}{\longrightarrow} (y^* \circ f_2)(\bar{x}), \\ (y^* \circ g_i)(u_{i,i,n}) - \langle u_{i,i,n}^*, u_{i,i,n} - h_i(\bar{x}) \rangle \underset{n \mapsto +\infty}{\longrightarrow} (y^* \circ g_i)(h_i(\bar{x})), \\ \langle x_{i,n}^*, x_{i,n} - \bar{x} \rangle + \langle w_{i,i,n}^*, w_{i,i,n} - h_i(\bar{x}) \rangle \underset{n \mapsto +\infty}{\longrightarrow} 0. \end{cases}$$

W can write (8) equivalently as

$$x_n^* + r_n^* + \sum_{i=1}^{p} x_{i,n}^* \overset{\|\cdot\|_{x^*}}{\underset{n \longmapsto +\infty}{\longrightarrow}} y^* \circ A \text{ and } z_{i,i,n}^* + v_{i,i,n}^* \overset{\|\cdot\|_{Z_i^*}}{\underset{n \longmapsto +\infty}{\longrightarrow}} 0.$$

Since $w_{j,n}$, $t_{j,n}$, $j = 1, \ldots, p$, $c_{i,n}$, $u_{j,i,n}$, $j \in \{1, \ldots, p\} \backslash \{i\}$, $i = 1, \ldots, p$, and $w_{j,i,n}$, $j \in \{1, \ldots, p\} \backslash \{i\}$, $i = 1, \ldots, p$, are superfluous, we put $z_{i,n} := w_{i,i,n}$, $v_{i,n} = u_{i,i,n}$, $z_{i,n}^* := w_{i,i,n}^*$ and $v_{i,n}^* := u_{i,i,n}^*$, $i = 1, \ldots, p$, which completes the proof. □

Remark 1. Let us note that if $\varepsilon = 0_Y$ in the above theorem, then $\alpha_1 = \alpha_2 = \lambda_i = \gamma_i = \eta_i = 0$ $(i = 1, \ldots, p)$.

4 Approximate optimality conditions for a constrained vector minimization problem

In this section, our main objective is to establish sequential approximate efficiency optimality conditions for the convex problem (P_1). In fact, by introducing the vector indicator mappings δ_C^v and $\delta_{-Z_i^+}^v$, $i = 1, \ldots, p$, the problem (P_1) becomes equivalent to the unconstrained vector minimization problem

$$(Q_1) \quad \min_{x \in X} \left(f + \delta_C^v + \sum_{i=1}^{p} \delta_{-Z_i^+}^v \circ h_i \right)(x).$$

Theorem 4. Let $f : X \longrightarrow Y \cup \{+\infty_Y\}$ be a proper, Y_+-convex and star Y_+-lower semicontinuous mapping, and let $h_i : X \longrightarrow Z_i \cup \{+\infty_{Z_i}\}$, $i = 1, \ldots, p$, be p proper, Z_i^+-convex and Z_i^+-epi-closed mappings. Let $\bar{x} \in \bigcap_{i=1}^p h_i^{-1}(-Z_i^+) \cap C$, $\varepsilon \in D^\sigma$ and let $\sigma \in \{w, p\}$. Suppose that Y_+ is pointed as $\sigma = p$ and that $C, Z_1^+, \ldots, Z_p^+$ are all closed convex cones. Then, $\bar{x}$ is a ε-σ-efficient solution of (P_1) if and only if there exist $y^* \in Y_+^\sigma$, $x_n \in \mathrm{dom} f$, $r_n \in C$, $(x_{i,n}, z_{i,n}) \in \mathrm{epi} h_i$, $v_{i,n} \in -Z_i^+$,

$x_n^*, r_n^*, x_{i,n}^* \in X^*$, $z_{i,n}^*, v_{i,n}^* \in Z_i^*$ and $\alpha_1, \alpha_2, \lambda_i, \gamma_i, \eta_i \geq 0$, $i = 1, \ldots, p$, satisfying

$$\begin{cases} \alpha_1 + \alpha_2 + \sum_{i=1}^p \lambda_i + \gamma_i + \eta_i = \langle y^*, \varepsilon \rangle, \\ \begin{cases} x_n \overset{\|\cdot\|_X}{\underset{n \longmapsto +\infty}{\longrightarrow}} \bar{x}, \quad r_n \overset{\|\cdot\|_X}{\underset{n \longmapsto +\infty}{\longrightarrow}} \bar{x}, \quad x_{i,n} \overset{\|\cdot\|_X}{\underset{n \longmapsto +\infty}{\longrightarrow}} \bar{x}, \\ z_{i,n} \overset{\|\cdot\|_{Z_i}}{\underset{n \longmapsto +\infty}{\longrightarrow}} h_i(\bar{x}), \quad v_{i,n} \overset{\|\cdot\|_{Z_i}}{\underset{n \longmapsto +\infty}{\longrightarrow}} h_i(\bar{x}), \end{cases} \\ x_n^* \in \partial_{\alpha_1}(y^* \circ f)(x_n), r_n^* \in N_{\alpha_2}(r_n, C), x_{i,n}^* \in \partial_{\lambda_i}(-z_{i,n}^* \circ h_i)(x_{i,n}), \\ v_{i,n}^* \in (Z_i^+)^* \text{ and } \langle v_{i,n}^*, -v_{i,n} \rangle \leq \eta_i, \\ -z_{i,n}^* \in (Z_i^+)^*, \langle z_{i,n}^*, h_i(x_{i,n}) - z_{i,n} \rangle \leq \gamma_i, \end{cases}$$

and

$$\begin{cases} x_n^* + \sum_{i=1}^p x_{i,n}^* \overset{\|\cdot\|_{x^*}}{\underset{n \longmapsto +\infty}{\longrightarrow}} 0 \text{ and } z_{i,n}^* + v_{i,n}^* \overset{\|\cdot\|_{Z_i^*}}{\underset{n \longmapsto +\infty}{\longrightarrow}} 0, \\ (y^* \circ f)(x_n) - \langle x_n^*, x_n - \bar{x} \rangle \underset{n \longmapsto +\infty}{\longrightarrow} (y^* \circ f)(\bar{x}), \\ \langle x_{i,n}^*, x_{i,n} - \bar{x} \rangle + \langle z_{i,n}^*, z_{i,n} - h_i(\bar{x}) \rangle \underset{n \longmapsto +\infty}{\longrightarrow} 0, \\ \langle r_n^*, r_n - \bar{x} \rangle \underset{n \longmapsto +\infty}{\longrightarrow} 0, \langle v_{i,n}^*, v_{i,n} - h_i(\bar{x}) \rangle \underset{n \longmapsto +\infty}{\longrightarrow} 0. \end{cases}$$

Proof. Since the problem (P_1) is equivalent to the unconstrained problem (Q_1), we have $\bar{x}$ is a ε-σ-efficient solution of (P_1) if and only if

$$0 \in \partial_\varepsilon^\sigma (f + \delta_C^v + \textstyle\sum_{i=1}^p \delta_{-Z_i^+}^v \circ h_i)(\bar{x}).$$

The vector indicator mappings δ^v_C and $\delta^v_{-Z_1^+}$ $(i=1,\dots,p)$ are Y_+-convex and star Y_+-lower semicontinuous. Let us recall that $\delta^v_{-Z_i^+}$ $(i=1,\dots,p)$ are (Z_i^+,Y_+)-nondecreasing (see [7]) and since the mappings $f_1=f$, $f_2=\delta^v_C$, $g_i=\delta^v_{-Z_i^+}$ and h_i satisfy together all the assumptions of Theorem 3, therefore there exist $y^*\in Y_+^\sigma$, $x_n\in \mathrm{dom} f$, $r_n\in C$, $(x_{i,n},z_{i,n})\in \mathrm{epi} h_i$, $v_{i,n}\in -Z_i^+$, $x_n^*,r_n^*,x_{i,n}^*\in X^*$, $z_{i,n}^*,v_{i,n}^*\in Z_i^*$, and $\alpha_1,\alpha_2,\lambda_i,\gamma_i,\eta_i\geq 0$, $i=1,\dots,p$, satisfying

$$
\begin{cases}
\alpha_1+\alpha_2+\displaystyle\sum_{i=1}^{p}\lambda_i+\gamma_i+\eta_i=\langle y^*,\varepsilon\rangle,\\
\begin{cases} x_n \underset{n\longmapsto+\infty}{\overset{\|\cdot\|_X}{\longrightarrow}} \bar{x}, \quad r_n \underset{n\longmapsto+\infty}{\overset{\|\cdot\|_X}{\longrightarrow}} \bar{x}, \quad x_{i,n} \underset{n\longmapsto+\infty}{\overset{\|\cdot\|_X}{\longrightarrow}} \bar{x},\\ z_{i,n} \underset{n\longmapsto+\infty}{\overset{\|\cdot\|_{Z_i}}{\longrightarrow}} h_i(\bar{x}), \quad v_{i,n} \underset{n\longmapsto+\infty}{\overset{\|\cdot\|_{Z_i}}{\longrightarrow}} h_i(\bar{x}),\end{cases}\\
x_n^*\in\partial_{\alpha_1}(y^*\circ f)(x_n), r_n^*\in N_{\alpha_2}(r_n,C), v_{i,n}^*\in N_{\eta_i}(v_{i,n},\\
-Z_i^+), x_{i,n}^*\in\partial_{\lambda_i}(-z_{i,n}^*\circ h_i)(x_{i,n})-z_{i,n}^*\in (Z_i^+)^*, \langle z_{i,n}^*,h_i(x_{i,n})-z_{i,n}\rangle\leq\gamma_i,
\end{cases}
$$

and

$$
\begin{cases}
x_n^*+\displaystyle\sum_{i=1}^{p}x_{i,n}^* \underset{n\longmapsto+\infty}{\overset{\|\cdot\|_{x^*}}{\longrightarrow}} 0 \text{ and } z_{i,n}^*+v_{i,n}^* \underset{n\longrightarrow+\infty}{\overset{\|\cdot\|_{Z_i^*}}{\longrightarrow}} 0,\\
(y^*\circ f)(x_n)-\langle x_n^*,x_n-\bar{x}\rangle \underset{n\mapsto+\infty}{\longrightarrow} (y^*\circ f)(\bar{x}),\\
\langle x_{i,n}^*,x_{i,n}-\bar{x}\rangle+\langle z_{i,n}^*,z_{i,n}-h_i(\bar{x})\rangle \underset{n\mapsto+\infty}{\longrightarrow} 0,\\
\langle r_n^*,r_n-\bar{x}\rangle \underset{n\mapsto+\infty}{\longrightarrow} 0, \langle v_{i,n}^*,v_{i,n}-h_i(\bar{x})\rangle \underset{n\mapsto+\infty}{\longrightarrow} 0.
\end{cases}
$$

It is easy to see that the condition $v_{i,n}^*\in N_{\eta_i}(v_{i,n},-Z_i^+)$ is equivalent to $v_{i,n}^*\in(Z_i^+)^*$ and $\langle v_{i,n}^*,-v_{i,n}\rangle\leq\eta_i$. Now the proof is complete. □

5 Applications to a constrained multiobjective fractional programming problems

In this section, by applying the previous results we present, without any constraint qualification, sequential weak approximate efficiency optimality conditions for the following multiobjective fractional programming problem

$$(P_2) \qquad \begin{cases} \min\left\{\dfrac{f_1(x)}{g_1(x)}, \ldots, \dfrac{f_s(x)}{g_s(x)}\right\} \\ x \in C, \\ h_1(x) \in -Z_1^+, \\ \vdots \\ h_p(x) \in -Z_p^+, \end{cases}$$

where f_j, $-g_j$: $X \longrightarrow \mathbb{R}$ $(j = 1, \ldots, s)$ are convex and lower semicontinuous functions and $h_i : X \longrightarrow Z_i \cup \{+\infty_{Z_i}\}$ $(i = 1, \ldots, p)$ are proper, Z_i^+-convex and Z_i^+-epi-closed mappings. Also, C and Z_i^+ $(i = 1, \ldots, p)$ are nonempty closed convex. Moreover, we suppose that $f_j(x) \geq 0$ and $g_j(x) > 0$ $(j = 1, \ldots, s)$ for any $x \in \bigcap_{i=1}^p h_i^{-1}(-Z_i^+) \cap C$. The finite-dimensional space $Y := \mathbb{R}^s$ is equipped with its natural order induced by the positive cone $Y := \mathbb{R}_+^s = \{(y_1, \ldots, y_s) \in \mathbb{R}^s,\ y_j \geq 0,\ \text{for all } j = 1, \ldots, s\}$. The following notations will be used in what follows:

$$\begin{aligned} &\varepsilon := (\varepsilon_1, \ldots, \varepsilon_s) \in \mathbb{R}_+^s, \\ &\bar{\varepsilon} := (\varepsilon_1 g_1(\bar{x}), \ldots, \varepsilon_s g_s(\bar{x})), \\ &\nu_j := \tfrac{f_j(\bar{x})}{g_j(\bar{x})} - \varepsilon_j \geq 0 \text{ and } \nu = (\nu_1, \ldots, \nu_s) \in \mathbb{R}_+^s. \end{aligned}$$

Now, we recall the definition of weakly ε-efficient solution of (P_2), which can be found in [10].

Definition 3. A point $\bar{x} \in \bigcap_{i=1}^p h_i^{-1}(-Z_i^+) \cap C$ is said to be weakly ε-efficient solution of (P_2) if there does not exist $x \in \bigcap_{i=1}^p h_i^{-1}(-Z_i^+) \cap C$ such that

$$\frac{f_j(x)}{g_j(x)} < \frac{f_j(\bar{x})}{g_j(\bar{x})} - \varepsilon_j \qquad (\text{ for all } j = 1, \ldots, s).$$

By using parametric approach of Dinkelbach [4], we can transform the problem (P_2) into a vector convex nonfractional programming problem de-

fined as follows:

$$(P_\nu) \qquad \begin{cases} \min F(x) \\ x \in C, \\ h_1(x) \in -Z_1^+, \\ \vdots \\ h_p(x) \in -Z_p^+, \end{cases}$$

where $F : X \longrightarrow \mathbb{R}^s$ is defined for any $x \in X$ by

$$F(x) := (f_1(x) - \nu_1 g_1(x), \ldots, f_s(x) - \nu_s g_s(x)).$$

Lemma 3. ([10]) A point $\bar{x} \in \bigcap_{i=1}^p h_i^{-1}(-Z_i^+) \cap C$ is a weakly ε-efficient solution of (P_2) if and only if $\bar{x}$ is a weakly $\bar{\varepsilon}$-efficient solution of (P_ν).

Theorem 5. Let $\bar{x} \in \bigcap_{i=1}^p h_i^{-1}(-Z_i^+) \cap C$, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_s) \in \mathbb{R}_+^s$ and $\nu_j = \frac{f_j(\bar{x})}{g_j(\bar{x})} - \varepsilon_j \geq 0$, $j = 1, \ldots, s$. Suppose that $C, Z_1^+, \ldots, Z_p^+$ are all closed convex cones. Then $\bar{x}$ is a weakly ε-efficient solution of (P_2) if and only if there exist $y^* = (y_1, \ldots, y_s) \in \mathbb{R}_+^s \setminus \{0\}$, $x_n \in X$, $r_n \in C$, $(x_{i,n}, z_{i,n}) \in \mathrm{epi} h_i$, $v_{i,n} \in -Z_i^+$,

$x_n^*, r_n^*, x_{i,n}^* \in X^*$, $z_{i,n}^*, v_{i,n}^* \in Z_i^*$ and $\alpha_1, \alpha_2, \lambda_i, \gamma_i, \eta_i \geq 0$, $i = 1, \ldots, p$, satisfying

$$\begin{cases} \alpha_1 + \alpha_2 + \sum_{i=1}^p \lambda_i + \gamma_i + \eta_i = \sum_{i=1}^s y_i \varepsilon_i, \\ \begin{cases} x_n \xrightarrow[n \longmapsto +\infty]{\|\cdot\|_X} \bar{x}, \quad r_n \xrightarrow[n \longmapsto +\infty]{\|\cdot\|_X} \bar{x}, \quad x_{i,n} \xrightarrow[n \longmapsto +\infty]{\|\cdot\|_X} \bar{x}, \\ z_{i,n} \xrightarrow[n \longmapsto +\infty]{\|\cdot\|_{Z_i}} h_i(\bar{x}), \quad v_{i,n} \xrightarrow[n \longmapsto +\infty]{\|\cdot\|_{Z_i}} h_i(\bar{x}), \end{cases} \\ x_n^* \in \partial_{\alpha_1} \left(\sum_{j=1}^s y_j \big(f_j + \nu_j(-g_j)\big) \right)(x_n), r_n^* \in N_{\alpha_2}(r_n, C), \\ x_{i,n}^* \in \partial_{\lambda_i}(-z_{i,n}^* \circ h_i)(x_{i,n}), v_{i,n}^* \in (Z_i^+)^* \text{ and } \langle v_{i,n}^*, -v_{i,n} \rangle \leq \eta_i, \\ -z_{i,n}^* \in (Z_i^+)^*, \langle z_{i,n}^*, h_i(x_{i,n}) - z_{i,n} \rangle \leq \gamma_i, \end{cases}$$

and

$$\begin{cases} x_n^* + r_n^* + \sum_{i=1}^{p} x_{i,n}^* \xrightarrow[n\longmapsto+\infty]{\|\cdot\|_{x^*}} 0 \text{ and } z_{i,n}^* + v_{i,n}^* \xrightarrow[n\longmapsto+\infty]{\|\cdot\|_{Z_i^*}} 0, \\ \Big(\sum_{j=1}^{s} y_j\big(f_j + \nu_j(-g_j)\big)\Big)(x_n) - \langle x_n^*, x_n - \bar{x}\rangle \\ \qquad\qquad \longrightarrow_{n\mapsto+\infty} \Big(\sum_{j=1}^{s} y_j\big(f_j + \nu_j(-g_j)\big)\Big)(\bar{x}), \\ \langle x_{i,n}^*, x_{i,n} - \bar{x}\rangle + \langle z_{i,n}^*, z_{i,n} - h_i(\bar{x})\rangle \longrightarrow_{n\mapsto+\infty} 0, \\ \langle r_n^*, r_n - \bar{x}\rangle \longrightarrow_{n\mapsto+\infty} 0, \langle v_{i,n}^*, v_{i,n} - h_i(\bar{x})\rangle \longrightarrow_{n\mapsto+\infty} 0. \end{cases}$$

Proof. According to Lemma 3, $\bar{x}$ is a weakly ε-efficient solution of (P_2) if and only if $\bar{x}$ is a weakly $\bar{\varepsilon}$-efficient solution of (P_v). The mappings $f := F$ and h_i satisfy together all the assumptions of Theorem 4, then it follows that there exist $y^* = (y_1, \ldots, y_s) \in (\mathbb{R}_+^s)^* \setminus \{0\} = \mathbb{R}_+^s \setminus \{0\}$, $x_n \in \mathrm{dom}F = X$, $r_n \in C$, $(x_{i,n}, z_{i,n}) \in \mathrm{epi}h_i$, $v_{i,n} \in -Z_i^+$,

$x_n^*, r_n^*, x_{i,n}^* \in X^*$, $z_{i,n}^*, v_{i,n}^* \in Z_i^*$ and $\alpha_1, \alpha_2, \lambda_i, \gamma_i, \eta_i \geq 0$, $i = 1, \ldots, p$, satisfying

$$\begin{cases} \alpha_1 + \alpha_2 + \sum_{i=1}^{p} \lambda_i + \gamma_i + \eta_i = \sum_{i=1}^{s} y_i\varepsilon_i, \\ \begin{cases} x_n \xrightarrow[n\longmapsto+\infty]{\|\cdot\|_X} \bar{x}, \quad r_n \xrightarrow[n\longmapsto+\infty]{\|\cdot\|_X} \bar{x}, \quad x_{i,n} \xrightarrow[n\longmapsto+\infty]{\|\cdot\|_X} \bar{x}, \\ z_{i,n} \xrightarrow[n\longmapsto+\infty]{\|\cdot\|_{Z_i}} h_i(\bar{x}), \quad v_{i,n} \xrightarrow[n\longmapsto+\infty]{\|\cdot\|_{Z_i}} h_i(\bar{x}), \end{cases} \\ x_n^* \in \partial_{\alpha_1}\Big(\sum_{j=1}^{s} y_j\big(f_j + \nu_j(-g_j)\big)\Big)(x_n), r_n^* \in N_{\alpha_2}(r_n, C), \\ x_{i,n}^* \in \partial_{\lambda_i}(-z_{i,n}^* \circ h_i)(x_{i,n}), v_{i,n}^* \in (Z_i^+)^* \text{ and } \langle v_{i,n}^*, -v_{i,n}\rangle \leq \eta_i, \\ -z_{i,n}^* \in (Z_i^+)^*, \langle z_{i,n}^*, h_i(x_{i,n}) - z_{i,n}\rangle \leq \gamma_i, \end{cases}$$

and

$$\begin{cases} x_n^* + r_n^* + \sum_{i=1}^{p} x_{i,n}^* \underset{n\longmapsto+\infty}{\overset{\|\cdot\|_{x^*}}{\longrightarrow}} 0 \text{ and } z_{i,n}^* + v_{i,n}^* \underset{n\longmapsto+\infty}{\overset{\|\cdot\|_{Z_i^*}}{\longrightarrow}} 0, \\ \Big(\sum_{j=1}^{s} y_j\big(f_j + \nu_j(-g_j)\big)\Big)(x_n) - \langle x_n^*, x_n - \bar{x}\rangle \\ \qquad\qquad \underset{n\longmapsto+\infty}{\longrightarrow} \Big(\sum_{j=1}^{s} y_j\big(f_j + \nu_j(-g_j)\big)\Big)(\bar{x}), \\ \langle x_{i,n}^*, x_{i,n} - \bar{x}\rangle + \langle z_{i,n}^*, z_{i,n} - h_i(\bar{x})\rangle \underset{n\longmapsto+\infty}{\longrightarrow} 0, \\ \langle r_n^*, r_n - \bar{x}\rangle \underset{n\longmapsto+\infty}{\longrightarrow} 0, \langle v_{i,n}^*, v_{i,n} - h_i(\bar{x})\rangle \underset{n\longmapsto+\infty}{\longrightarrow} 0. \end{cases}$$

□

In what follows, we present an important sub-class of such problems showing the applicability of the suggested conditions of Theorem 5. This is the class of multiobjective linear fractional programming problems, which have significant application in different real life areas such as production planning, financial sector, health care and all engineering fields. This type of problems is modeled as follows (see [12]):

$$(P_3) \qquad \begin{cases} \min\{K_1(x), \ldots, K_s(x)\} \\ x \in \mathbb{R}_+^n, \\ \mathcal{A}x - b \in -\mathbb{R}_+^p, \end{cases}$$

where $\mathcal{A}$ is a matrix $(p \times n)$, $b \in \mathbb{R}^p$, $K_j(x) = \frac{f_j(x)}{f_j(x)} = \frac{a_j^t x + b_j}{c_j^t x + d_j}$ $(a_j, c_j, x \in \mathbb{R}^n, b_j, d_j \in \mathbb{R}, j = 1, \ldots, s)$, and t denotes the transpose operation. Note that $\mathcal{S} = \{x \in \mathbb{R}_+^n, \mathcal{A}x - b \in -\mathbb{R}_+^p, b \in \mathbb{R}^p\}$ is the feasible set in decision space. We assume that for each feasible solution x, $f_j(x) \geq 0$ and $g_j(x) > 0$ $(j = 1, \ldots, s)$.

By taking $C := \mathbb{R}_+^n$ and $h(x) := \mathcal{A}x - b$, we observe that all assumptions of Theorem 5 are satisfied. Then we deduce the following result.

Theorem 6. Let $\bar{x} \in \mathcal{S}$, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_s) \in \mathbb{R}_+^s$ and let $\nu_j = \frac{a_j^t \bar{x} + b_j}{c_j^t \bar{x} + d_j} - \varepsilon_j \geq 0$, $j = 1, \ldots, s$. Then $\bar{x}$ is a weakly ε-efficient solution of (P_3) if and only if there

exist $y^* = (y_1, \ldots, y_s) \in \mathbb{R}^s_+ \setminus \{0\}$, $x_n \in \mathbb{R}^n$, $r_n \in \mathbb{R}^n_+$, $(w_n, z_n) \in \operatorname{epi}(\mathcal{A}(.)+b)$, $v_n \in -\mathbb{R}^p_+$,

$x_n^*, r_n^*, w_n^* \in \mathbb{R}^n$, $z_n^*, v_n^* \in \mathbb{R}^p$ and $\alpha_1, \alpha_2, \lambda, \gamma, \eta \geq 0$ satisfying

$$\begin{cases} \alpha_1 + \alpha_2 + \lambda + \gamma + \eta = \sum_{i=1}^{s} y_i \varepsilon_i, \\ \begin{cases} x_n \overset{\|\cdot\|_{\mathbb{R}^n}}{\underset{n\longmapsto+\infty}{\longrightarrow}} \bar{x}, \quad r_n \overset{\|\cdot\|_{\mathbb{R}^n}}{\underset{n\longmapsto+\infty}{\longrightarrow}} \bar{x}, \\ w_n \overset{\|\cdot\|_{\mathbb{R}^n}}{\underset{n\longmapsto+\infty}{\longrightarrow}} \bar{x}, \quad z_n \overset{\|\cdot\|_{\mathbb{R}^p}}{\underset{n\longmapsto+\infty}{\longrightarrow}} \mathcal{A}\bar{x} + b, \end{cases} \\ x_n^* \in \partial_{\alpha_1}\Big(\sum_{j=1}^{s} y_j\big(f_j + \nu_j(-g_j)\big)\Big)(x_n), r_n^* \in N_{\alpha_2}(r_n, \mathbb{R}^n_+), \\ w_n^* \in \partial_\lambda(-z_n^* \circ (\mathcal{A}(.) + b))(w_n), v_n^* \in \mathbb{R}^p_+ \text{ and } \langle v_n^*, -v_n\rangle \leq \eta, \\ -z_n^* \in \mathbb{R}^p_+, \langle z_n^*, \mathcal{A}w_n + b - z_n\rangle \leq \gamma, \end{cases}$$

and

$$\begin{cases} x_n^* + r_n^* + w_n^* \overset{\|\cdot\|_{\mathbb{R}^n}}{\underset{n\longmapsto+\infty}{\longrightarrow}} 0 \text{ and } z_n^* + v_n^* \overset{\|\cdot\|_{\mathbb{R}^p}}{\underset{n\longmapsto+\infty}{\longrightarrow}} 0, \\ \Big(\sum_{j=1}^{s} y_j\big(f_j + \nu_j(-g_j)\big)\Big)(x_n) - \langle x_n^*, x_n - \bar{x}\rangle \\ \qquad\qquad \underset{n\mapsto+\infty}{\longrightarrow} \Big(\sum_{j=1}^{s} y_j\big(f_j + \nu_j(-g_j)\big)\Big)(\bar{x}), \\ \langle w_n^*, w_n - \bar{x}\rangle + \langle z_n^*, z_n - \mathcal{A}\bar{x} + b\rangle \underset{n\mapsto+\infty}{\longrightarrow} 0, \\ \langle r_n^*, r_n - \bar{x}\rangle \underset{n\mapsto+\infty}{\longrightarrow} 0, \langle v_n^*, v_n - \mathcal{A}\bar{x} + b\rangle \underset{n\mapsto+\infty}{\longrightarrow} 0. \end{cases}$$

Next, we close this section by presenting an example showing that the Moreau–Rockafellar and Attouch–Brézis constraint qualifications (see [6]) fail and also the sequential ε-optimality conditions of Theorem 5 hold.

Example 1. Let us consider the following multiobjective fractional programming problem:

$$(Q) \qquad \begin{cases} \min f(x,y), \\ (x,y) \in C, \\ \sqrt{x^2+y^2} - y \leq 0, \\ (-x,-x) \leq_{\mathbb{R}^+_2} (0,0), \end{cases}$$

where $f : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$ is defined by

$$f(x,y) := \left(\frac{x+1}{2y}, \frac{-x}{2(y+1)} \right) \text{ and } C := \{0\} \times [1,2].$$

Let $f_1(x,y) := \frac{x+1}{2}$, $f_2(x,y) := \frac{-x}{2}$, $g_1(x,y) := y$, $g_2(x,y) := y+1$, $h_1(x,y) := \sqrt{x^2+y^2} - y$ and $h_2(x,y) := (-x,-x)$. Let $\varepsilon := (\varepsilon_1, \varepsilon_2) = (\frac{1}{2}, 0)$, $(\overline{x}, \overline{y}) := (0,1)$ be a feasible point, $\nu_1 = \frac{f_1(\overline{x},\overline{y})}{g_1(\overline{x},\overline{y})} - \varepsilon_1 = 0$, $\nu_2 = \frac{f_2(\overline{x},\overline{y})}{g_2(\overline{x},\overline{y})} - \varepsilon_2 = 0$ and $\overline{\varepsilon} = (\varepsilon_1 g_1(\overline{x},\overline{y}), \varepsilon_2 g_2(\overline{x},\overline{y})) = (\frac{1}{2}, 0)$. It is easy to check that $(\overline{x}, \overline{y})$ is a ε-weakly efficient solution of (Q). Indeed, putt $h := (h_1, h_2)$. Then we observe that $h(C) \cap (-\text{int}\mathbb{R}^3_+) = \emptyset$ and that $\mathbb{R}_+[\mathbb{R}^3_+ + h(C)] = \mathbb{R}^3_+$, which is not a subspace of $\mathbb{R}^3_+$. Hence, Moreau–Rockafellar and Attouch–Brézis qualification conditions are not satisfied.

For each $n \in \mathbb{N}$, we take $\alpha_1 = \frac{1}{2}$, $\alpha_2 = 0$, $\lambda_i = \gamma_i = \eta_i = 0$ $(i = 1,2)$, $x_n = (0,1)$, $x_{i,n} = (0,1)$ $(i = 1,2)$, $r_n = (0,1)$, $z_{1,n} = v_{1,n} = 0$, $z_{2,n} = v_{2,n} = (0,0)$. Let $x^*_n = (0,0)$, $x^*_{i,n} = (0,0)$ $(i = 1,2)$, $r^*_n = (0,0)$, $z^*_{1,n} = 0 \in -\mathbb{R}_+$, $v^*_{1,n} = 0 \in \mathbb{R}_+$, $z^*_{2,n} = (0,0) \in -\mathbb{R}^2_+$, $v^*_{2,n} = (0,0) \in \mathbb{R}^2_+$, and $(y_1, y_2) = (1,1) \in \mathbb{R}^2_+ \setminus \{(0,0)\}$. We have

$$\begin{cases} \alpha_1 + \alpha_2 + \sum_{i=1}^{2} \lambda_i + \gamma_i + \eta_i = \frac{1}{2} = \sum_{i=1}^{2} y_i \varepsilon_i g_i(0,1), \\ x_n \underset{n \mapsto +\infty}{\longrightarrow} (0,1), \quad r_n \underset{n \mapsto +\infty}{\longrightarrow} (0,1), \quad x_{i,n} \underset{n \mapsto +\infty}{\longrightarrow} (0,1), \\ z_{1,n} \underset{n \mapsto +\infty}{\longrightarrow} h_1(0,1) = 0, \quad v_{1,n} \underset{n \mapsto +\infty}{\longrightarrow} h_1(0,1) = 0, \\ z_{2,n} \underset{n \mapsto +\infty}{\longrightarrow} h_2(0,1) = (0,0), \quad v_{2,n} \underset{n \mapsto +\infty}{\longrightarrow} h_2(0,1) = (0,0). \end{cases}$$

It is immediate to see that $x^*_n \in \partial_{\alpha_1} \left(\sum_{i=1}^{2} y_i \big(f_i + \nu_i(-g_i)\big) \right)(x_n) = \{(0,0)\}$. Moreover, $r^*_n \in N_{\alpha_2}(r_n, C)$, $\langle v^*_{i,n}, -v_{i,n} \rangle = 0 \leq \eta_i = 0$ $(i = 1,2)$, $x^*_{i,n} \in \partial_{\lambda_i}(-z^*_{i,n} \circ h_i)(x_{i,n}) = \{(0,0)\}$ $(i = 1,2)$, $\langle z^*_{i,n}, h_i(x_{i,n}) - z_{i,n} \rangle = 0 \leq \gamma_i = 0$ $(i = 1,2)$ and

$$\begin{cases} x_n^* + r_n^* + \sum_{i=1}^{2} x_{i,n}^* \underset{n\mapsto+\infty}{\longrightarrow} (0,0) \text{ and } z_{i,n}^* + v_{i,n}^* \underset{n\mapsto+\infty}{\longrightarrow} (0,0), \\ \Big(\sum_{i=1}^{2} y_i\big(f_i + \nu_i(-g_i)\big)\Big)(x_n) - \langle x_n^*, x_n - (0,0)\rangle = \frac{1}{2} \\ \qquad\qquad \underset{n\mapsto+\infty}{\longrightarrow} \Big(\sum_{i=1}^{2} y_i\big(f_i + \nu_i(-g_i)\big)\Big)(0,1) = \frac{1}{2}, \\ \langle x_{i,n}^*, x_{i,n} - (0,1)\rangle + \langle z_{i,n}^*, z_{i,n} - h_i(0,1)\rangle = 0 \underset{n\mapsto+\infty}{\longrightarrow} 0, \\ \langle r_n^*, r_n - (0,1)\rangle \underset{n\mapsto+\infty}{\longrightarrow} 0, \langle v_{i,n}^*, v_{i,n} - h_i(0,1)\rangle = 0 \underset{n\mapsto+\infty}{\longrightarrow} 0. \end{cases}$$

Then, by Theorem 5 the point $(\overline{x}, \overline{y})$ is a weakly approximate solution for the problem (Q).

6 Conclusion

It is well known that the constraint qualifications are required to obtain approximate optimality conditions but sometimes these constraint qualifications become very difficult to compute. In this work, we focused to establish sequential approximate optimality conditions without any constraint qualification for a constrained convex vector minimization problem (P_1) via scalarization process in terms of the approximate subdifferentials of the associated functions. As an application, we derive sequential weakly approximate optimality conditions to a constrained multiobjective fractional programming problem (P_2).

Acknowledgments. The authors would like to express their sincere thanks to anonymous reviewers for their valuable comments and suggestions.

References

[1] Bai, F.S., Wu, Z.Y. and Zhu, D.L. *Sequential Lagrange multiplier condition for ε-optimal solution in convex programming*, Optimization. 57 (2008), 669–680.

[2] Bard, J.F. *Practical Bilevel Optimization: Algorithms and Applications*, Springer Science and Business Media. 30, 2013.

[3] Boţ, R.I., Csetnek, E.R. and Wanka, G. *Sequential optimality conditions for composed convex optimization problems*, J. Math. Anal. Appl. 342(2) (2008), 1015–1025.

[4] Dinkelbach, W. *On nonlinear fractional programming*, Manage. Sci. 13 (1967), 492–498.

[5] EL Maghri, M. *Pareto-Fenchel ε-subdifferential sum rule ε-efficiency*, Optim. Lett. 6(4) (2012), 763–781.

[6] El Maghri, M. *Pareto-Fenchel ε-subdifferential composition rule and ε-efficiency*, Numer. Funct. Anal. Optim. 35(1) (2014), 1–19.

[7] El Maghri, M. and Laghdir, M. *Pareto subdifferential calculus for convex vector mappings and applications to vector optimization*, SIAM J. Optim. 19 (2008), 1970–1994.

[8] Jeyakumar, V., Lee, G.M. and Dinh, N. *New sequential Lagrange multiplier conditions characterizing optimality without constraint qualifications for convex programs*, SIAM J. Optim. 14(2) (2003), 534–547.

[9] Jeyakumar, V., Lee, G.M. and Dinh, N. *Liberating the subgradient optimality conditions from constraint qualifications*, J. Glob. Optim. 36 (2006), 127–137.

[10] Kim, M.H., Kim, G.S. and Lee, G.M. *On ε-optimality conditions for multiobjective fractional optimization problems*, Fixed Point Theory Appl.(2011), 1–13.

[11] Kohli, B. *Sequential optimality conditions for multiobjective fractional programming problems*, Math Sci. 8(2) (2014), 128.

[12] Luhandjula, M.K. *Fuzzy approach for multiple objective linear fractional programming problems*, Fuzzy Sets Syst. 13(1) (1984), 11–23.

[13] Moustaid, M.B., Laghdir, M., Dali, I. and Rikouane, A. *Sequential optimality conditions for multi objective fractional programming problems via sequential subdifferential calculus*, Le Matematiche. 76(1) (2021), 79–96.

[14] Moustaid, M.B., Rikouane, A., Dali, I. and Laghdir, M. *Sequential approximate weak optimality conditions for multiobjective fractional programming problems via sequential calculus rules for the Brndsted-Rockafellar approximate subdifferential*, Rend. Circ. Mat. Palermo, II. Ser. (2022), 337–754.

[15] Stancu-Minasian, I.M. *Fractional programming : Theory, Methodsand Applica- tions*, Springer Science and Business Media. 409, 2012.

[16] Stancu-Minasian, I.M. *A ninth bibliography of fractional programming*, Optimization. 68(11) (2019) 2125–2169.

[17] Sun, G.Z. *An economic approach to some classical theorem in optimization theory*, Optim. Lett. 2(2) (2008), 281–286.

[18] Sun, X.K., Long, X.J. and Chai, Y. *Sequential optimality conditions for fractional optimization with applications to vector optimization*, J Optim. Theory. Appl. 164(2) (2015), 479–499.

[19] Thibault, L. *Sequential convex subdifferential calculus and sequential Lagrange multipliers*, SIAM J. Control. Optim. 35(4) (1997), 1434–1444.

[20] Zappone, A. and Jorswieck, E. *Energy efficiency in wireless networks via fractional programming theory*, Found. Trends Commun. Inf. Theory, 11 (2015), 185–396.

Iranian Journal of Numerical Analysis and Optimization
Vol. 15, No. 2, 2025, pp 704–727
https://doi.org/10.22067/ijnao.2025.91265.1571
https://ijnao.um.ac.ir/

Research Article

Numerical design of nonstationary wavelets: Enhanced filter design andapplications in image compression

A. Boussaad, Y. Fourar* and K. Melkemi

Abstract

In this study, we propose a novel method for computing both primal and dual filters for nonstationary biorthogonal wavelets, offering an advanced approach to wavelet filter design. The key challenge in image compression that this study addresses is the inefficiency of conventional stationary wavelets, which rely on fixed filter banks that do not adapt to local variations in an image. This limitation results in suboptimal compression

*Corresponding author

Received 22 December 2024; revised 20 February 2025; accepted 2 March 2025

Abdelmalik Boussaad
Department of Mathematics, Faculty of Mathematics And Computer Science, University of Batna 2, Algeria. e-mail: a.boussaad2001@gmail.com

Yasmine Fourar
Department of Mathematics, Faculty of Mathematics And Computer Science, University of Batna 2, Algeria. e-mail: ya.fourar@univ-batna2.dz

Khaled Melkemi
Department of Mathematics, Faculty of Mathematics And Computer Science, University of Batna 2, Algeria. e-mail: k.melkemi@univ-batna2.dz

How to cite this article
Boussaad, A. Fourar, Y. and Melkemi, K., Numerical design of nonstationary wavelets: Enhanced filter design andapplications in image compression. *Iran. J. Numer. Anal. Optim.*, 2025; 15(2): 704-727. https://doi.org/10.22067/ijnao.2025.91265.1571

performance, particularly for images with varying statistical properties and localized features. To address this, we use a nonstationary biorthogonal filter banks, which modify basis functions at different scaling levels, leading to improved frequency resolution, signal representation, and compression efficiency.

Our technique employs cardinal Chebyshev B-splines to derive explicit formulas for the primal filters, enabling precise calculation of filter coefficients essential for wavelet transforms. Additionally, we enforce normality and biorthogonality conditions within nonstationary multiresolution analysis to maintain the relationship between primal and dual wavelet filters at each scaling level. This structured approach allows for explicit formulation of the dual filters while ensuring accurate decomposition and reconstruction. Experimental results confirm that the proposed method improves compression efficiency over conventional Daubechies biorthogonal filters, increasing the number of zero coefficients in compressed images. This leads to better visual quality and reduced storage requirements while maintaining computational efficiency. Such improvements are particularly beneficial in applications requiring high-fidelity image reconstruction, such as medical imaging, satellite data processing, and video compression. MATLAB simulations validate the effectiveness of the approach, making it a promising alternative for image processing and data compression applications.

AMS subject classifications (2020): Primary 42C40; Secondary 65T60, 42C10.

Keywords: Cardinal Chebyshev B-splines; Biorthogonality; Wavelets; Image compression.

1 Introduction

Biorthogonal wavelets form a distinct class of wavelets that, unlike orthogonal wavelets [6, 4], employ separate functions for analysis and reconstruction. This fundamental distinction allows for greater flexibility in signal representation and improved frequency resolution, making them particularly useful in various signal processing applications. Key properties of biorthogonal wavelets include:

- **Symmetry**: They can be designed with varying degrees of symmetry, which is beneficial for specific applications.

- **Smoothness**: Greater smoothness corresponds to more vanishing moments, enabling better approximations of signals with smooth variations.

- **Compact Support**: Their scaling and wavelet functions are nonzero only over a finite interval, ensuring computational efficiency in fast algorithms like the discrete wavelet transform.

Nonstationary biorthogonal wavelets extend these advantages by adaptively adjusting their basis functions at each scaling level, based on the signal's properties. This adaptability enhances their ability to represent localized features or varying statistical properties, making them particularly effective in image processing, audio signal processing, and biomedical signal analysis. In contrast, stationary biorthogonal wavelets rely on a fixed set of basis functions, which may not be optimal for signals with complex characteristics.

A key advancement in this field is the use of *cardinal Chebyshev B-splines* [8, 7, 2] as a basis for constructing nonstationary biorthogonal wavelets. These splines offer excellent time-frequency localization properties, making them highly effective for analyzing nonstationary signals. Previous studies have explored their applications in wavelet construction. For example, Lee and Yoon [5] introduced a lifting scheme-based algorithm that enhances wavelet approximation properties and computational efficiency. Similarly, Vonesch, Blu, and Unser [10] developed a family of generalized Daubechies wavelets [3] incorporating Chebyshev B-splines, and Boxing. Zhang et al. [11] proposed a construction method based on exponential pseudo-splines.

While significant progress has been made, current methods for designing nonstationary biorthogonal wavelets still face several challenges, particularly in computational efficiency and systematic filter design. Many approaches rely on iterative or heuristic techniques, which can be computationally intensive and lack a structured framework. Additionally, solving the underlying systems of equations during wavelet construction remains a computational bottleneck, limiting practical applications in real-time or resource-constrained environments.

To address these challenges, this paper introduces a novel approach that systematically derives a linear system based on the normality and biorthogonality conditions of nonstationary multiresolution analysis. Our key contributions include:

- **Novel Design Method**: We propose a new method for constructing nonstationary biorthogonal wavelets using fundamental conditions from multiresolution analysis [9].
- **Efficient Numerical Solution**: By formulating a linear system, we enable the use of direct numerical methods such as Gauss–Jordan elimination and LU factorization, improving computational efficiency.
- **Explicit Filter Formulation**: Our approach provides explicit formulas for filters and dual filters at all scaling levels, ensuring a systematic and practical wavelet design process.
- **Comprehensive Framework**: We integrate theoretical insights with practical applications, making our method accessible to both researchers and practitioners in signal processing.

The remainder of this paper is structured as follows: Section 2 provides an overview of nonstationary biorthogonal multiresolution analysis, including key definitions and properties. Section 3 explores the use of cardinal Chebyshev B-splines in constructing nonstationary biorthogonal wavelets. By the conclusion, readers will gain a comprehensive understanding of the theory, practical applications, and our novel design method for nonstationary biorthogonal wavelets.

2 Definitions and properties of biorthogonal nonstationary wavelets

Let $(\mathcal{V}_j)_{j\leq j_0}$ and $(\widetilde{\mathcal{V}}_j)_{j\leq j_0}$ be a pair of **NSRMAs** [1] generated by a pair of dual scaling functions φ_j and $\widetilde{\varphi}_j$, $j \leq j_0$, such that

$$\langle \varphi_j, \widetilde{\varphi}_j(. - k)\rangle = \delta_{0,k}, \qquad k \in \mathbb{Z}. \tag{1}$$

The concept of biorthogonal wavelets is to find complementary spaces $\mathcal{W}_j$ and $\widetilde{\mathcal{W}}_j$ satisfying

$$\mathcal{V}_{j+1} = \mathcal{V}_j + \mathcal{W}_j, \text{ and } \widetilde{\mathcal{V}}_{j+1} = \widetilde{\mathcal{V}}_j + \widetilde{\mathcal{W}}_j, \tag{2}$$

$$\mathcal{W}_j \perp \widetilde{\mathcal{V}}_j, \text{ and } \widetilde{\mathcal{W}}_j \perp \mathcal{V}_j. \tag{3}$$

The nesting property for the subspaces $\mathcal{V}_j$ and $\widetilde{\mathcal{V}}_j$, implies the existence of filters $h_j = (h_j[k])_{k\in\mathbb{Z}} \in \ell^2(\mathbb{Z})$, $\widetilde{h}_j = (\widetilde{h}_j[k])_{k\in\mathbb{Z}} \in \ell^2(\mathbb{Z})$, $j \le j_0$, which satisfy

$$\varphi_j(t) = \sqrt{2}\sum_{k\in\mathbb{Z}} h_j[k]\varphi_{j+1}(2t-k), \quad \widetilde{\varphi}_j(t) = \sqrt{2}\sum_{k\in\mathbb{Z}} \widetilde{h}_j[k]\widetilde{\varphi}_{j+1}(2t-k). \tag{4}$$

By the Fourier transform, we have

$$\widehat{\varphi}_j(\omega) = \mathcal{H}_j(e^{i\frac{\omega}{2}})\widehat{\varphi}_{j+1}\left(\frac{\omega}{2}\right), \tag{5}$$

$$\widehat{\widetilde{\varphi}}_j(\omega) = \widetilde{\mathcal{H}}_j(e^{i\frac{\omega}{2}})\widehat{\widetilde{\varphi}}_{j+1}\left(\frac{\omega}{2}\right), \tag{6}$$

with

$$\mathcal{H}_j(e^{i\omega}) = \frac{1}{\sqrt{2}}\sum_{k\in\mathbb{Z}} h_j[k]e^{-ik\omega}, \tag{7}$$

$$\widetilde{\mathcal{H}}_j(e^{i\omega}) = \frac{1}{\sqrt{2}}\sum_{k\in\mathbb{Z}} \widetilde{h}_j[k]e^{-ik\omega}. \tag{8}$$

We assume that $z = e^{-i\omega}$.

Proposition 1. The filters $\mathcal{H}_j$ and $\widetilde{\mathcal{H}}_j$ associated for φ_j and $\widetilde{\varphi}_j$ satisfy the scaling condition:

$$\mathcal{H}_j(z)\widetilde{\mathcal{H}}_j(z^{-1}) + \mathcal{H}_j(-z)\widetilde{\mathcal{H}}_j(-z^{-1}) = 1. \tag{9}$$

For all $k \in \mathbb{Z}$, the preceding condition is equivalent to

$$\sum_{\ell\in\mathbb{Z}} \widetilde{h}_j[\ell]h_j[\ell-2k] = \delta_{k,0}. \tag{10}$$

Definition 1. The biorthogonal wavelet functions are given as follows:

$$\psi_j(t) = \sqrt{2}\sum_{k\in\mathbb{Z}} g_j[k]\varphi_{j+1}(2t-k), \quad \widetilde{\psi}_j(t) = \sqrt{2}\sum_{k\in\mathbb{Z}} \widetilde{g}_j[k]\widetilde{\varphi}_{j+1}(2t-k), \tag{11}$$

such that

$$\widetilde{g}_j[k] = (-1)^k h_j[1-k], \qquad g_j[k] = (-1)^k \widetilde{h}_j[1-k], \tag{12}$$

where

$$g_j = (g_j[k])_{k\in\mathbb{Z}}, \qquad \widetilde{g}_j = (\widetilde{g}_j[k])_{k\in\mathbb{Z}}, \tag{13}$$

are the scaling filters associated for ψ_j and $\widetilde{\psi}_j$, respectively.

For all $k \in \mathbb{Z}$, we have

$$\sum_{k\in\mathbb{Z}} h_j[k] = \sqrt{2}, \quad \sum_{k\in\mathbb{Z}} \widetilde{h}_j[k] = \sqrt{2}, \tag{14}$$

$$\sum_{k\in\mathbb{Z}} g_j[k] = 0, \quad \sum_{k\in\mathbb{Z}} \widetilde{g}_j[k] = 0. \tag{15}$$

As in the stationary case for a function $f \in L^2(\mathbb{R})$, a wavelet decomposition at level j_0 can be introduced, as follows:

$$f = \sum_{k\in\mathbb{Z}} a_{j_0}[k]\varphi_{j_0,k} + \sum_{j\le j_0}\sum_{k\in\mathbb{Z}} d_j[k]\psi_{j,k}, \tag{16}$$

where

$$a_{j_0}[k] = \langle f, \widetilde{\varphi}_{j_0,k}\rangle, \quad d_j[k] = \langle f, \widetilde{\psi}_{j,k}\rangle, \tag{17}$$

and

$$\widetilde{\varphi}_{j,k}(t) = 2^{j/2}\widetilde{\varphi}_j(2^j t - k), \qquad \widetilde{\psi}_{j,k}(t) = 2^{j/2}\widetilde{\psi}_j(2^j t - k). \tag{18}$$

3 Construction of scaling filters and dual scaling filters

The function we choose as a scaling function is the cardinal Chebyshev B-splines. Let us first recall the definition of the exponential Chebyshev B-splines mentioned in [1]. Let $\lambda = (\lambda_1, \ldots, \lambda_n)$ be a complex vector.

Definition 2. Suppose that $\alpha \in \mathbb{R}$. The function N_α^1 is defined by

$$N_\alpha^1(t) = e^{\alpha t}\chi_{[0,1]}(t), \qquad t \in \mathbb{R}. \tag{19}$$

Function N_λ^n is defined by convolution as follows:

$$N_{\lambda}^{n} = N_{\lambda_1}^{1} * \cdots * N_{\lambda_n}^{1}. \tag{20}$$

The function $N_{2^{-j}\lambda}^{n}(t)$ can generate a Riesz basis only if the vector λ does not have any distinct purely imaginary components λ_i and λ_j, where the difference between them equals $i2\pi k$ for some integer k.

For $n = 1$,

$$N_{\alpha}^{1}(t) = N_{\frac{\alpha}{2}}^{1}(2t) + e^{\frac{\alpha}{2}} N_{\frac{\alpha}{2}}^{1}(2t-1).$$

By the Fourier transform, we have

$$\widehat{N_{\alpha}^{1}}(\omega) = \left(\frac{1 + e^{\frac{\alpha - i\omega}{2}}}{2}\right) \widehat{N_{\frac{\alpha}{2}}^{1}}\left(\frac{\omega}{2}\right).$$

In the following, we assume that $\alpha \in [0, 10] \setminus \{2\pi\}$ with $2\pi \approx 6.2831$. This restriction ensures that the denominator remains nonzero for all the filters of the primal and dual functions.

3.1 Second order cardinal Chebyshev B-splines

Let $\lambda = (\lambda_1, \lambda_2) = (-i\alpha, i\alpha)$. We assume that

$$\varphi_j(t) = N_{2^{-j}\lambda}^{2}(t). \tag{21}$$

Then

$$\begin{aligned}
&\widehat{\varphi_j}(\omega) \\
&= \widehat{N_{2^{-j}\lambda_1}^{1}}(\omega) \widehat{N_{2^{-j}\lambda_2}^{1}}(\omega) \\
&= \frac{1}{4}\left(1 + (e^{2^{-(j+1)}\lambda_1} + e^{2^{-(j+1)}\lambda_2}) e^{\frac{-i\omega}{2}} + e^{2^{-(j+1)}(\lambda_1+\lambda_2)} e^{-i\omega}\right) \widehat{N_{2^{-(j+1)}\lambda}^{2}}\left(\frac{\omega}{2}\right) \\
&= \mathcal{H}_j\left(\frac{\omega}{2}\right) \widehat{\varphi_{j+1}}\left(\frac{\omega}{2}\right).
\end{aligned}$$

In accordance with (5), we obtain

$$h_j[0] = h_j[2] = \frac{\sqrt{2}}{4}, \qquad h_j[1] = \frac{2\sqrt{2}\cos(2^{-(j+1)}\alpha)}{4}.$$

However, these coefficients are not normalized. To normalize them, we simply divide by $\mathcal{H}_j(0)$, yielding

$$h_j[0] = h_j[2] = \frac{\sqrt{2}}{2 + 2\cos(2^{-(j+1)}\alpha)}, \qquad h_j[1] = \frac{2\sqrt{2}\cos(2^{-(j+1)}\alpha)}{2 + 2\cos(2^{-(j+1)}\alpha)}.$$

We will now introduce a method to explicitly determine the filters of dual scaling functions. Let the filters coefficients of the dual scaling functions $(\widetilde{\varphi}_j)_{j\le j_0}$ be $\widetilde{h}_j[-1]$, $\widetilde{h}_j[0]$, $\widetilde{h}_j[1]$, $\widetilde{h}_j[2]$, $\widetilde{h}_j[3]$. First, we assume that the dual scaling functions are symmetric, which means

$$\widetilde{h}_j[-1] = \widetilde{h}_j[3], \qquad \widetilde{h}_j[0] = \widetilde{h}_j[2].$$

The relation (14) implies

$$2\widetilde{h}_j[3] + 2\widetilde{h}_j[2] + \widetilde{h}_j[1] = \sqrt{2}. \tag{22}$$

Using (10) for $k = 0$, we obtain

$$\cos\left(2^{-(j+1)}\alpha\right)\widetilde{h}_j[1] + \widetilde{h}_j[2] = \frac{2 + 2\cos(2^{-(j+1)}\alpha)}{2\sqrt{2}}, \tag{23}$$

and for $k = 1$, we have

$$2\cos\left(2^{-(j+1)}\alpha\right)\widetilde{h}_j[3] + \widetilde{h}[2] = 0. \tag{24}$$

As an additional condition to simplify the calculation, we use (15) to get

$$\sum_{k=-2}^{2} (-1)^k \widetilde{h}_j[1-k] = 0,$$

which leads to

$$2\widetilde{h}_j[3] - 2\widetilde{h}_j[2] + \widetilde{h}_j[1] = 0. \tag{25}$$

To determine $\widetilde{h}_j[1]$, $\widetilde{h}_j[2]$, and $\widetilde{h}_j[3]$, the above system is solved using the Gaussian elimination.

Therefore,

$$\begin{cases} \widetilde{h}_j[-1] = \widetilde{h}_j[3] = -\dfrac{\sqrt{2}}{8\cos(2^{-(j+1)}\alpha)}, \\ \widetilde{h}_j[0] = \widetilde{h}_j[2] = \dfrac{\sqrt{2}}{4}, \\ \widetilde{h}_j[1] = \dfrac{2\sqrt{2}\cos(2^{-(j+1)}\alpha) + \sqrt{2}}{4\cos(2^{-(j+1)}\alpha)}. \end{cases}$$

3.2 Third order cardinal Chebyshev B-splines

Let $\lambda = (\lambda_1, \lambda_2, \lambda_3) = (0, -i\alpha, i\alpha)$, let $\widetilde{\lambda} = (\lambda_2, \lambda_3)$, and let

$$N^3_{\lambda}(x) = N^1_{\lambda_1} * N^2_{\widetilde{\lambda}}(x).$$

We assume that

$$\varphi_j(t) = N^3_{2^{-j}\lambda}(t). \tag{26}$$

Then

$$\begin{aligned} \widehat{\varphi_j}(\omega) &= \widehat{N^2_{2^{-j}\widetilde{\lambda}}}(\omega)\widehat{N^1_{2^{-j}\lambda_1}}(\omega), \\ &= \left(\frac{1}{8} + \frac{1+2\cos(2^{-(j+1)}\alpha)}{8}e^{-i\frac{\omega}{2}} + \frac{1+2\cos(2^{-(j+1)}\alpha)}{8}e^{-i\omega} + \frac{1}{8}e^{-i\frac{3\omega}{2}}\right) \\ &\quad \times \widehat{N^3_{2^{-(j+1)}\gamma}}\left(\frac{\omega}{2}\right) \\ &= \mathcal{H}_j\left(\frac{\omega}{2}\right)\widehat{\varphi_{j+1}}\left(\frac{\omega}{2}\right). \end{aligned}$$

The normalized coefficients are derived using the same method as described earlier:

$$\begin{aligned} h_j[0] = h_j[3] &= \frac{\sqrt{2}}{4 + 4\cos(2^{-(j+1)}\alpha)}, \\ h_j[1] = h_j[2] &= \frac{\sqrt{2}(2\cos(2^{-(j+1)}\alpha) + 1)}{4 + 4\cos(2^{-(j+1)}\alpha)}. \end{aligned}$$

Using the same steps as for $n = 2$, we assume that $\widetilde{h}_j[-1], \widetilde{h}_j[0], \widetilde{h}_j[1], \widetilde{h}_j[2]$, $\widetilde{h}_j[3]$, $\widetilde{h}_j[4]$, are the filters coefficients of the dual scaling functions$(\widetilde{\varphi}_j)_{j\leq j_0}$, subject to the condition that they are symmetric:

$$\widetilde{h}_j[-1] = \widetilde{h}_j[4], \qquad \widetilde{h}_j[0] = \widetilde{h}_j[3], \qquad \widetilde{h}_j[1] = \widetilde{h}_j[2].$$

The relation (14) implies

$$\widetilde{h}_j[2] + \widetilde{h}_j[3] + \widetilde{h}_j[4] = \frac{\sqrt{2}}{2}. \tag{27}$$

Using (10), we can find the following equations:

For $k = 0$,

$$(2\cos(2^{-(j+1)}\alpha) + 1)\widetilde{h}_j[2] + \widetilde{h}_j[3] = \frac{4 + 4\cos(2^{-(j+1)}\alpha)}{2\sqrt{2}}, \tag{28}$$

and for $k = 1$,

$$\widetilde{h}_j[2] + \left(2\cos(2^{-(j+1)}\alpha) + 1\right)\widetilde{h}_j[3] + \left(2\cos(2^{-(j+1)}\alpha) + 1\right)\widetilde{h}_j[4] = 0. \tag{29}$$

We can use Gaussian elimination to solve the above linear system and obtain $\widetilde{h}_j[2]$, $\widetilde{h}_j[3]$, and $\widetilde{h}_j[4]$, with the following explicit formulas:

$$\begin{cases} \widetilde{h}_j[-1] = \widetilde{h}_j[4] = 0, \\ \widetilde{h}_j[0] = \widetilde{h}_j[3] = \dfrac{-\sqrt{2}}{4\cos(2^{-(j+1)}\alpha)}, \\ \widetilde{h}_j[1] = \widetilde{h}_j[2] = \dfrac{2\sqrt{2}\cos(2^{-(j+1)}\alpha) + \sqrt{2}}{4\cos(2^{-(j+1)}\alpha)}. \end{cases}$$

3.3 Fourth order cardinal Chebyshev B-splines

Let $\lambda = (\delta_1, \delta_2)$, where $\delta_1 = (\lambda_1, \lambda_2) = (0, 0)$, $\delta_2 = (\lambda_3, \lambda_4) = (-i\alpha, i\alpha)$, and

$$N^4_\lambda(x) = N^2_{\delta_1} * N^2_{\delta_2}(x).$$

Then, we assume that

$$\varphi_j(t) = N^4_{2^{-j}\lambda}(t). \tag{30}$$

Then

$$\widehat{\varphi_j}(\omega) = \widehat{N^2_{2^{-j}\delta_1}}(\omega)\widehat{N^2_{2^{-j}\delta_2}}(\omega),$$

$$
\begin{aligned}
&= \left(\frac{1}{16} + \frac{2 + 2\cos(2^{-(j+1)}\alpha)}{16} e^{-i\frac{\omega}{2}} + \frac{2 + 4\cos(2^{-(j+1)}\alpha)}{16} e^{-i\omega} \right. \\
&\quad \left. + \frac{2 + 2\cos(2^{-(j+1)}\alpha)}{16} e^{-i\frac{3\omega}{2}} + \frac{1}{16} e^{2i\omega} \right) \widehat{N^4_{2^{-(j+1)}\lambda}} \left(\frac{\omega}{2} \right) \\
&= \mathcal{H}_j \left(\frac{\omega}{2} \right) \widehat{\varphi_{j+1}} \left(\frac{\omega}{2} \right),
\end{aligned}
$$

Following the same steps as before, we obtain the normalized coefficients:

$$
\begin{cases}
h_j[0] = h_j[4] = \dfrac{\sqrt{2}}{8 + 8(\cos(2^{-(j+1)}\alpha)}, \\
h_j[1] = h_j[3] = \dfrac{2\sqrt{2}(\cos(2^{-(j+1)}\alpha) + 1)}{8 + 8(\cos(2^{-(j+1)}\alpha)}, \\
h_j[2] = \dfrac{2\sqrt{2}(2\cos(2^{-(j+1)}\alpha) + 1)}{8 + 8(\cos(2^{-(j+1)}\alpha)}.
\end{cases}
$$

As before, let us assume that $\widetilde{h}_j[-1], \widetilde{h}_j[0], \widetilde{h}_j[1], \widetilde{h}_j[2], \widetilde{h}_j[3], \widetilde{h}_j[4], \widetilde{h}_j[5]$ are the filters coefficients of dual scaling functions. Moreover, to ensure that the dual scaling functions are symmetric, we set

$$
\widetilde{h}_j[-1] = \widetilde{h}_j[5], \quad \widetilde{h}_j[0] = \widetilde{h}_j[4], \quad \widetilde{h}_j[1] = \widetilde{h}_j[3].
$$

The relation (14) implies

$$
2\widetilde{h}_j[5] + 2\widetilde{h}_j[4] + 2\widetilde{h}_j[3] + \widetilde{h}_j[2] = \sqrt{2}. \tag{31}
$$

Using (10), we find, for $k = 0$,

$$
\begin{aligned}
&\widetilde{h}_j[4] + 2\left(\cos(2^{-(j+1)}\alpha) + 1\right)\widetilde{h}_j[3] + \left(2\cos(2^{-(j+1)}\alpha) + 1\right)\widetilde{h}_j[2] \\
&\qquad = \frac{8 + 8(\cos(2^{-(j+1)}\alpha)}{2\sqrt{2}},
\end{aligned} \tag{32}
$$

for $k = 1$,

$$
\begin{aligned}
&\widetilde{h}_j[2] + 2\left(\cos(2^{-(j+1)}\alpha) + 1\right)\widetilde{h}_j[3] + 2\left(2\cos(2^{-(j+1)}\alpha) + 1\right)\widetilde{h}_j[4] \\
&\qquad + 2\left(\cos(2^{-(j+1)}\alpha) + 1\right)\widetilde{h}_j[5] = 0,
\end{aligned} \tag{33}
$$

and for $k = 2$,

$$
\widetilde{h}[4] + 2\left(\cos(2^{-(j+1)}\alpha) + 1\right)\widetilde{h}_j[5] = 0. \tag{34}
$$

As in $n = 2$, we take the following additional condition to simplify the computation:

$$\sum_{k=-4}^{2} (-1)^k \widetilde{h}_j[1-k] = 0,$$

which implies

$$2\widetilde{h}_j[5] - 2\widetilde{h}_j[4] + 2\widetilde{h}_j[3] - \widetilde{h}_j[2] = 0. \tag{35}$$

Finally, we find $\widetilde{h}_j[2], \widetilde{h}_j[3], \widetilde{h}_j[4]$ and $\widetilde{h}_j[5]$ by the following formulas:

$$\begin{cases} \widetilde{h}_j[-1] = \widetilde{h}_j[5] = \dfrac{\sqrt{2}\cos(2^{-(j+1)}\alpha) + 2\sqrt{2}}{16\cos(2^{-(j+1)}\alpha) + 8\cos(2^{-j}\alpha) + 8}, \\ \widetilde{h}_j[0] = \widetilde{h}_j[4] = \dfrac{-6\sqrt{2}\cos(2^{-(j+1)}\alpha) - \sqrt{2}\cos(2^{-j}\alpha) - 5\sqrt{2}}{16\cos(2^{-(j+1)}\alpha) + 8\cos(2^{-j}\alpha) + 8}, \\ \widetilde{h}_j[1] = \widetilde{h}_j[3] \\ \qquad = \dfrac{8\sqrt{2}\cos(2^{-(j+1)}\alpha) + 7\sqrt{2}\cos(2^{-j}\alpha) + 2\sqrt{2}\cos(3\,2^{-(j+1)}\alpha) + 3\sqrt{2}}{56\cos(2^{-(j+1)}\alpha) + 32\cos(2^{-j}\alpha) + 8\cos(3\,2^{-(j+1)}\alpha) + 32}, \\ \widetilde{h}_j[2] = \dfrac{10\sqrt{2}\cos(2^{-(j+1)}\alpha) + 3\sqrt{2}\cos(2^{-j}\alpha) + 7\sqrt{2}}{8\cos(2^{-(j+1)}\alpha) + 4\cos(2^{-j}\alpha) + 4}. \end{cases}$$

3.4 Fifth order cardinal Chebyshev B-splines

Let $\lambda = (\delta_1, \delta_2)$, where $\delta_1 = (\lambda_1, \lambda_2, \lambda_3) = (0, 0, 0)$, $\delta_2 = (\lambda_4, \lambda_5) = (-i\alpha, i\alpha)$, and

$$N_\lambda^5(x) = N_{\delta_1}^3 * N_{\delta_2}^2(x).$$

Then we assume that

$$\varphi_j(t) = N_{2^{-j}\lambda}^5(t). \tag{36}$$

Therefore,

$$\begin{aligned} \widehat{\varphi_j}(\omega) &= \widehat{N_{2^{-j}\delta_1}^3}(\omega)\widehat{N_{2^{-j}\delta_2}^2}(\omega), \\ &= \frac{1}{32}\Big(1 + (1 + 4\cos(2^{-(j+1)}\alpha)e^{-i\frac{\omega}{2}} \end{aligned}$$

$$
\begin{aligned}
&+(4+4\cos(2^{-(j+1)}\alpha)+2\cos(2^{-j}\alpha))e^{-i\omega}\\
&+(4+4\cos(2^{-(j+1)}\alpha)+2\cos(2^{-j}\alpha))e^{-i\frac{3\omega}{2}}\\
&+(1+4\cos(2^{-(j+1)}\alpha)e^{-2i\omega}+e^{-i\frac{5\omega}{2}}\Big)\widehat{N^5_{2^{-j}\lambda}}\left(\frac{\omega}{2}\right)\\
&=\mathcal{H}_j\left(\frac{\omega}{2}\right)\widehat{\varphi_{j+1}}\left(\frac{\omega}{2}\right),
\end{aligned}
$$

So, the normalized coefficients are given by

$$
\begin{cases}
h_j[0]=h_j[5]=\dfrac{\sqrt{2}}{12+16\cos(2^{-(j+1)}\alpha)+4\cos(2^{-j}\alpha)},\\
h_j[1]=h_j[4]=\dfrac{\sqrt{2}(1+4\cos(2^{-(j+1)}\alpha)}{12+16\cos(2^{-(j+1)}\alpha)+4\cos(2^{-j}\alpha)},\\
h_j[2]=h_j[3]=\dfrac{2\sqrt{2}(2+2\cos(2^{-(j+1)}\alpha)+\cos(2^{-j}\alpha))}{12+16\cos(2^{-(j+1)}\alpha)+4\cos(2^{-j}\alpha)}.
\end{cases}
$$

We assume that $\widetilde{h}_j[-1],\widetilde{h}_j[0],\widetilde{h}_j[1],\widetilde{h}_j[2],\widetilde{h}_j[3],\widetilde{h}_j[4],\widetilde{h}_j[5]$, and $\widetilde{h}_j[6]$ are the filters coefficients of dual scaling functions. Moreover, to ensure that the dual scaling functions are symmetric, we put

$$
\widetilde{h}_j[-1]=\widetilde{h}_j[6],\quad \widetilde{h}_j[0]=\widetilde{h}_j[5],\quad \widetilde{h}_j[1]=\widetilde{h}_j[4],\quad \widetilde{h}_j[2]=\widetilde{h}_j[3].
$$

The relation (14) implies

$$
2\widetilde{h}_j[3]+2\widetilde{h}_j[4]+2\widetilde{h}_j[5]+2\widetilde{h}_j[6]=\sqrt{2}. \tag{37}
$$

Using (10), we find, for $k=0$,

$$
\begin{aligned}
&(4+4\cos(2^{-(j+1)}\alpha)+2\cos(2^{-j}\alpha))\widetilde{h}_j[3]+(1+4\cos(2^{-(j+1)}\alpha)\widetilde{h}_j[4]+\widetilde{h}_j[5]\\
&\qquad=\frac{12+16\cos(2^{-(j+1)}\alpha)+4\cos(2^{-j}\alpha)}{2\sqrt{2}},
\end{aligned} \tag{38}
$$

for $k=1$,

$$
\begin{aligned}
&(2+4\cos(2^{-(j+1)}\alpha)\widetilde{h}_j[3]+(4+4\cos(2^{-(j+1)}\alpha)+2\cos(2^{-j}\alpha))\widetilde{h}_j[4]\\
&+(4+4\cos(2^{-(j+1)}\alpha)+2\cos(2^{-j}\alpha))\widetilde{h}_j[5]+(1+4\cos(2^{-(j+1)}\alpha)\widetilde{h}_j[6]\\
&\quad=0,
\end{aligned} \tag{39}
$$

and for $k=2$,

$$\widetilde{h}[4] + (1 + 4\cos(2^{-(j+1)}\alpha)\widetilde{h}_j[5] + (4 + 4\cos(2^{-(j+1)}\alpha) + 2\cos(2^{-j}\alpha))\widetilde{h}_j[6] = 0. \tag{40}$$

Finally, we find $\widetilde{h}_j[3], \widetilde{h}_j[4], \widetilde{h}_j[5]$, and $\widetilde{h}_j[6]$ by the following formulas:

$$\begin{cases} \widetilde{h}_j[-1] = \widetilde{h}_j[6] = 0, \\ \widetilde{h}_j[0] = \widetilde{h}_j[5] = \dfrac{2\sqrt{2}\cos(2^{-(j+1)}\alpha) + \sqrt{2}}{12\cos(2^{-(j+1)}\alpha) + 4\cos(3\,2^{-(j+1)}\alpha)}, \\ \widetilde{h}_j[1] = \widetilde{h}_j[4] = \dfrac{-6\sqrt{2}\cos(2^{-(j+1)}\alpha) - 4\sqrt{2}\cos(2^{-j}\alpha) - 5\sqrt{2}}{12\cos(2^{-(j+1)}\alpha) + 4\cos(3\,2^{-(j+1)}\alpha)}, \\ \widetilde{h}_j[2] = \widetilde{h}_j[3] \\ \qquad = \dfrac{5\sqrt{2}\cos(2^{-(j+1)}\alpha) + 2\sqrt{2}\cos(2^{-j}\alpha) + \sqrt{2}\cos(3\,2^{-(j+1)}\alpha) + 2\sqrt{2}}{6\cos(2^{-(j+1)}\alpha) + 2\cos(3\,2^{-(j+1)}\alpha)}. \end{cases}$$

3.5 Sixth order cardinal Chebyshev B-splines

Let $\lambda = (\delta_1, \delta_2)$, where $\delta_1 = (\lambda_1, \lambda_2, \lambda_3) = (0, 0, 0)$, $\delta_2 = (\lambda_4, \lambda_5, \lambda_6) = (0, -i\alpha, i\alpha)$, and

$$N_\lambda^6(x) = N_{\delta_1}^3 * N_{\delta_2}^3(x).$$

Then, we assume that

$$\varphi_j(t) = N_{2^{-j}\lambda}^6(t). \tag{41}$$

Therefore,

$$\begin{aligned} \widehat{\varphi_j}(\omega) &= \widehat{N_{2^{-j}\delta_1}^3}(\omega)\widehat{N_{2^{-j}\delta_2}^3}(\omega), \\ &= \frac{1}{64}\Big(1 + (2 + 4\cos(2^{-(j+1)}\alpha))e^{-i\frac{\omega}{2}} + (5 + 8\cos(2^{-(j+1)}\alpha) \\ &\quad + 2\cos(2^{-j}\alpha))e^{-i\omega} + (8 + 8\cos(2^{-(j+1)}\alpha) + 4\cos(2^{-j}\alpha))e^{-i\frac{3\omega}{2}} \\ &\quad + (5 + 8\cos(2^{-(j+1)}\alpha) + 2\cos(2^{-j}\alpha))e^{-2i\omega} \\ &\quad + (2 + 4\cos(2^{-(j+1)}\alpha))e^{-i\frac{5\omega}{2}} + e^{-3i\omega}\Big)\widehat{N_{2^{-j}\lambda}^6}\left(\frac{\omega}{2}\right) \\ &= \mathcal{H}_j\left(\frac{\omega}{2}\right)\widehat{\varphi_{j+1}}\left(\frac{\omega}{2}\right). \end{aligned}$$

Thus, the normalized coefficients are expressed as

$$\begin{cases} h_j[0] = h_j[6] = \dfrac{\sqrt{2}}{24 + 32\cos(2^{-(j+1)}\alpha) + 8\cos(2^{-j}\alpha)}, \\ h_j[1] = h_j[5] = \dfrac{2\sqrt{2}\left(1 + 2\cos(2^{-(j+1)}\alpha)\right)}{24 + 32\cos(2^{-(j+1)}\alpha) + 8\cos(2^{-j}\alpha)}, \\ h_j[2] = h_j[4] = \dfrac{\sqrt{2}\left(5 + 8\cos(2^{-(j+1)}\alpha) + 2\cos(2^{-j}\alpha)\right)}{24 + 32\cos(2^{-(j+1)}\alpha) + 8\cos(2^{-j}\alpha)}, \\ h_j[3] = \dfrac{2\sqrt{2}\left(4 + 4\cos(2^{-(j+1)}\alpha) + 2\cos(2^{-j}\alpha)\right)}{24 + 32\cos(2^{-(j+1)}\alpha) + 8\cos(2^{-j}\alpha)}. \end{cases}$$

As in previous instances, let us assume that $\widetilde{h}_j[-1]$, $\widetilde{h}_j[0]$, $\widetilde{h}_j[1]$, $\widetilde{h}_j[2]$, $\widetilde{h}_j[3]$, $\widetilde{h}_j[4]$, $\widetilde{h}_j[5]$, $\widetilde{h}_j[6]$, and $\widetilde{h}_j[7]$, where

$$\widetilde{h}_j[-1] = \widetilde{h}_j[7], \quad \widetilde{h}_j[0] = \widetilde{h}_j[6], \quad \widetilde{h}_j[1] = \widetilde{h}_j[5], \quad \widetilde{h}_j[2] = \widetilde{h}_j[4].$$

The relation (14) implies

$$\widetilde{h}_j[3] + 2\widetilde{h}_j[4] + 2\widetilde{h}_j[5] + 2\widetilde{h}_j[6] + 2\widetilde{h}_j[7] = \sqrt{2}. \tag{42}$$

Using (10), we find, for $k = 0$

$$\begin{aligned} &\left(4 + 4\cos(2^{-(j+1)}\alpha) + 2\cos(2^{-j}\alpha)\right)\widetilde{h}_j[3] + \left(5 + 8\cos(2^{-(j+1)}\alpha) + 2\cos(2^{-j}\alpha)\right)\widetilde{h}_j[4] \\ &\quad + \left(2 + 4\cos(2^{-(j+1)}\alpha)\right)\widetilde{h}_j[5] + \widetilde{h}_j[6] = \frac{24 + 32\cos(2^{-(j+1)}\alpha) + 8\cos(2^{-j}\alpha)}{2\sqrt{2}}, \end{aligned} \tag{43}$$

for $k = 1$,

$$\begin{aligned} &\left(2 + 4\cos(2^{-(j+1)}\alpha)\right)\widetilde{h}_j[3] + \left(6 + 8\cos(2^{-(j+1)}\alpha) + 2\cos(2^{-j}\alpha)\right)\widetilde{h}_j[4] \\ &+ \left(8 + 8\cos(2^{-(j+1)}\alpha) + 4\cos(2^{-j}\alpha)\right)\widetilde{h}_j[5] \\ &+ \left(5 + 8\cos(2^{-(j+1)}\alpha) + 2\cos(2^{-j}\alpha)\right)\widetilde{h}_j[6] \\ &+ \left(2 + 4\cos(2^{-(j+1)}\alpha)\right)\widetilde{h}_j[7] = 0, \end{aligned} \tag{44}$$

for $k = 2$,

$$\begin{aligned} &\widetilde{h}_j[4] + \left(2 + 4\cos(2^{-(j+1)}\alpha)\right)\widetilde{h}_j[5] \\ &+ \left(5 + 8\cos(2^{-(j+1)}\alpha) + 2\cos(2^{-j}\alpha)\right)\widetilde{h}_j[6] \end{aligned}$$

$$+\left(8+8\cos(2^{-(j+1)}\alpha)+4\cos(2^{-j}\alpha)\right)\widetilde{h}_j[7]=0, \tag{45}$$

and for $k=3$,

$$\widetilde{h}_j[6]+\left(2+4\cos(2^{-(j+1)}\alpha)\right)\widetilde{h}_j[7]=0. \tag{46}$$

We take the following additional condition to simplify the computation:

$$\sum_{k=-6}^{2}(-1)^k\widetilde{h}_j[1-k]=0.$$

Finally, we can derive explicit formulas of $\widetilde{h}_j[3], \widetilde{h}_j[4], \widetilde{h}_j[5], \widetilde{h}_j[6]$, and $\widetilde{h}_j[7]$ as follows:

$$\begin{cases}
\widetilde{h}_j[-1]=\widetilde{h}_j[7]=\dfrac{-8\sqrt{2}\beta^2-9\sqrt{2}\beta-\sqrt{2}\beta\beta_2-2\sqrt{2}}{64\beta^2+48\beta+16\beta\beta_2^2+64\beta^2\beta_2+64\beta\beta_2},\\
\widetilde{h}_j[0]=\widetilde{h}_j[6]=\dfrac{16\sqrt{2}\beta^3+26\sqrt{2}\beta^2+13\sqrt{2}\beta+2\sqrt{2}\beta^2\beta_2+\sqrt{2}\beta\beta_2+2\sqrt{2}}{32\beta^2+24\beta+8\beta\beta_2^2+32\beta^2\beta_2+32\beta\beta_2},\\
\widetilde{h}_j[1]=\widetilde{h}_j[5]=\dfrac{-8\sqrt{2}\beta-\sqrt{2}\beta_2-5\sqrt{2}}{8\beta_2+8},\\
\widetilde{h}_j[2]=\widetilde{h}_j[4]\\
\qquad=\dfrac{-16\sqrt{2}\beta^3-18\sqrt{2}\beta^2-7\sqrt{2}\beta+2\sqrt{2}\beta\beta_2^2+6\sqrt{2}\beta^2\beta_2+7\sqrt{2}\beta\beta_2-2\sqrt{2}}{32\beta^2+24\beta+8\beta\beta_2^2+32\beta^2\beta_2+32\beta\beta_2},\\
\widetilde{h}_j[3]=\dfrac{64\sqrt{2}\beta^3+112\sqrt{2}\beta^2+51\sqrt{2}\beta+6\sqrt{2}\beta\beta_2^2+40\sqrt{2}\beta^2\beta_2+33\sqrt{2}\beta\beta_2+2\sqrt{2}}{32\beta^2+24\beta+8\beta\beta_2^2+32\beta^2\beta_2+32\beta\beta_2},
\end{cases}$$

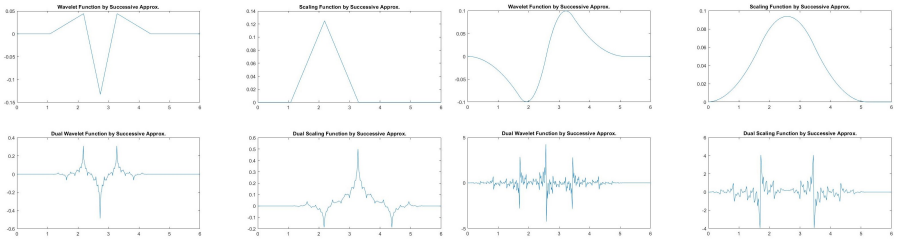
where

$$\beta=\cos\left(2^{-(j+1)}\alpha\right),\quad \beta_2=\cos\left(2^{-j}\alpha\right).$$

The figures presented below (Figure 1) illustrate the scaling and wavelet functions associated with the previously obtained biorthogonal filter banks.

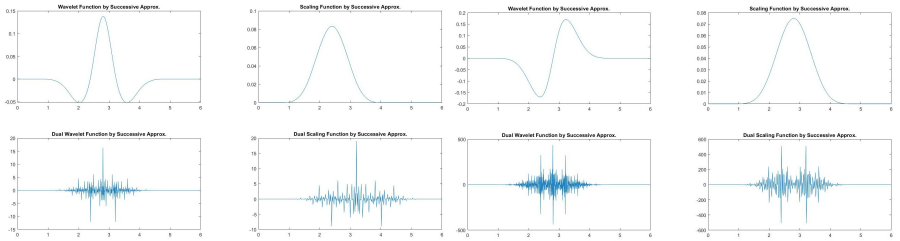
4 Application of nonstationary biorthogonal filter banks for image compression

In this section of our comparative study, we analyze the performance of nonstationary biorthogonal filter banks, derived in the previous section with parameters $\mathbf{n}=\mathbf{6}$, in comparison to the widely used Daubechies biorthogonal



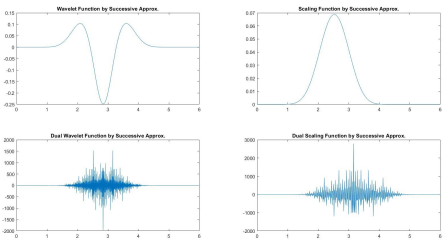
(a) φ_j, $\widetilde{\varphi}_j$, ψ_j, and $\widetilde{\varphi}_j$ for $\alpha = 0$, $j = 0$, and $n = 2$

(b) φ_j, $\widetilde{\varphi}_j$, ψ_j and $\widetilde{\varphi}_j$ for $\alpha = 0$, $j = 0$ and $n = 3$



(c) φ_j, $\widetilde{\varphi}_j$, ψ_j, and $\widetilde{\varphi}_j$ for $\alpha = 0$, $j = 0$, and $n = 4$

(d) φ_j, $\widetilde{\varphi}_j$, ψ_j, and $\widetilde{\varphi}_j$ for $\alpha = 0$, $j = 0$, and $n = 5$



(e) φ_j, $\widetilde{\varphi}_j$, ψ_j, and $\widetilde{\varphi}_j$ for $\alpha = 0$, $j = 0$, and $n = 6$

Figure 1: Curves of scaling functions and wavelet functions

filter bior **bior 3.3**, **bior 4.4**, and **bior 5.5** [3]. The evaluation is conducted using the peak signal-to-noise ratio (PSNR) quality measure. The images considered in our study include pepper, bamboo, Barbara, lighthouse, boat, X-ray, and a textured image. The results of our analysis are presented in Table 1:

Table 1: Comparison of PSNR values for different images using stationary and nonstationary biorthogonal filter banks **(PS-NS)**.

Filter / Image	**bior 3.3**	**bior 4.4**	**bior 5.5**	**PS-NS**	**Optimal value of α**
Pepper image	373.0289	321.8209	317.6718	377.4309	**0.4**
Bamboo image	373.3974	312.1773	310.2757	376.0176	**1.4**
Barbara image	370.5101	312.7138	311.3224	373.5105	**1.4**
Lighthouse image	368.791	315.1341	312.1767	372.6814	**3**
Boat image	369.5577	316.4587	312.8915	372.7301	**0.1**
X-ray image	374.7837	321.2239	317.7600	379.4312	**0.4**
Textured image	372.4533	309.4582	305.5483	375.9778	**1.3**

Nonstationary filters optimized with an adaptive parameter α consistently offer better image reconstruction quality, as measured by a higher PSNR. Although the improvement varies depending on the image, this demonstrates the potential of nonstationary filters to outperform traditional biorthogonal approaches in terms of reconstruction fidelity.

In this section, we further analyze the effectiveness of these filter banks in image compression. Beyond the PSNR measure discussed earlier, we evaluate compression performance by quantifying the number of zeros introduced in the processed image. A higher count of zeros signifies greater compression efficiency, as it indicates a substantial reduction in redundant information

while retaining key image details. This metric is crucial for minimizing file size while maintaining satisfactory visual quality.

For each image, the parameter α is optimized to maximize the number of zeros, ensuring the best compression performance for the specific image. Thresholds are selected adaptively based on the characteristics of each image, striking a balance between maintaining visual quality and enhancing compression efficiency.

Table 2: Comparison of the number of zeros in the image after compression using stationary and nonstationary biorthogonal filter banks **(PS-NS)**.

Image	**Filter** / **Threshold**	**bior 3.3**	**bior 4.4**	**bior 5.5**	**PS-NS**	**Optimal value of α**
Pepper image	0.01	200271	195713	197236	202269	**3.4**
	0.05	201226	202617	204126	203773	**5.5**
	0.1	201408	202871	204126	205864	**5.6**
	0.15	201698	203119	204618	207672	**5.6**
	0.2	202298	203660	205135	207453	**5.4**
Bamboo image	0.1	205408	203566	203099	203737	**0.1**
	0.2	217290	217299	218389	218952	**0.1**
	0.3	223279	223417	224895	226805	**7.9**
	0.4	227922	228209	229755	232402	**7.9**
	0.5	232562	233066	234571	236767	**7.9**
Barbara image	0.1	193788	193613	194121	195190	**0.1**
	0.2	199199	199351	200434	203116	**0.2**
	0.3	203980	203945	204867	208590	**7.8**
	0.4	209261	209146	210240	214263	**7.9**
	0.5	213573	213464	214480	218510	**7.9**
Lighthouse image	0.1	200554	200951	201971	202509	**0.2**
	0.15	201904	202386	203638	204443	**2.3**
	0.2	202665	203358	204573	205806	**2.5**
	0.25	203531	204284	205498	206993	**3**
	0.3	204524	205309	206534	208083	**3**

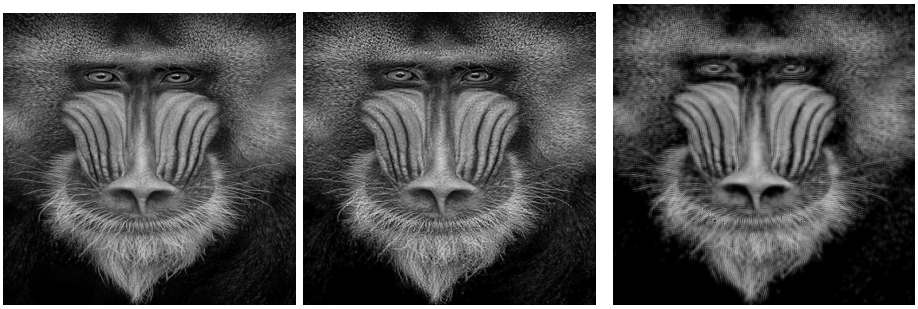
Boat image	0.15	202593	203857	205311	206093	**5.3**
	0.3	206465	207617	208973	210035	**7.4**
	0.45	209932	210944	212315	213802	**7.3**
	0.6	212121	213224	214562	217194	**7.3**
	0.75	214717	215921	217330	220750	**7.2**
X-ray image	0.05	201017	202083	203617	203654	**4.6**
	0.1	201263	202711	204266	206469	**4.8**
	0.15	201330	202797	204349	209284	**4.8**
	0.2	232677	234175	235987	238573	**4.5**
	0.25	238801	240088	241793	244086	**7.4**
Textured image	0.05	202284	197225	197596	206985	**0.1**
	0.1	211203	208244	207600	212592	**0.1**
	0.15	216815	215130	214592	217219	**0.1**
	0.2	221580	220335	219984	221975	**0.1**
	0.25	225770	225134	225145	226346	**0.1**

The nonstationary filter consistently outperforms all other filters, achieving the highest number of zeros across all threshold values and images. This performance is particularly notable when compared to **bior 5.5**, highlighting its superior effectiveness in compression (Table 2).

The figures presented below (Figure 2–8) provide a qualitative evaluation of the reconstructed images after compression using the nonstationary biorthogonal filter bank **PS-NS**. These visualizations validate the filter's capability to achieve high compression efficiency without compromising perceptual image quality.

(a) Original pepper image. (b) Reconstructed pepper im-(c) Reconstructed pepper image for $th = 0.01$ and $\alpha = 3.4$.age for $th = 0.2$ and $\alpha = 5$.

Figure 2: Visualization of the original pepper image and its reconstructions with nonstationary filter.



(a) Original mandrill image. (b) Reconstructed mandrill image for $th = 0.1$ and $\alpha = 0.1$.(c) Reconstructed mandrill image for $th = 0.5$ and $\alpha = 7.9$.

Figure 3: Visualization of the original mandrill image and its reconstructions with nonstationary filter.

(a) Original Barbara image. (b) Reconstructed Barbara image for $th = 0.1$ and $\alpha = 0.1$. (c) Reconstructed Barbara image for $th = 0.5$ and $\alpha = 7.9$.

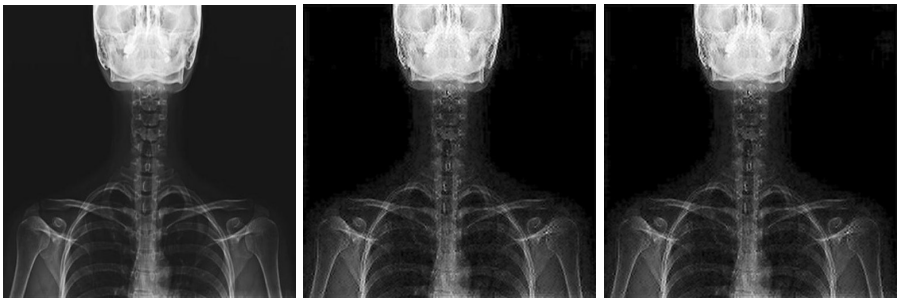
Figure 4: Visualization of the original Barbara image and its reconstructions with nonstationary filter.

(a) Original lighthouse image. (b) Reconstructed lighthouse image for $th = 0.1$ and $\alpha = 0.2$.(c) Reconstructed lighthouse image for $th = 0.3$ and $\alpha = 3$.

Figure 5: Visualization of the original lighthouse image and its reconstructions with nonstationary filter.

(a) Original boat image. (b) Reconstructed boat image for $th = 0.15$ and $\alpha = 5.3$. (c) Reconstructed boat image for $th = 0.75$ and $\alpha = 7.2$.

Figure 6: Visualization of the original boat image and its reconstructions with nonstationary filter.



(a) Original X-ray image. (b) Reconstructed X-ray image for $th = 0.2$ and $\alpha = 4.5$. (c) Reconstructed X-ray image for $th = 0.25$ and $\alpha = 7.4$.

Figure 7: Visualization of the original X-ray image and its reconstructions with nonstationary filter.

(a) Original textured image. (b) Reconstructed textured image for $th = 0.2$ and $\alpha = 0.1$. (c) Reconstructed textured image for $th = 0.25$ and $\alpha = 0.1$.

Figure 8: Visualization of the original textured image and its reconstructions with nonstationary filter.

5 Conclusion

In conclusion, nonstationary biorthogonal wavelets are powerful tools for signal analysis, with recent advancements showcasing the utility of cardinal Chebyshev B-splines in their construction. The proposed method simplifies filter design by explicitly formulating filters and dual filters at all scaling levels, enabling efficient implementation through numerical methods like Gauss elimination or LU factorization. This streamlined approach, combined with the practical applications of nonstationary wavelets, underscores their importance in image processing.

Filters based on cardinal Chebyshev B-splines offer potential improvements in compression quality and computational efficiency, although performance may depend on the specific characteristics of the images processed.

Future research could focus on optimizing the method for high-dimensional data and integrating it into machine learning workflows, broadening the applicability of nonstationary wavelets across scientific and engineering domains.

Acknowledgements

The authors sincerely appreciate the constructive feedback and valuable suggestions provided by the anonymous referees and the editor.

References

[1] Boor, C., DeVore, R.A. and Ron, A. *On the construction of multivariate (pre) wavelets,* Constr. Approx. 9 (1993) 123–166.

[2] Brahimi, M., Melkemi, K. and Boussaad, A. *Design of nonstationary wavelets through the positive solution of Bezout's equation,* J. Interdiscip. Math. 24(3), (2021) 553–565.

[3] Daubechies, I. Ten lectures on wavelets, SIAM. 1992.

[4] Gasquet, C. and Witomski, P. *Analyse de Fourier et applications: filtrage, calcul numérique, ondelettes,* Elsevier Masson, 1990.

[5] Lee, Y.J. and Yoon, J. *Analysis of compactly supported nonstationary biorthogonal wavelet systems based on exponential B-splines,*In Abstract and Applied Analysis, vol. 2011, no. 1, p. 593436. Hindawi Publishing Corporation, 2011.

[6] Mallat, S.G. *Multiresolution approximations and wavelet orthonormal bases of* $L^2(\mathbb{R})$, Trans. Am. Math. Soc. 315(1) (1989) 69–87.

[7] Mazure, M.L., and Melkemi, K. Orthonormality of Cardinal Chebyshev B-spline Bases in Weighted Sobolev Spaces, Constr. Approx. 18(3) (2002) 387–415.

[8] Melkemi, K. *Orthogonalité des B-splines de Chebyshev cardinales dans un espace de Sobolev pondéré,* PhD diss., University Joseph-Fourier-Grenoble I, 1999.

[9] Soman, K. P. Insight into wavelets: from theory to practice, PHI Learning Pvt. Ltd, 2010.

[10] Vonesch, C., Blu, T., and Unser, M. Generalized Daubechies wavelet families, IEEE Trans. Signal Process. 55(9) (2007) 4415-4429.

[11] Zhang, B., Zheng, H., Zhou, J., and Pan, L. Construction of a family of non-stationary biorthogonal wavelets, J. Inequ. Appl. 1 (2019) 1–15.

Iranian Journal of Numerical Analysis and Optimization
Vol. 15, No. 2, 2025, pp 728–769
https://doi.org/10.22067/ijnao.2025.90299.1536
https://ijnao.um.ac.ir/

Research Article

A parameter uniform hybrid approach for singularly perturbed two-parameter parabolic problem with discontinuous data

N. Roy*, A. Jha

Abstract

In this article, we address singularly perturbed two-parameter parabolic problem of the reaction-convection-diffusion type in two dimensions. These problems exhibit discontinuities in the source term and convection coefficient at particular domain points, which result in the formation of interior layers. The presence of two perturbation parameters leads to the formation of boundary layers with varying widths. Our primary focus is to address these layers and develop a scheme that is uniformly convergent. So we propose a hybrid monotone difference scheme for the spatial direction, im-

*Corresponding author

Received 16 October 2024; revised 6 Febryary 2025; accepted 5 March 2025

Nirmali Roy
Department of Science and Mathematics, Indian Institute of Information Technology Guwahari, Bongora, Assam,781015. e-mail: nirmali@iiitg.ac.in

Anuradha Jha
Department of Science and Mathematics, Indian Institute of Information Technology Guwahari, Bongora, Assam,781015. e-mail: anuradha@iiitg.ac.in

How to cite this article
Roy, N. and Jha, A., A parameter uniform hybrid approach for singularly perturbed two-parameter parabolic problem with discontinuous data. *Iran. J. Numer. Anal. Optim.*, 2025; 15(2): 728-769. https://doi.org/10.22067/ijnao.2025.90299.1536

plemented on a specially designed piece-wise uniform Shishkin mesh, combined with the Crank–Nicolson method on a uniform mesh for the temporal direction. The resulting scheme is proven to be uniformly convergent, with an order of almost two in the spatial direction and exactly two in the temporal direction. Numerical experiments support the theoretically proven higher order of convergence and show that our approach results in better accuracy and convergence compared to other existing methods in the literature.

AMS subject classifications (2020): 65M06, 65M12, 65M15

Keywords: Interior layers; Hybrid Method; Singularly Perturbed; Parabolic problem; Shishkin mesh.

1 Introduction

We have considered a two-parameter singularly perturbed parabolic initial boundary value problem with nonsmooth data:

$$\begin{aligned}
&\mathcal{L}u(x,t) \equiv (\epsilon u_{xx} + \mu a u_x - bu - u_t)(x,t) = f(x,t), \qquad (x,t) \in (\Omega^- \cup \Omega^+),\\
&u(0,t) = p(t), \qquad (0,t) \in \Gamma_l = \{(0,t) | 0 \le t \le T\},\\
&u(1,t) = r(t), \qquad (1,t) \in \Gamma_r = \{(1,t) | 0 \le t \le T\},\\
&u(x,0) = q(x), \qquad (x,0) \in \Gamma_b = \{(x,0) | 0 \le x \le 1\},
\end{aligned} \tag{1}$$

where $0 < \epsilon \ll 1$ and $0 < \mu \le 1$ are two singular perturbation parameters. Let $d \in \Gamma = (0,1)$, $\Gamma^- = (0,d)$, $\Gamma^+ = (d,1)$, $\Gamma_c = \Gamma_l \cup \Gamma_b \cup \Gamma_r$ and $\Omega = (0,1) \times (0,T]$, $\Omega^- = (0,d) \times (0,T]$, $\Omega^+ = (d,1) \times (0,T]$. The convection coefficient $a(x,t)$ and source term $f(x,t)$ experience discontinuities at $(d,t) \in \Omega$ for all t. Moreover, $a(x,t) \le -\alpha_1 < 0$ for $(x,t) \in \Omega^-$ and $a(x,t) \ge \alpha_2 > 0$ for $(x,t) \in \Omega^+$, where α_1, α_2 are positive constants. These functions are sufficiently smooth on $\Omega^- \cup \Omega^+$. We assume that the jumps of $a(x,t)$ and $f(x,t)$ at (d,t) satisfy $|[a](d,t)| < C$ and $|[f](d,t)| < C$, where the jump of ω at (d,t) is $[\omega](d,t) = \omega(d+,t) - \omega(d-,t)$. Additionally, the coefficient $b(x,t)$ is assumed to be a sufficiently smooth function on Ω such that $b(x,t) \ge \beta > 0$.

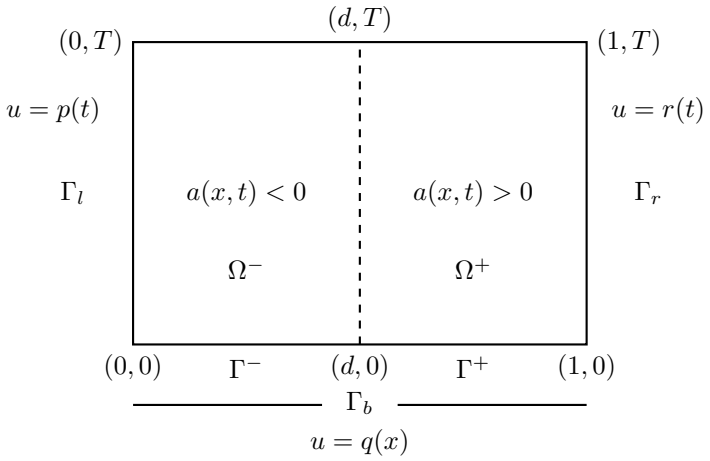


Figure 1: Domain

The boundary data $p(t), r(t)$ and initial data $q(x)$ are sufficiently smooth on the domain and satisfy the compatibility condition at the two corner points $(0,0)$ and $(1,0)$:

$$q(0) = p(0),$$

$$q(1) = r(0),$$

as well as

$$\epsilon\frac{\partial^2 q(0)}{\partial x^2} + \mu a(0)\frac{\partial q(0)}{\partial x} - b(0)q(0) - f(0,0) = \frac{\partial p(0)}{\partial t},$$

$$\epsilon\frac{\partial^2 q(1)}{\partial x^2} + \mu a(1)\frac{\partial q(1)}{\partial x} - b(1)q(1) - f(1,0) = \frac{\partial r(0)}{\partial t}.$$

Similar compatibility conditions apply at the transition point $(d,0)$; see Figure 1. Under these assumptions, the problem (1) has a unique continuous solution in the domain $\bar{\Omega}$. The solution contains strong interior layers at the points of discontinuity (d,t) for all $t \in (0,T]$ due to the discontinuity in the convection coefficient and source term. Additionally, the solution exhibits boundary layers at Γ_l and Γ_r due to the presence of perturbation parameters ϵ and μ. Singularly perturbed two-parameter parabolic problems are not extensively studied, and there is a scarcity of well-established numerical methods tailored to solve them effectively, especially in comparison to the abundance

of methods available for one-parameter singularly perturbed parabolic problems. A number of researchers have developed a numerical technique for a system singularly perturbed two-parameter parabolic problems with smooth data [1, 10, 11, 14, 16, 19, 24]. Singh and Kumar [21] developed a numerical techniques for a system of singularly perturbed parabolic convection-diffusion equations. Finding a uniform numerical solution for singularly perturbed parabolic equations with discontinuous data is more challenging. For the problem (1), Chandru et al. [3, 4] showed that the accuracy of the upwind scheme on Shishkin mesh in space and the backward Euler scheme on uniform mesh in time is nearly first-order. Kumar and Kumari [12] used the Crank–Nicolson method in time on a uniform mesh and the upwind method in space on an appropriately defined Shishkin mesh to achieve a second order accurate method in time and almost first order accurate in space. Singh, Choudhary, and Kumar [22] presented an implicit scheme based on the Crank–Nicolson method in time and trigonometric B-spline basis functions on Shishkin mesh in space. This parameter uniform method is proved to be almost first order convergent in space and second order convergent in time.

With uniform meshes, standard numerical approaches cannot yield effective solution approximations in layer regions. Stynes and Roos [23] introduced a hybrid technique for steady-state singular perturbation problems (SPPs) with continuous data, which is almost of order two on a Shishkin mesh. The hybrid numerical scheme for solving singularly perturbed boundary value problems with discontinuous convection coefficients introduced by Cen [2] achieves nearly second-order accuracy across the entire domain $[0, 1]$ when the perturbation parameter ϵ satisfies $\epsilon \leq N^{-1}$. Also, Mukherjeee and Natesan [15] proposed a hybrid difference scheme on Shishkin mesh in space and Euler-backward difference in time for singularlyperturbed parabolic problems with nonsmooth data. They achieve the convergence almost two in space and first order in time.

Many problems of significance like chemical flow in reactor theory [17], fluid dynamics [20] are modeled as singularly perturbed problems. In particular, these equations are used in time-dependent semiconductor models [13]. In a drift-diffusion model, jump discontinuities in the doping profile at a p-n junction create sharp interior layers in carrier densities and elec-

tric potential. These layers significantly impact the transient behavior of semiconductor devices, such as transistor switching and charge transport.

In this paper, we have used the Crank–Nicolson scheme [5] on time on a uniform mesh and in space a hybrid scheme on piece-wise uniform Shishkin mesh. The Crank–Nicolson method in time gives second order convergence in time in supremum norm. The hybrid scheme we propose combines a modified central difference scheme, the midpoint upwind scheme, and the upwind scheme. This blending offers a key advantage: it surpasses the limitations of the classical central difference scheme. The classical central difference scheme on a uniform mesh has been found to generate nonphysical oscillations in the discrete solution when ϵ is small, unless the mesh diameter is prohibitively small. On the other hand, the midpoint upwind approach shows second-order uniform convergence beyond the boundary layer region when used with a uniform mesh. The upwind method, though first order, does not have spurious oscillations. Furthermore, the Shishkin mesh uses piece-wise uniform meshes. Thus we choose a fine mesh within the layer regions and a coarse mesh in the outer regions. Leveraging this characteristic, we utilize the second-order classical central difference scheme in the interior layer regions which is second order accurate under certain conditions. The midpoint upwind scheme and upwind scheme are used suitably in the outer region to get a monotone method. This approach allows us to accomplish nearly second-order convergence in space for the proposed hybrid scheme. Thus, the overall method has almost second order parameter uniform convergence.

Contribution in context: The main contribution of the present paper is that the numerical scheme presented here achieves order two in time and almost order two in space in the supremum norm. Also, the method is parameter uniform. The orders of convergence obtained in space in this paper is an improvement of the result of Singh, Choudhary, and Kumar [22]. The method presented there had uniform convergence of order one up to a logarithmic factor in space and order two in time in the supremum norm.

The article is arranged as follows. In Section 2, we present a *priori* bounds on the analytical solution and its derivatives. The Shishkin mesh and the detailed construction of the suggested hybrid finite difference scheme are provided in Section 3. In Section 4, the truncation error of the scheme and

uniform error bound are given. In Section 5, we conduct a few numerical experiments to validate the theoretical findings and showcase the efficiency and accuracy of the proposed scheme. The primary findings are summarized in Section 6.

Notation: The norm used is the maximum norm given by

$$\|u\|_{\Omega} = \max_{(x,t)\in\Omega} |u(x,t)|.$$

In this study, C, C_1, or C_2 are used as a positive constants that are independent of the perturbation parameters ϵ, μ, and mesh size.

2 Analytical aspects of the solution

The analytical properties of the solution to the problem (1) are discussed in this section. The following provides details regarding the solution to (1), including the existence of a singularly perturbed solution, the minimum principle, stability bounds and a priori bounds.

Lemma 1. [12] The solution $u(x,t) \in C^0(\bar{\Omega}) \cap C^1(\Omega) \cap C^2(\Omega^- \cup \Omega^+)$ exists for (1).

Lemma 2. [12] If $u(x,t) \in C^0(\bar{\Omega}) \cap C^1(\Omega) \cap C^2(\Omega^- \cup \Omega^+)$ such that $u(x,t) \geq 0$, for all $(x,t) \in \Gamma_c, \mathcal{L}u(x,t) \leq 0$, for all $(x,t) \in (\Omega^- \cup \Omega^+)$ and $[u_x](d,t) \leq 0, t > 0$, then $u(x,t) \geq 0$ for all $(x,t) \in \bar{\Omega}$.

Lemma 3. [12] The bounds on $u(x,t)$ is given by

$$\|u\|_{\bar{\Omega}} \leq \|u\|_{\Gamma_c} + \frac{\|f\|_{\Omega^- \cup \Omega^+}}{\theta}.$$

where $\theta = \min\{\alpha_1/d, \alpha_2/(1-d)\}$.

The following bounds are parameter-explicit and satisfied by the derivatives of the solution to this problem (1).

Lemma 4. [3] The solution $u(x,t)$ and its derivatives for all $(x,t) \in \Omega^- \cup \Omega^+$ satisfy the following bounds for nonnegative integers k, m such that $1 \leq k + m \leq 3$:
If $\sqrt{\alpha}\mu \leq \sqrt{\rho\epsilon}$, then

$$\left\| \frac{\delta^{k+m} u}{\delta x^k \delta t^m} \right\|$$
$$\leq C\epsilon^{\frac{-k}{2}} \max\left\{ \|u\|_{\bar{\Omega}}, \sum_{i+2j=0}^{2} \epsilon^{i/2} \left\| \frac{\delta^{i+j} f}{\delta x^i \delta t^j} \right\|, \sum_{i=0}^{4} \left[\epsilon^{i/2} \left\| \frac{d^i p}{dt^i} \right\|_{\Gamma_l} + \left\| \frac{d^i q}{dx^i} \right\|_{\Gamma_b} + \left\| \frac{d^i r}{dt^i} \right\|_{\Gamma_r} \right] \right\}.$$

If $\sqrt{\alpha}\mu > \sqrt{\rho\epsilon}$, then

$$\left\| \frac{\delta^{k+m} u}{\delta x^k \delta t^m} \right\|$$
$$\leq C\left(\frac{\mu}{\epsilon}\right)^k \left(\frac{\mu^2}{\epsilon}\right)^m \max\left\{ \|u\|_{\bar{\Omega}}, \sum_{i+2j=0}^{2} \left(\frac{\epsilon}{\mu}\right)^i \left(\frac{\epsilon}{\mu^2}\right)^{j+1} \left\| \frac{\delta^{i+j} f}{\delta x^i \delta t^j} \right\|, \right.$$
$$\left. \sum_{i=0}^{4} \left[\left(\frac{\epsilon}{\mu}\right)^i \left\| \frac{d^i p}{dt^i} \right\|_{\Gamma_l} + \left(\frac{\epsilon}{\mu^2}\right)^{j+1} \left\| \frac{d^i q}{dx^i} \right\|_{\Gamma_b} + \left(\frac{\epsilon}{\mu^2}\right)^{j+1} \left\| \frac{d^i r}{dt^i} \right\|_{\Gamma_r} \right] \right\},$$

where C is a constant independent of ϵ and μ.

To obtain sharper bounds on the solution, the solution $u(x,t)$ of (1) is decomposed into regular and layers components as $u(x,t) = v(x,t) + w_l(x,t) + w_r(x,t)$. The regular component $v(x,t)$ satisfies the following equation:

$$\begin{aligned} &\mathcal{L}v(x,t) = f(x,t), && (x,t) \in \Omega^- \cup \Omega^+, \\ &v(x,0) = u(x,0), && \text{for all } x \in \Gamma^- \cup \Gamma^+, \end{aligned} \tag{2}$$

and $v(0,t), v(1,t), v(d-,t), v(d+,t)$ are chosen suitably for all t in $(0,T]$. Also, $v(d-,t) = \lim_{x \to d^-} v(x,t)$ and $v(d+,t) = \lim_{x \to d^+} v(x,t)$. The regular component $v(x,t)$ can further be decomposed as

$$v(x,t) = \begin{cases} v^-(x,t), & (x,t) \in \Omega^-, \\ v^+(x,t), & (x,t) \in \Omega^+, \end{cases}$$

where $v^-(x,t)$ and $v^+(x,t)$ are the left and right regular components, respectively.

The singular components $w_l(x,t)$ and $w_r(x,t)$ are the solutions of

$$\begin{aligned} &\mathcal{L}w_l(x,t) = 0, && (x,t) \in \Omega^- \cup \Omega^+, \\ &w_l(0,t) = u(0,t) - v(0,t) - w_r(0,t), && \text{for all } t \in (0,T], \\ &w_l(1,t) \text{ is chosen suitably} && \text{for all } t \in (0,T], \\ &w_l(x,0) = 0, && \text{for all } x \in \Gamma^- \cup \Gamma^+, \end{aligned} \tag{3}$$

and

$$\begin{aligned}
&\mathcal{L}w_r(x,t) = 0, \qquad (x,t) \in \Omega^- \cup \Omega^+,\\
&w_r(0,t) \text{ is chosen suitably} \qquad \text{for all } t \in (0,T],\\
&w_r(1,t) = u(1,t) - v(1,t) - w_l(1,t), \qquad \text{for all } t \in (0,T],\\
&w_r(x,0) = 0, \qquad \text{for all } x \in \Gamma^- \cup \Gamma^+,
\end{aligned} \tag{4}$$

respectively. The regular v and layer w_l, w_r components may be discontinuous at (d,t), for all $t \in (0,T]$ but their sum u is continuous at (d,t), for all $t \in (0,T]$. Furthermore, the singular components $w_l(x,t)$ and $w_r(x,t)$ are decomposed as

$$w_l(x,t) = \begin{cases} w_l^-(x,t), & (x,t) \in \Omega^-,\\ w_l^+(x,t), & (x,t) \in \Omega^+, \end{cases} \quad w_r(x,t) = \begin{cases} w_r^-(x,t), & (x,t) \in \Omega^-,\\ w_r^+(x,t), & (x,t) \in \Omega^+. \end{cases}$$

Hence, the unique solution $u(x,t)$ to (1) is written as

$$u(x,t) = \begin{cases} (v^- + w_l^- + w_r^-)(x,t), & (x,t) \in \Omega^-,\\ (v^- + w_l^- + w_r^-)(d-,t) & \\ = (v^+ + w_l^+ + w_r^+)(d+,t), & (x,t) = (d,t), \text{ for all } t \in (0,T],\\ (v^+ + w_l^+ + w_r^+)(x,t), & (x,t) \in \Omega^+. \end{cases}$$

Lemma 5. The bounds on the components $v(x,t), w_l(x,t), w_r(x,t)$ and their derivatives are given as follows for $\sqrt{\alpha}\mu \le \sqrt{\rho\epsilon}$ and $0 \le i + 2j \le 4$:

$$\left\| \frac{\delta^{i+j} v}{\delta x^i \delta t^j} \right\|_{\Omega^- \cup \Omega^+} \le C(1 + \epsilon^{(3-i)/2}),$$

$$\left\| \frac{\delta^{i+j} w_l}{\delta x^i \delta t^j} \right\|_{\Omega^- \cup \Omega^+} \le C\epsilon^{-i/2} \begin{cases} e^{-\theta_2 x}, & (x,t) \in \Omega^-,\\ e^{-\theta_1 (x-d)}, & (x,t) \in \Omega^+, \end{cases}$$

$$\left\| \frac{\delta^{i+j} w_r}{\delta x^i \delta t^j} \right\|_{\Omega^- \cup \Omega^+} \le C\epsilon^{-i/2} \begin{cases} e^{-\theta_1 (d-x)}, & (x,t) \in \Omega^-,\\ e^{-\theta_2 (1-x)}, & (x,t) \in \Omega^+, \end{cases}$$

where

$$\theta_1 = \frac{\sqrt{\rho\alpha}}{\sqrt{\epsilon}}, \qquad \theta_2 = \frac{\sqrt{\rho\alpha}}{2\sqrt{\epsilon}}, \tag{5}$$

and C is a constant independent of ϵ and μ.

Proof. The proof follows by following the approach given in Gracia and O'Riordan [9]. □

Lemma 6. The bounds on the components $v(x,t), w_l(x,t), w_r(x,t)$ and their derivatives are given as follows for $\sqrt{\alpha}\mu > \sqrt{\rho\epsilon}$ and $0 \le i + 2j \le 4$,:

$$\left\| \frac{\delta^{i+j} v}{\delta x^i \delta t^j} \right\|_{\Omega^- \cup \Omega^+} \le C(1 + (\epsilon/\mu)^{(3-i)}),$$

$$\left\| \frac{\delta^{i+j} w_l}{\delta x^i \delta t^j} \right\|_{\Omega^- \cup \Omega^+} \le C \left(\frac{\mu}{\epsilon} \right)^i \begin{cases} e^{-\theta_2 x}, & (x,t) \in \Omega^-, \\ e^{-\theta_1 (x-d)}, & (x,t) \in \Omega^+, \end{cases}$$

$$\left\| \frac{\delta^{i+j} w_r}{\delta x^i \delta t^j} \right\|_{\Omega^- \cup \Omega^+} \le C \left(\frac{1}{\mu} \right)^i \begin{cases} e^{-\theta_1 (d-x)}, & (x,t) \in \Omega^-, \\ e^{-\theta_2 (1-x)}, & (x,t) \in \Omega^+, \end{cases}$$

where

$$\theta_1 = \frac{\alpha\mu}{\epsilon}, \qquad \theta_2 = \frac{\rho}{2\mu}, \tag{6}$$

and C is a constant independent of ϵ and μ.

Proof. For the proof, refer to the idea given in Gracia and O'Riordan [9]. □

3 Discretization

We will first discretize (1) in time using the Crank–Nicolson method [5]. This results in a semidiscrete system of equations. We then discretize in space.

3.1 Discretization in time

We begin by demonstrating the semidiscretization of the problem in the temporal direction on a uniform mesh using the Crank–Nicolson method [5]. For a fixed time T, the interval $[0,T]$ is partitioned uniformly as $\Lambda^M = \{t_j = j\Delta t : j = 0, 1, \ldots, M, \Delta t = \frac{T}{M}\}$. The semidiscretization yields the following system of linear ordinary differential equations:

$$\begin{aligned} &\epsilon U_{xx}^{j+\frac{1}{2}}(x) + \mu a^{j+\frac{1}{2}}(x) U_x^{j+\frac{1}{2}}(x) - b^{j+\frac{1}{2}}(x) U^{j+\frac{1}{2}}(x) \\ &= f^{j+\frac{1}{2}}(x) + \frac{U^{j+1}(x) - U^j(x)}{\Delta t}, \end{aligned}$$

$$x \in (\Omega^- \cup \Omega^+),\ 0 \le j \le M-1,$$
$$U^{j+1}(0) = u(0, t_{j+1}), \qquad U^{j+1}(1) = u(1, t_{j+1}), \quad 0 \le j \le M-1,$$
$$U^0(x) = u(x, 0), \qquad x \in \Gamma_b,$$

where $U^{j+1}(x)$ is the approximation of $u(x, t_{j+1})$ of (1) at $(j+1)$th time level and $p^{j+\frac{1}{2}} = \dfrac{p^{j+1}(x) + p^j(x)}{2}$. Upon simplification, we have

$$\left.\begin{aligned} &\widetilde{\mathcal{L}} U^{j+1}(x) = g(x, t_{j+1}), \ x \in (\Gamma^- \cup \Gamma^+), \qquad 0 \le j \le M-1, \\ &U^{j+1}(0) = u(0, t_{j+1}), \quad 0 \le j \le M-1, \\ &U^{j+1}(1) = u(1, t_{j+1}), \quad 0 \le j \le M-1, \\ &U^0(x) = u(x, 0), \qquad x \in \Gamma, \end{aligned}\right\} \tag{7}$$

where the operator $\widetilde{\mathcal{L}}$ is defined as

$$\widetilde{\mathcal{L}} \equiv \epsilon \frac{d^2}{dx^2} + \mu a^{j+\frac{1}{2}} \frac{d}{dx} - c^{j+\frac{1}{2}} I,$$

and

$$g(x, t_{j+1}) = 2f^{j+\frac{1}{2}}(x) - \epsilon U^j_{xx}(x) - \mu a^{j+\frac{1}{2}}(x) U^j_x(x) + d^{j+\frac{1}{2}}(x) U^j(x),$$
$$c^{j+\frac{1}{2}}(x) = b^{j+\frac{1}{2}}(x) + \frac{2}{\Delta t},$$
$$d^{j+\frac{1}{2}}(x) = b^{j+\frac{1}{2}}(x) - \frac{2}{\Delta t}.$$

The error in temporal semidiscretization is defined by $e^{j+1} = u(x, t_{j+1}) - \hat{U}^{j+1}(x)$, where $u(x, t_{j+1})$ is the solution of (1). Moreover, $\hat{U}^{j+1}(x)$ is the solution of semidiscrete scheme (7), when $u(x, t_j)$ is taken instead of U^j to find solution at (x, t_{j+1}).

Theorem 1. The local truncation error $T_{j+1} = \widetilde{\mathcal{L}}(e^{j+1})$ satisfies

$$\|T_{j+1}\| \le C(\Delta t)^3, \qquad 0 \le j \le M-1.$$

Proof. The Taylor's approximation gives

$$u(x, t_{j+1}) = u(x, t_{j+\frac{1}{2}}) + \frac{1}{2}\Delta t u_t(x, t_{j+\frac{1}{2}}) + \frac{1}{8}\Delta t^2 u_{tt}(x, t_{j+\frac{1}{2}}) + \mathcal{O}(\Delta t)^3,$$
$$u(x, t_j) = u(x, t_{j+\frac{1}{2}}) - \frac{1}{2}\Delta t u_t(x, t_{j+\frac{1}{2}}) + \frac{1}{8}\Delta t^2 u_{tt}(x, t_{j+\frac{1}{2}}) + \mathcal{O}(\Delta t)^3.$$

Using these equations, we have

$$\begin{aligned}\frac{u(x,t_{j+1})-u(x,t_j)}{\Delta t} =& u_t(x,t_{j+\frac{1}{2}})+\mathcal{O}(\Delta t)^2\\ =& \epsilon u_{xx}(x,t_{j+\frac{1}{2}})+\mu a(x,t_{j+\frac{1}{2}})u_x(x,t_{j+\frac{1}{2}})\\ & -b(x,t_{j+\frac{1}{2}})u(x,t_{j+\frac{1}{2}})-f(x,t_{j+\frac{1}{2}})+\mathcal{O}(\Delta t)^2.\end{aligned}$$

Further simplification gives

$$\begin{aligned}\widetilde{\mathcal{L}}u_x^j &= \epsilon u_{xx}(x,t_{j+1})+\mu a^{j+\frac{1}{2}}u_x(x,t_{j+1})-c^{j+\frac{1}{2}}u(x,t_{j+1})\\ &= g(x,t_{j+1})+\mathcal{O}(\Delta t)^3.\end{aligned}$$

Hence the local truncation error satisfies

$$\begin{aligned}\widetilde{\mathcal{L}}(e^{j+1}) &= \mathcal{O}(\Delta t)^3,\\ e^{j+1}(0) &= e^{j+1}(0)=0.\end{aligned}$$

We arrive at the result using Lemma 3. [14]) $e^{j+1}(x)\leq \mathcal{O}(\Delta t)^3$. □

Theorem 2. The global error $E^j = u(x,t_j)-U^j(x)$ is estimated as

$$\|E^j\|\leq C(\Delta t)^2, \qquad 0\leq j\leq M-1,$$

where $U^j(x)$ is the solution of (7).

Proof. For proof, refer to [22]. $\|\prod_{k=0}^{i-1}R_{j-k}\|\leq C$, for all $i=1,2,3,...j, 1\leq j\leq M$. □

3.2 Discretization in space

We understand from the relevant literature on SPPs that a uniform mesh is insufficient for achieving uniform convergence of (1) due to the existence of layer regions. The differential equation (1) exhibits strong interior layers including boundary layers near the boundaries. To resolve these layers, we have discretized this problem by using a piece-wise uniform Shishkin mesh.

Let the interior points of the spatial mesh be denoted by $\Gamma^N=\{x_i: 1\leq i\leq \frac{N}{2}-1\}\cup\{x_i: \frac{N}{2}+1\leq i\leq N-1\}$. The equation $\bar{\Gamma}^N=\{x_i\}_0^N\cup\{d\}$ denotes

the mesh points with $x_0 = 0, x_N = 1$ and the point of discontinuity at point $x_{\frac{N}{2}} = d$. We also introduce the notation $\Gamma^{N-} = \{x_i\}_0^{\frac{N}{2}-1}, \Gamma^{N+} = \{x_i\}_{\frac{N}{2}+1}^{N-1}$, $\Omega^{N-} = \Gamma^{N-} \times \Lambda^M$, $\Omega^{N+} = \Gamma^{N+} \times \Lambda^M$, $\Omega^{N,M} = \Gamma^N \times \Lambda^M$, $\bar{\Omega}^{N,M} = \bar{\Gamma}^N \times \Lambda^M$, $\Gamma_c^N = \bar{\Omega}^{N,M} \cap \Gamma_c$, $\Gamma_l^N = \Gamma_c^N \cap \Gamma_l$, $\Gamma_b^N = \Gamma_c^N \cap \Gamma_b$ and $\Gamma_r^N = \Gamma_c^N \cap \Gamma_r$. The domain $[0,1]$ is subdivided into six subintervals as

$$\bar{\Gamma} = [0,\tau_1] \cup [\tau_1, d-\tau_2] \cup [d-\tau_2, d] \cup [d, d+\tau_3] \cup [d+\tau_3, 1-\tau_4] \cup [1-\tau_4, 1] = \bigcup_{i=1}^{6} I_i.$$

The transition points in $\bar{\Gamma}$ are

$$\tau_1 = \min\left\{\frac{d}{4}, \frac{2}{\theta_2}\ln N\right\}, \qquad \tau_2 = \min\left\{\frac{d}{4}, \frac{2}{\theta_1}\ln N\right\},$$

$$\tau_3 = \min\left\{\frac{1-d}{4}, \frac{2}{\theta_1}\ln N\right\}, \qquad \tau_4 = \min\left\{\frac{1-d}{4}, \frac{2}{\theta_2}\ln N\right\}.$$

On the subintervals $[0,\tau_1], [d-\tau_2, d]$, $[d, d+\tau_3]$, and $[1-\tau_4, 1]$, a uniform mesh of $\left(\frac{N}{8}+1\right)$ mesh points and on $[\tau_1, d-\tau_2]$ and $[d+\tau_3, 1-\tau_4]$ a uniform mesh of $\left(\frac{N}{4}+1\right)$ mesh points are taken. The mesh points are given by

$$x_i = \begin{cases} \dfrac{8i\tau_1}{N}, & 0 \le i \le \frac{N}{8}, \\ \tau_1 + \dfrac{4(d-\tau_1-\tau_2)}{N}\left(i - \dfrac{N}{8}\right), & \frac{N}{8} \le i \le \frac{3N}{8}, \\ d - \tau_2 + \dfrac{8\tau_2}{N}\left(i - \dfrac{3N}{8}\right), & \frac{3N}{8} \le i \le \frac{N}{2}, \\ d + \dfrac{8\tau_3}{N}\left(i - \dfrac{N}{2}\right), & \frac{N}{2} \le i \le \frac{5N}{8}, \\ d + \tau_3 + \dfrac{4(1-d-\tau_3-\tau_4)}{N}\left(i - \dfrac{5N}{8}\right), & \frac{5N}{8} \le i \le \frac{7N}{8}, \\ 1 - \tau_4 + \dfrac{8\tau_4}{N}\left(i - \dfrac{7N}{8}\right), & \frac{7N}{8} \le i \le N. \end{cases}$$

Let $h_i = x_i - x_{i-1}$ denote the step size and let $h_1 = \frac{8\tau_1}{N}, h_2 = \frac{4(d-\tau_1-\tau_2)}{N}, h_3 = \frac{8\tau_2}{N}, h_4 = \frac{8\tau_3}{N}, h_5 = \frac{4(1-d-\tau_3-\tau_4)}{N}, h_6 = \frac{8\tau_4}{N}$ denote the step sizes in the interval I_1, I_2, I_3, I_4, I_5, and I_6, respectively.

Assumption 1. Throughout the paper, we shall also assume that $\epsilon \le N^{-1}$ as is generally the case for discretization of convection-dominated problems.

We have presented a hybrid difference approach for the discretization of the problem (1) in the spatial variable. On the above piece-wise uniform Shishkin mesh, this hybrid difference scheme comprises of an upwind, a central difference and a midpoint upwind scheme. Here, we introduce the hybrid difference method for the ratio $\sqrt{\alpha}\mu \leq \sqrt{\rho\epsilon}$:

$$\mathcal{L}_h^{N,M}U^{j+1}(x_i) = \begin{cases} \mathcal{L}_c^{N,M}U^{j+1}(x_i), & x_i \in (0,1)\backslash\{\tau_1, d-\tau_2, d, d+\tau_3, 1-\tau_4\}, \\ & \text{with } \mu\|a\|h_k < 2\epsilon,\ 1 \leq k \leq 6, \\ \mathcal{L}_t^{N,M}U^{j+1}(x_{\frac{N}{2}}), & x_i = d \text{ with } h_k(\|b\| + \frac{2}{\Delta t}) < 2\mu\alpha,\ k = 3,4. \end{cases}$$

For the ratio $\sqrt{\alpha}\mu > \sqrt{\rho\epsilon}$, we use

$$\mathcal{L}_h^{N,M}U^{j+1}(x_i) = \begin{cases} \mathcal{L}_m^{N,M}U^{j+1}(x_i), & x_i \in (0,\tau_1) \cup (\tau_1, d-\tau_2), \\ & \text{with } (\|b\| + \frac{2}{\Delta t})h_k \leq 2\alpha\mu,\ k = 1,2, \\ \mathcal{L}_u^{N,M}U^{j+1}(x_i), & x_i \in (0,\tau_1) \cup (\tau_1, d-\tau_2), \\ & \text{with } (\|b\| + \frac{2}{\Delta t})h_k > 2\alpha\mu,\ k = 1,2, \\ \mathcal{L}_c^{N,M}U^{j+1}(x_i), & x_i \in (d-\tau_2, d) \cup (d, d+\tau_3), \\ & \text{with } \mu\|a\|h_k < 2\epsilon,\ k = 3,4, \\ \mathcal{L}_m^{N,M}U^{j+1}(x_i), & x_i \in (d+\tau_3, 1-\tau_4) \cup (1-\tau_4, 1), \\ & \text{with } (\|b\| + \frac{2}{\Delta t})h_k \leq 2\alpha\mu,\ k = 5,6, \\ \mathcal{L}_u^{N,M}U^{j+1}(x_i), & x_i \in (d+\tau_3, 1-\tau_4) \cup (1-\tau_4, 1), \\ & \text{with } (\|b\| + \frac{2}{\Delta t})h_k > 2\alpha\mu,\ k = 5,6, \\ \mathcal{L}_t^{N,M}U^{j+1}(x_{\frac{N}{2}}), & x_i = d \text{ with } h_k(\|b\| + \frac{2}{\Delta t}) < 2\mu\alpha,\ k = 3,4. \end{cases}$$

At the transition points for both case

$$\mathcal{L}_h^{N,M}U^{j+1}(x_i) = \begin{cases} \mathcal{L}_m^{N,M}U^{j+1}(x_i),\ (\|b\| + \frac{2}{\Delta t})h_k \leq 2\alpha\mu, & k = 2,3,5,6, \\ \mathcal{L}_u^{N,M}U^{j+1}(x_i), & \text{otherwise.} \end{cases}$$

The fully discretize scheme is given by

$$\mathcal{L}_h^{N,M}U^{j+1}(x_i) \equiv r_i^- U^{j+1}(x_{i-1}) + r_i^c U^{j+1}(x_i) + r_i^+ U^{j+1}(x_{i+1}) = g^{j+1}(x_i). \tag{8}$$

At the point of discontinuity $x_{N/2} = d$, we have used five point second order difference scheme

$$\mathcal{L}_t^{N,M} U^{j+1}(x_{\frac{N}{2}}) = \frac{-U^{j+1}(x_{N/2+2}) + 4U^{j+1}(x_{\frac{N}{2}+1}) - 3U^{j+1}(x_{\frac{N}{2}})}{2h_4} - \frac{U^{j+1}(x_{\frac{N}{2}-2}) - 4U^{j+1}(x_{\frac{N}{2}-1}) + 3U^{j+1}(x_{\frac{N}{2}})}{2h_3} = 0. \tag{9}$$

Associated with each of these finite difference operators, we have the following finite difference scheme:

$$\mathcal{L}_c^{N,M} U^{j+1}(x_i) \equiv \epsilon\delta^2 U^{j+1}(x_i) + \mu a^{j+\frac{1}{2}}(x_i) D^0 U^{j+1}(x_i) - b^{j+\frac{1}{2}}(x_i) U^{j+1}(x_i),$$
$$\mathcal{L}_m^{N,M} U^{j+1}(x_i) \equiv \epsilon\delta^2 U^{j+1}(x_i) + \mu \bar{a}^{j+\frac{1}{2}}(x_i) D^* U^{j+1}(x_i) - \bar{b}^{j+\frac{1}{2}}(x_i) \bar{U}^{j+1}(x_i),$$
$$\mathcal{L}_u^{N,M} U^{j+1}(x_i) \equiv \epsilon\delta^2 U^{j+1}(x_i) + \mu a^{j+\frac{1}{2}}(x_i) D^* U^{j+1}(x_i) - b^{j+\frac{1}{2}}(x_i) U^{j+1}(x_i),$$

and

$$g^{j+1}(x_i) = \begin{cases} 2f^{j+\frac{1}{2}}(x_i) - \epsilon\delta_x^2 U^j(x_i) - \mu a^{j+\frac{1}{2}}(x_i) D_x^0 U^j(x_i) + d^{j+\frac{1}{2}}(x) U^j(x_i), & \mathcal{L}_h^{N,M} = \mathcal{L}_c^{N,M}, \\ 2f^{j+\frac{1}{2}}(x_i) - \epsilon\delta_x^2 U^j(x_i) - \mu \bar{a}^{j+\frac{1}{2}}(x_i) D_x^* U^j(x_i) + \bar{d}^{j+\frac{1}{2}}(x) U^j(x_i), & \mathcal{L}_h^{N,M} = \mathcal{L}_m^{N,M}, \\ 2f^{j+\frac{1}{2}}(x_i) - \epsilon\delta_x^2 U^j(x_i) - \mu a^{j+\frac{1}{2}}(x_i) D_x^* U^j(x_i) + d^{j+\frac{1}{2}}(x) U^j(x_i), & \mathcal{L}_h^{N,M} = \mathcal{L}_u^{N,M}, \end{cases}$$

where

$$D^+ U^{j+1}(x_i) = \frac{U^{j+1}(x_{i+1}) - U^{j+1}(x_i)}{x_{i+1} - x_i}, \qquad D^- U^{j+1}(x_i) = \frac{U^{j+1}(x_i) - U^{j+1}(x_{i-1})}{x_i - x_{i-1}},$$

$$D^0 U^{j+1}(x_i) = \frac{U^{j+1}(x_{i+1}) - U^{j+1}(x_{i-1})}{h_{i+1} + h_i},$$

$$\delta^2 U^{j+1}(x_i) = \frac{2(D^+ U^{j+1}(x_i) - D^- U^{j+1}(x_i))}{x_{i+1} - x_{i-1}},$$

$$D^* U^{j+1}(x_i) = \begin{cases} D^- U^{j+1}(x_i), \, i < \frac{N}{2}, \\ D^+ U^{j+1}(x_i), \, i > \frac{N}{2}, \end{cases}$$

$\hbar_i = \frac{h_i + h_{i+1}}{2}$, $\bar{w}^{j+1}(x_i) = \dfrac{w^{j+1}(x_i) + w^{j+1}(x_{i-1})}{2}$ in Ω^{N-},
$\bar{w}^{j+1}(x_i) = \dfrac{w^{j+1}(x_i) + w^{j+1}(x_{i+1})}{2}$ in Ω^{N+}, $\bar{a}^{j+\frac{1}{2}}(x_i) = \dfrac{\bar{a}^{j+1}(x_i) + \bar{a}^j(x_i)}{2}$,
$\bar{b}^{j+\frac{1}{2}}(x_i) = \dfrac{\bar{b}^{j+1}(x_i) + \bar{b}^j(x_i)}{2}$ and $\bar{d}^{j+\frac{1}{2}}(x_i) = \dfrac{\bar{d}^{j+1}(x_i) + \bar{d}^j(x_i)}{2}$.

The matrix associated with the above equations (8) is not an M-matrix. We transform (9) to establish the monotonicity property by estimating $U^{j+1}_{\frac{N}{2}-2}$ and $U^{j+1}_{\frac{N}{2}+2}$ for $\mathcal{L}^{N,M}_t U^{j+1}(x_i)$.

For $\sqrt{\alpha}\mu \le \sqrt{\rho}\epsilon$ and $\sqrt{\alpha}\mu > \sqrt{\rho}\epsilon$, from the operator $\mathcal{L}^{N,M}_c$, we get

$$
\begin{aligned}
&U^{j+1}_{N/2-2}\\
&= \frac{2h_3^2}{2\epsilon - \mu h_3 a^{j+\frac{1}{2}}_{N/2-1}}\Bigg\{g^{j+1}_{N/2-1} - U^{j+1}_{N/2-1}\left(\frac{-2\epsilon - h_3^2(b^{j+\frac{1}{2}}_{N/2-1} + \frac{2}{\Delta t})}{h_3^2}\right)\\
&\quad - U^{j+1}_{N/2}\left(\frac{2\epsilon + \mu h_3 a^{j+\frac{1}{2}}_{N/2-1}}{2h_3^2}\right)\Bigg\}\\
&U^{j+1}_{N/2+2}\\
&= \frac{2h_4^2}{2\epsilon + \mu h_4 a^{j+\frac{1}{2}}_{N/2+1}}\Bigg\{g^{j+1}_{N/2+1} - U^{j+1}_{N/2+1}\left(\frac{-2\epsilon - h_4^2(b^{j+\frac{1}{2}}_{N/2+1} + \frac{2}{\Delta t})}{h_4^2}\right)\\
&\quad - U^{j+1}_{N/2}\left(\frac{2\epsilon - \mu h_4 a^{j+\frac{1}{2}}_{N/2+1}}{2h_4^2}\right)\Bigg\}.
\end{aligned}
$$

Inserting the above expression in $\mathcal{L}^{N,M}_t U^{j+1}(x_{N/2})$, we get a tridiagonal system of equations of the following form:

$$
\begin{aligned}
\mathcal{L}^{N,M}_t U^{j+1}_{N/2} =& \left(\frac{2\epsilon - h_4\mu a^{j+\frac{1}{2}}_{N/2+1}}{2\epsilon + h_4\mu a^{j+\frac{1}{2}}_{N/2+1}}h_3 - 3(h_3+h_4) + \frac{2\epsilon + h_3\mu a^{j+\frac{1}{2}}_{N/2-1}}{2\epsilon - h_3\mu a^{j+\frac{1}{2}}_{N/2-1}}h_4\right)U^{j+1}_{N/2}\\
&+ \left(\frac{-4\epsilon - 2h_4^2(b^{j+\frac{1}{2}}_{N/2+1} + \frac{2}{\Delta t})}{2\epsilon + h_4\mu a^{j+\frac{1}{2}}_{N/2+1}} + 4\right)h_3 U^{j+1}_{N/2+1}\\
&+ \left(\frac{-4\epsilon - 2h_3^2(b^{j+\frac{1}{2}}_{N/2-1} + \frac{2}{\Delta t})}{2\epsilon - h_3\mu a^{j+\frac{1}{2}}_{N/2-1}} + 4\right)h_4 U^{j+1}_{N/2-1}\\
=& \frac{2h_3^2 h_4 g^{j+1}_{N/2-1}}{2\epsilon - h_3\mu a^{j+\frac{1}{2}}_{N/2-1}} + \frac{2h_4^2 h_3 g^{j+1}_{N/2+1}}{2\epsilon + h_4\mu a^{j+\frac{1}{2}}_{N/2+1}}.
\end{aligned}
$$

The element of the system matrix $\mathcal{L}^{N,M}_h$ is as follows:
In $\Omega^- \cup \Omega^+$, if $\mathcal{L}^{N,M}_h = \mathcal{L}^{N,M}_c$, then

$$r_i^- = \frac{\epsilon}{h_i\hbar_i} - \frac{\mu a^{j+\frac{1}{2}}(x_i)}{2\hbar_i}, \qquad r_i^+ = \frac{\epsilon}{h_{i+1}\hbar_i} + \frac{\mu a^{j+\frac{1}{2}}(x_i)}{2\hbar_i},$$

$$r_i^c = -r_i^- - r_i^+ - \left(b^{j+\frac{1}{2}}(x_i) + \frac{2}{\Delta t}\right).$$

In Ω^-, if $\mathcal{L}_h^{N,M} = \mathcal{L}_m^{N,M}$, then

$$r_i^- = \frac{\epsilon}{h_i\hbar_i} - \frac{\mu\bar{a}^{j+\frac{1}{2}}(x_i)}{h_i} - \left(\frac{\bar{b}^{j+\frac{1}{2}}(x_i)}{2} + \frac{1}{\Delta t}\right), \qquad r_i^+ = \frac{\epsilon}{h_{i+1}\hbar_i},$$

$$r_i^c = -r_i^- - r_i^+ - \frac{1}{2}\left(\bar{b}^{j+\frac{1}{2}}(x_i) + \frac{2}{\Delta t}\right).$$

In Ω^+, if $\mathcal{L}_h^{N,M} = \mathcal{L}_m^{N,M}$, then

$$r_i^- = \frac{\epsilon}{h_i\hbar_i}, \qquad r_i^+ = \frac{\epsilon}{h_{i+1}\hbar_i} - \frac{\mu\bar{a}^{j+\frac{1}{2}}(x_i)}{h_{i+1}} - \left(\frac{\bar{b}^{j+\frac{1}{2}}(x_i)}{2} + \frac{1}{\Delta t}\right),$$

$$r_i^c = -r_i^- - r_i^+ - \frac{1}{2}\left(\bar{b}^{j+\frac{1}{2}}(x_i) + \frac{2}{\Delta t}\right).$$

In Ω^-, if $\mathcal{L}_h^{N,M} = \mathcal{L}_u^{N,M}$, then

$$r_i^- = \frac{\epsilon}{h_i\hbar_i} - \frac{\mu a^{j+\frac{1}{2}}(x_i)}{h_i}, \qquad r_i^+ = \frac{\epsilon}{h_{i+1}\hbar_i},$$

$$r_i^c = -r_i^- - r_i^+ - \left(b^{j+\frac{1}{2}}(x_i) + \frac{2}{\Delta t}\right).$$

In Ω^+, if $\mathcal{L}_h^{N,M} = \mathcal{L}_u^{N,M}$, then

$$r_i^- = \frac{\epsilon}{h_i\hbar_i}, \qquad r_i^+ = \frac{\epsilon}{h_{i+1}\hbar_i} + \frac{\mu a^{j+\frac{1}{2}}(x_i)}{h_{i+1}},$$

$$r_i^c = -r_i^- - r_i^+ - \left(b^{j+\frac{1}{2}}(x_i) + \frac{2}{\Delta t}\right).$$

At $x_{N/2} = d$,

we have

$$r_{N/2}^- = \left(\frac{-4\epsilon - 2h_3^2(b_{N/2-1}^{j+\frac{1}{2}} + \frac{2}{\Delta t})}{2\epsilon - h_3\mu a_{N/2-1}^{j+\frac{1}{2}}} + 4\right)h_4,$$

$$r_{N/2}^+ = \left(\frac{-4\epsilon - 2h_4^2(b_{N/2+1}^{j+\frac{1}{2}} + \frac{2}{\Delta t})}{2\epsilon + h_4\mu a_{N/2+1}^{j+\frac{1}{2}}} + 4\right)h_3,$$

$$r^c_{N/2} = \left(\frac{2\epsilon - h_4 \mu a^{j+\frac{1}{2}}_{N/2+1}}{2\epsilon + h_4 \mu a^{j+\frac{1}{2}}_{N/2+1}} h_3 - 3(h_3 + h_4) + \frac{2\epsilon + h_3 \mu a^{j+\frac{1}{2}}_{N/2-1}}{2\epsilon - h_3 \mu a^{j+\frac{1}{2}}_{N/2-1}} h_4 \right).$$

To guarantee that the operator $\mathcal{L}^{N,M}_h$ is monotone, we impose the following assumption:

$$N(\ln N)^{-1} > 16 \max \left\{ \frac{\|b\|}{\alpha}, \frac{(\|b\| + \frac{2}{\Delta t})}{\alpha \rho} \right\}. \tag{10}$$

To find the error estimates for the scheme (8) defined above, we first decompose the discrete solution $U(x_i, t_{j+1})$ into the discrete regular and discrete singular components.

Let

$$U^{j+1}(x_i) = V^{j+1}(x_i) + W^{j+1}_l(x_i) + W^{j+1}_r(x_i).$$

The discrete regular components is

$$V^{j+1}(x_i) = \begin{cases} V^{-(j+1)}(x_i), & (x_i, t_{j+1}) \in \Omega^{N-}, \\ V^{+(j+1)}(x_i), & (x_i, t_{j+1}) \in \Omega^{N+}, \end{cases}$$

where $V^{-(j+1)}(x_i)$ and $V^{+(j+1)}(x_i)$ approximate $v^-(x_i, t_{j+1})$ and $v^+(x_i, t_{j+1})$, respectively. They satisfy the following equations:

$$\left.\begin{array}{ll} \mathcal{L}^{N,M}_h V^{-(j+1)}(x_i) = g(x_i, t_{j+1}), & \text{for all } (x_i, t_{j+1}) \in \Omega^{N-}, \\ V^{-(j+1)}(x_0) = v^-(0, t_{j+1}), V^{-(j+1)}(x_{\frac{N}{2}}) = v^-(d-, t_{j+1}), & \\ \mathcal{L}^{N,M}_h V^{+(j+1)}(x_i) = g(x_i, t_{j+1}), & \text{for all } (x_i, t_{j+1}) \in \Omega^{N+}, \\ V^{+(j+1)}(x_{\frac{N}{2}}) = v^+(d+, t_{j+1}), V^{+(j+1)}(x_N) = v^-(1, t_{j+1}). & \end{array}\right\} \tag{11}$$

The discrete singular components $W^{j+1}_l(x_i)$ and $W^{j+1}_r(x_i)$ are also decomposed as

$$W^{j+1}_l(x_i) = \begin{cases} W^{-(j+1)}_l(x_i), & (x_i, t_{j+1}) \in \Omega^{N-}, \\ W^{+(j+1)}_l(x_i), & (x_i, t_{j+1}) \in \Omega^{N+}, \end{cases}$$

$$W^{j+1}_r(x_i) = \begin{cases} W^{-(j+1)}_r(x_i), & (x_i, t_{j+1}) \in \Omega^{N-}, \\ W^{+(j+1)}_r(x_i), & (x_i, t_{j+1}) \in \Omega^{N+}, \end{cases}$$

where $W^{-(j+1)}_l(x_i), W^{+(j+1)}_l(x_i)$ approximates the layer components $w^-_l(x_i, t_{j+1})$ and $w^+_l(x_i, t_{j+1})$ and $W^{-(j+1)}_r(x_i), W^{+(j+1)}_r(x_i)$ approximates the layer components $w^-_r(x_i, t_{j+1})$ and $w^+_r(x_i, t_{j+1})$, respectively. These components satisfy the following equations:

$$\left.\begin{array}{l} \mathcal{L}_h^{N,M} W_l^{-(j+1)}(x_i) = 0, \qquad \text{for all } (x_i, t_{j+1}) \in \Omega^{N-}, \\ W_l^{-(j+1)}(x_0) = w_l^-(0, t_{j+1}), W_l^{-(j+1)}(x_{\frac{N}{2}}) = w_l^-(d, t_{j+1}), \\ \mathcal{L}_h^{N,M} W_l^{+(j+1)}(x_i) = 0, \qquad \text{for all } (x_i, t_{j+1}) \in \Omega^{N+}, \\ W_l^{+(j+1)}(x_{\frac{N}{2}}) = w_l^+(d, t_{j+1}), W_l^{+(j+1)}(x_N) = 0. \end{array}\right\} \tag{12}$$

$$\left.\begin{array}{l} \mathcal{L}_h^{N,M} W_r^{-(j+1)}(x_i) = 0, \qquad \text{for all } (x_i, t_{j+1}) \in \Omega^{N-}, \\ W_r^{-(j+1)}(x_0) = 0, W_r^{-(j+1)}(x_{\frac{N}{2}}) = w_r^-(d, t_{j+1}), \\ \mathcal{L}_h^{N,M} W_r^{+(j+1)}(x_i) = 0, \qquad \text{for all } (x_i, t_{j+1}) \in \Omega^{N+}, \\ W_r^{+(j+1)}(x_{\frac{N}{2}}) = w_r^+(d+, t_{j+1}), W_r^{+(j+1)}(x_N) = w_r(1, t_{j+1}). \end{array}\right\} \tag{13}$$

Hence, the complete discrete solution $U^{j+1}(x_i)$ is defined by

$$U^{j+1}(x_i) = \begin{cases} (V^{-(j+1)} + W_l^{-(j+1)} + W_r^{-(j+1)})(x_i), & (x_i, t_{j+1}) \in \Omega^{N-}, \\ (V^{-(j+1)} + W_l^{-(j+1)} + W_r^{-(j+1)})(d-) & \\ = (V^{+(j+1)} + W_l^{+(j+1)} + W_r^{+(j+1)})(d+), & (x_i, t_{j+1}) = (d, t_{j+1}), \\ (V^{+(j+1)} + W_l^{+(j+1)} + W_r^{+(j+1)})(x_i), & (x_i, t_{j+1}) \in \Omega^{N+}. \end{cases}$$

Lemma 7. Suppose that a mesh function $Y(x_i, t_j)$ satisfies $Y(x_i, t_j) \geq 0$, for all $(x_i, t_j) \in \Gamma_c^N$ and $\mathcal{L}_t^{N,M} Y(x_{N/2}) \leq 0$ for all $j = 0, \ldots, M$. If $\mathcal{L}_h^{N,M} Y(x_i, t_j) \leq 0$ for all $(x_i, t_j) \in \Omega^{N-} \cup \Omega^{N+}$, then $Y(x_i, t_j) \geq 0$, for all $(x_i, t_j) \in \bar{\Omega}^{N,M}$.

Proof. The condition (10) ensures that the matrix generated by $\mathcal{L}_h^{N,M}$ satisfies $r_i^- > 0$, $r_i^+ > 0$ and $r_i^- + r_i^c + r_i^+ < 0$ in layer regions. For the case when $\sqrt{\alpha}\mu \leq \sqrt{\rho\epsilon}$, the operator $\mathcal{L}_c^{N,M}$ is used in the region $(\tau_1, d - \tau_2)$ and $(d + \tau_3, 1 - \tau_4)$ except transition points with $\mu h_k \|a\| \leq 2\epsilon$, where $k = 2, 5$ to guarantee $r_i^- > 0$ and $r_i^+ > 0$.

In the case when $\sqrt{\alpha}\mu \leq \sqrt{\rho\epsilon}$, the operator $\mathcal{L}_m^{N,M}$ is applied in the region $(\tau_1, d - \tau_2)$ and $(d + \tau_3, 1 - \tau_4)$ with $\left(\|b\| + \frac{2}{\Delta t}\right) h_k < 2\mu\alpha$, $k = 2, 5$, otherwise $\mathcal{L}_u^{N,M}$ is used. At the transition point, $\mathcal{L}_m^{N,M}$ is applied with $\left(\|b\| + \frac{2}{\Delta t}\right) h_k < 2\mu\alpha$, $k = 2, 3, 5, 6$, otherwise $\mathcal{L}_u^{N,M}$ is used to guarantee $r_i^- > 0$ and $r_i^+ > 0$. At the point of discontinuity, the operator $\mathcal{L}_t^{N,M}$ is used with $\left(\|b\| + \frac{2}{\Delta t}\right) h_k < 2\mu\alpha$, $k = 3, 4$. The associated matrix with the operator $\mathcal{L}_h^{N,M} U^{j+1}(x_i)$ is negative of an M-matrix. The finite difference method is monotone and hence the finite difference operator satisfies the discrete minimum principle. □

Lemma 8. If $U(x_i,t_{j+1}),(x_i,t_{j+1})\in\bar{\Omega}^{N,M}$ is a mesh function satisfying the difference scheme (8), then $\|U\|_{\bar{\Omega}^{N,M}}\leq C$.

Proof. Define the mesh function $\psi(x_i,t_{j+1})=\phi(x_i,t_{j+1})\pm U(x_i,t_{j+1})$, where $\phi(x_i,t_{j+1})=\max\left\{|U^{j+1}(0)|,|U^{j+1}(1)|,\frac{\|g\|_{\bar{\Omega}^{N,M}}}{\beta}\right\}$. Clearly $\psi(x_i,t_{j+1})\geq 0$ for all $(x_i,t_{j+1})\in\Gamma_c^N$ and $\mathcal{L}_h^{N,M}\psi(x_i,t_{j+1})=\mathcal{L}_h^{N,M}\phi(x_i,t_{j+1})\pm\mathcal{L}_h^{N,M}U(x_i,t_{j+1})\leq 0$.

At the point of discontinuity, we have

$$\mathcal{L}_t^{N,M}\psi(x_{\frac{N}{2}},t_{j+1})=\mathcal{L}_t^{N,M}\phi(x_{\frac{N}{2}},t_{j+1})\pm\mathcal{L}_t^{N,M}U(x_{\frac{N}{2}},t_{j+1})\leq 0.$$

Therefore, $\psi(x_i,t_{j+1})\geq 0$, for all $(x_i,t_{j+1})\in\bar{\Omega}^{N,M}$. This leads to the required result

$$\|U\|_{\bar{\Omega}^{N,M}}\leq C$$

□

4 Truncation error analysis

Lemma 9. The singular component $W_l^-(x_i,t_{j+1}),W_l^+(x_i,t_{j+1}),W_r^-(x_i,t_{j+1})$ and $W_r^+(x_i,t_{j+1})$ satisfy the following bounds:

$$|W_l^-(x_i,t_{j+1})|\leq C\gamma_{l,i}^-,\quad \gamma_{l,i}^-=\prod_{n=1}^{i}(1+\theta_2h_n)^{-1},\quad \gamma_{l,0}^-=C_1,\ i=0,1,\ldots,\frac{N}{2},$$

$$|W_l^+(x_i,t_{j+1})|\leq C\gamma_{l,i}^+,\quad \gamma_{l,i}^+=\prod_{n=\frac{N}{2}+1}^{i}(1+\theta_1h_n)^{-1},\quad \gamma_{l,\frac{N}{2}}^+=C_1,\ i=\frac{N}{2}+1,\ldots,N,$$

$$|W_r^-(x_i,t_{j+1})|\leq C\gamma_{r,i}^-,\quad \gamma_{r,i}^-=\prod_{n=i+1}^{\frac{N}{2}}(1+\theta_1h_n)^{-1},\quad \gamma_{r,\frac{N}{2}}^-=C_1,\ i=0,1,\ldots,\frac{N}{2},$$

$$|W_r^+(x_i,t_{j+1})|\leq C\gamma_{r,i}^+,\quad \gamma_{r,i}^+=\prod_{n=i+1}^{N}(1+\theta_2h_n)^{-1},\quad \gamma_{r,N}^+=C_1,\ i=\frac{N}{2}+1,\ldots,N.$$

Proof. Define the barrier function for the left layer term as

$$\eta_{l,i}^{-(j+1)}=C\gamma_{l,i}^{-(j+1)}\pm W_l^{-(j+1)}(x_i),\quad 0\leq i\leq\frac{N}{2},\ 0\leq j\leq M-1,$$

where

$$\gamma_{l,i}^{-(j+1)} = \begin{cases} \prod_{k=1}^{i}(1+\theta_2 h_k)^{-1}, & 1 \le i \le \frac{N}{2}, \\ 1, & i = 0. \end{cases}$$

For the both cases when $\sqrt{\alpha}\mu \le \sqrt{\rho\epsilon}$ and $\sqrt{\alpha}\mu > \sqrt{\rho\epsilon}$, for large enough C, $\eta_{l,0}^{-(j+1)} \ge 0, \eta_{r,i}^{-(0)} \ge 0$ and $\eta_{l,N/2}^{-(j+1)} \ge 0$.

Now,

$$\begin{aligned} \mathcal{L}_c^{N,M}\eta_{l,i}^{-(j+1)} &= \mathcal{L}_c^{N,M}\gamma_{l,i}^{-(j+1)} \pm \mathcal{L}_c^{N,M}W_l^{-(j+1)}(x_i) \\ &= \gamma_{l,i+1}^{-(j+1)}\left(2\epsilon\theta_2^2\left(\frac{h_{i+1}}{h_{i+1}+h_i}-1\right) + 2\epsilon\theta_2^2 - \mu a^{j+\frac{1}{2}}(x_i)\theta_2 \right. \\ &\quad \left. -\mu a^{j+\frac{1}{2}}(x_i)\theta_2^2\frac{h_i h_{i+1}}{h_i+h_{i+1}} - b^{j+\frac{1}{2}}(x_i)\right). \end{aligned}$$

On simplification, we get

$$\begin{aligned} \mathcal{L}_c^{N,M}\eta_{l,i}^{-(j+1)} &\le \gamma_{l,i+1}^{-(j+1)}\left(2\epsilon\theta_2^2 - \mu a^{j+\frac{1}{2}}(x_i)\theta_2 - b^{j+\frac{1}{2}}(x_i)\right) \\ &\le \gamma_{l,i+1}^{-(j+1)}(\rho|a^{j+\frac{1}{2}}(x_i)| - b^{j+\frac{1}{2}}(x_i)) \le 0, \end{aligned}$$

$$\begin{aligned} \mathcal{L}_m^{N,M}\eta_{l,i}^{-(j+1)} &\le \gamma_{l,i+1}^{-(j+1)}\left(2\epsilon\theta_2^2 - \mu \bar{a}^{j+\frac{1}{2}}(x_i)\theta_2 - \bar{b}^{j+\frac{1}{2}}(x_i)\right) \\ &\le \gamma_{l,i+1}^{-(j+1)}(\rho|\bar{a}^{j+\frac{1}{2}}(x_i)| - \bar{b}^{j+\frac{1}{2}}(x_i)) \le 0, \end{aligned}$$

$$\begin{aligned} \mathcal{L}_u^{N,M}\eta_{l,i}^{-(j+1)} &\le \gamma_{l,i+1}^{-(j+1)}\left(2\epsilon\theta_2^2 - \mu a^{j+\frac{1}{2}}(x_i)\theta_2 - b^{j+\frac{1}{2}}(x_i)\right) \\ &\le \gamma_{l,i+1}^{-(j+1)}(\rho|a^{j+\frac{1}{2}}(x_i)| - b^{j+\frac{1}{2}}(x_i)) \le 0. \end{aligned}$$

By the discrete minimum principle for the continuous case [18], we obtain

$$\eta_{l,i}^{-(j+1)} \ge 0 \implies |W_l^{-(j+1)}(x_i)| \le C\prod_{k=1}^{i}(1+\theta_2 h_k)^{-1}, \quad 1 \le i \le \frac{N}{2},\ 0 \le j \le M-1.$$

Similarly, we prove the bound for $W_l^{+(j+1)}(x_i)$ for $\frac{N}{2}+1 \le i \le N-1,\ 0 \le j \le M-1$.

Now, to prove the bound for $W_r^{-(j+1)}(x_i)$, we define the barrier function for the right layer component as

$$\eta_{r,i}^{-(j+1)} = C\gamma_{r,i}^{-(j+1)} \pm W_r^{-(j+1)}(x_i), \quad 0 \le i \le \frac{N}{2},\ 0 \le j \le M-1,$$

where

$$\gamma_{r,i}^{-(j+1)} = \begin{cases} \prod_{k=i+1}^{\frac{N}{2}} (1+\theta_1 h_k)^{-1},\ 0 \le i < \frac{N}{2}, \\ 1, \qquad\qquad\qquad i = \frac{N}{2}. \end{cases}$$

For large enough C, we have $\eta_{r,0}^{-(j+1)} \ge 0, \eta_{r,i}^{-(0)} \ge 0$ and $\eta_{r,N/2}^{-(j+1)} \ge 0$.
Now,

$$\begin{aligned} \mathcal{L}_c^{N,M}\eta_{r,i}^{-(j+1)} &= \mathcal{L}_c^{N,M}\gamma_{r,i}^{-(j+1)} \pm \mathcal{L}_c^{N,M}W_r^{-(j+1)}(x_i) \\ &= \frac{\gamma_{r,i}^{-(j+1)}}{1+\theta_1 h_i}\left(2\epsilon\theta_1^2\big(\frac{h_i}{h_{i+1}+h_i}-1\big) + 2\epsilon\theta_1^2 + \mu a^{j+\frac{1}{2}}(x_i)\theta_1 \right. \\ &\quad \left. + \frac{\theta_1^2\mu a^{j+\frac{1}{2}}(x_i)h_{i+1}}{h_i+h_{i+1}} - b^{j+\frac{1}{2}}(x_i)\right) \\ &\le \frac{\gamma_{r,i}^{-(j+1)}}{1+\theta_1 h_i}\left(2\epsilon\theta_1^2 + \mu a^{j+\frac{1}{2}}(x_i)\theta_1 - b^{j+\frac{1}{2}}(x_i)\right), \quad \text{as } \frac{h_{i+1}}{h_{i+1}+h_i} - 1 \le 0. \end{aligned}$$

For the case when $\sqrt{\alpha}\mu \le \sqrt{\rho\epsilon}$, on simplification, we get

$$\mathcal{L}_c^{N,M}\eta_{r,i}^{-(j+1)} \le \frac{\gamma_{r,i}^{-(j+1)}}{1+\theta_1 h_i}\left(\rho|a^{j+\frac{1}{2}}(x_i)| - b^{j+\frac{1}{2}}(x_i)\right) \le 0,$$

$$\mathcal{L}_m^{N,M}\eta_{r,i}^{-(j+1)} \le \frac{\bar{\gamma_{r,i}}}{1+\theta_1 h_i}\left((\rho|\bar{a}^{j+\frac{1}{2}}(x_i)| - \bar{b}^{j+\frac{1}{2}}(x_i))\right) \le 0,$$

$$\mathcal{L}_u^{N,M}\eta_{r,i}^{-(j+1)} \le \frac{\bar{\gamma_{r,i}}}{1+\theta_1 h_i}\left((\rho|a^{j+\frac{1}{2}}(x_i)| - b^{j+\frac{1}{2}}(x_i))\right) \le 0,$$

and for the case when $\sqrt{\alpha}\mu > \sqrt{\rho\epsilon}$,

$$\mathcal{L}_c^{N,M}\eta_{r,i}^{-(j+1)} \le \frac{\gamma_{r,i}^{-(j+1)}}{1+\theta_1 h_i}\left(\frac{\alpha\mu^2}{2\epsilon}(\alpha + a^{j+\frac{1}{2}}(x_i)) - b^{j+\frac{1}{2}}(x_i)\right) \le 0,$$

$$\mathcal{L}_m^{N,M}\eta_{r,i}^{-(j+1)} \le \frac{\bar{\gamma_{r,i}}}{1+\theta_1 h_i}\left(\frac{\alpha\mu^2}{2\epsilon}(\alpha + \bar{a}^{j+\frac{1}{2}}(x_i)) - \bar{b}^{j+\frac{1}{2}}(x_i))\right) \le 0,$$

$$\mathcal{L}_u^{N,M}\eta_{r,i}^{-(j+1)} \le \frac{\gamma_{r,i}^{-}}{1+\theta_1 h_i}\left(\frac{\alpha\mu^2}{2\epsilon}(\alpha + a^{j+\frac{1}{2}}(x_i)) - b^{j+\frac{1}{2}}(x_i))\right) \le 0.$$

By the discrete minimum principle for the continuous case [18], we obtain

$$\eta_{r,i}^{-(j+1)} \ge 0 \implies |W_r^{-(j+1)}(x_i)| \le C\prod_{k=i+1}^{N/2}(1+\theta_1 h_k)^{-1}, \quad 1 \le i \le \frac{N}{2},\ 0 \le j \le M-1.$$

Similarly, we prove the bound for $W_r^{+(j+1)}(x_i)$ for $\frac{N}{2}+1 \le i \le N-1,\ 0 \le j \le M-1$. □

For the truncation error at the mesh point $(x_i, t_{j+1}) \in \Omega^{N,M}\backslash\{d\}$, when mesh is arbitrary, we have

$$|\mathcal{L}_h^{N,M}(e(x_i))| \le \begin{cases} |(\mathcal{L}_c^{N,M} - \mathcal{L})y_i| \le \epsilon\hbar_i\|y_{xxx}\| + \mu\hbar_i\|a\|\|y_{xx}\| + \|y_{tt}\| \\ |(\mathcal{L}_m^{N,M} - \mathcal{L})y_i| \le \epsilon\hbar_i\|y_{xxx}\| + C\mu h_{i+1}^2(\|y_{xxx}\| + \|y_{xx}\|) + \|y_{tt}\| \\ |(\mathcal{L}_u^{N,M} - \mathcal{L})y_i| \le \epsilon\hbar_i\|y_{xxx}\| + \mu h_{i+1}\|a\|\|y_{xx}\| + \|y_{tt}\|. \end{cases}$$

On a uniform mesh with step size h, we get

$$|\mathcal{L}_h^{N,M}(e(x_i))| \le \begin{cases} |(\mathcal{L}_c^{N,M} - \mathcal{L})y_i| \le \epsilon h^2\|y_{xxxx}\| + \mu h^2|a\|\|y_{xxx}\| + \|y_{tt}\| \\ |(\mathcal{L}_m^{N,M} - \mathcal{L})y_i| \le \epsilon h\|y_{xxx}\| + C\mu h_{i+1}^2(\|y_{xxx}\| + \|y_{xx}\|) + \|y_{tt}\| \\ |(\mathcal{L}_u^{N,M} - \mathcal{L})y_i| \le \epsilon h^2\|y_{xxxx}\| + \mu h\|a\|\|y_{xx}\| + \|y_{tt}\|, \end{cases}$$

where $C_{\|a\|,\|a'\|}$ is a positive constant depending on $\|a\|$ and $\|a'\|$.

Lemma 10. The discrete regular component $V^{j+1}(x_i)$ defined in (11) and $v(x,t)$ is solution of the problem (2). So, the error in the regular component satisfies the following estimate for $\sqrt{\alpha}\mu \le \sqrt{\rho\epsilon}$:

$$\|V - v\|_{\Omega^{N-}\cup\Omega^{N+}} \le C(N^{-2} + \Delta t^2).$$

Proof. For the regular part of the solution $y(x,t)$ of (1) for the case where $\sqrt{\alpha}\mu \le \sqrt{\rho\epsilon}$ and when mesh is uniform ($\tau_1 = \tau_2 = \frac{d}{4}$) in the domain $(x_i, t_{j+1}) \in \Omega^{N-}$, the truncation error is

$$\begin{aligned} |\mathcal{L}_h^{N,M}(V^{-(j+1)} - v^{-(j+1)})(x_i)| &= |\mathcal{L}_c^{N,M}(V^{-(j+1)} - v^{-(j+1)})(x_i)| \\ &\le \left|\epsilon\left(\delta^2 - \frac{d^2}{dx^2}\right)v^{-(j+1)}(x_i)\right| \end{aligned}$$

$$
\begin{aligned}
&+ \mu|a^{j+\frac{1}{2}}(x_i)|\left|\left(D^0 - \frac{d}{dx}\right)v^{-(j+1)}(x_i)\right| \\
&+ \left|\left(D_t^- - \frac{\delta}{\delta t}\right)v^{-(j+1)}(x_i)\right| \\
&\leq \epsilon h^2\|v^-_{xxxx}\| + \mu h^2\|a\|\|v^-_{xxx}\| + C_1\Delta t^2 \\
&\leq C(N^{-2} + \Delta t^2).
\end{aligned}
$$

When mesh is nonuniform in $(x_i, t_{j+1}) \in \Omega^{N-}$ for $\sqrt{\alpha}\mu \leq \sqrt{\rho\epsilon}$:

$$
|\mathcal{L}_c^{N,M}(V^{-(j+1)} - v^{-(j+1)})(x_i)| \leq \epsilon\hbar_i\|v^-_{xxx}\| + \mu\hbar_i\|a\|\|v^-_{xx}\| + C_1\Delta t^2 \leq C(N^{-2} + \Delta t^2).
$$

For the case where $\sqrt{\alpha}\mu > \sqrt{\rho\epsilon}$ and when mesh is uniform ($\tau_1 = \tau_2 = \frac{d}{4}$) in the domain $(x_i, t_{j+1}) \in \Omega^{N-}$:

$$
|\mathcal{L}_h^{N,M}(V^{-(j+1)} - v^{-(j+1)})(x_i)| \leq C(N^{-2} + \Delta t^2).
$$

When mesh is nonuniform:

$$
\begin{aligned}
|\mathcal{L}_c^{N,M}(V^{-(j+1)} - v^{-(j+1)})(x_i)| &\leq \epsilon\hbar_i\|v^-_{xxx}\| + \mu\hbar_i\|a\|\|v^-_{xx}\| + C_1\Delta t^2 \\
&\leq C(N^{-2} + \Delta t^2),
\end{aligned}
$$

$$
\begin{aligned}
|\mathcal{L}_m^{N,M}(V^{-(j+1)} - v^{-(j+1)})(x_i)| &\leq \epsilon\hbar_i\|v^-_{xxx}\| + C\mu h_{i+1}^2(\|v^-_{xxx}\| + \|v^-_{xx}\|) \\
&\leq C(N^{-2} + \Delta t^2),
\end{aligned}
$$

$$
\begin{aligned}
|\mathcal{L}_u^{N,M}(V^{-(j+1)} - v^{-(j+1)})(x_i)| &\leq \epsilon\hbar_i\|v^-_{xxx}\| + C\mu h_{i+1}(\|v^-_{xx}\|) \\
&\leq C(N^{-2} + \Delta t^2).
\end{aligned}
$$

At the transition point τ_1:

$$
|\mathcal{L}_m^{N,M}(V^{-(j+1)} - v^{-(j+1)})(x_i)| \leq C(N^{-2} + \Delta t^2).
$$

Define the barrier function in $(x_i, t_{j+1}) \in \Omega^{N-}$:

$$
\psi^{j+1}(x_i) = C(N^{-2} + \Delta t^2) \pm (V^{-(j+1)} - v^{-(j+1)})(x_i).
$$

For large C, $\psi^{j+1}(0) \geq 0$, $\psi^{j+1}(x_{\frac{N}{2}}) \geq 0, \psi^0(x_i) \geq 0$ and $\mathcal{L}_h^{N,M}\psi^{j+1}(x_i) \leq 0$. Hence, using the approach given in [8], we get $\psi^{j+1}(x_i) \geq 0$ and

$$|(V^{-(j+1)} - v^{-(j+1)})(x_i)|_{\Omega^{N-}} \leq C(N^{-2} + \Delta t^2). \tag{14}$$

Similarly, we can estimate the error bounds for right regular part in the domain $(x_i, t_{j+1}) \in \Omega^{N+}$:

$$|(V^{+(j+1)} - v^{+(j+1)})(x_i)|_{\Omega^{N+}} \leq C(N^{-2} + \Delta t^2). \tag{15}$$

Combining the above results (14) and (15), we obtain

$$\|V - v\|_{\Omega^{N-} \cup \Omega^{N+}} \leq C(N^{-2} + \Delta t^2).$$

□

Lemma 11. Let $W_l(x_i, t_{j+1}), w_l(x, t)$ be solution of the problem (12) and (3), respectively. The left singular component of the truncation error satisfies the following estimate:

$$|\|W_l - w_l\|_{\Omega^{N-} \cup \Omega^{N+}} \leq \begin{cases} C((N^{-1} \ln N)^2 + \Delta t^2), & \sqrt{\alpha}\mu \leq \sqrt{\rho\epsilon}, \\ C(N^{-2}(\ln N)^3 + \Delta t^2), & \sqrt{\alpha}\mu > \sqrt{\rho\epsilon}. \end{cases} \tag{16}$$

Proof. The truncation error in case of uniform ($\tau_1 = \tau_2 = \frac{d}{4}$) for the case $\sqrt{\alpha}\mu \leq \sqrt{\rho\epsilon}$ in the domain $(x_i, t_{j+1}) \in \Omega^{N-}$ is

$$\begin{aligned} |\mathcal{L}_c^{N,M}(W_l^{-(j+1)} - w_l^{-(j+1)})(x_i)| &\leq CN^{-2}(\epsilon\|w_{l,xxxx}^{-(j+1}\| + \mu\|w_{l,xxx}^{-(j+1)}\|) + C_1\Delta t^2\|w_{tt}\| \\ &\leq C(N^{-2}(\ln N)^2 + \Delta t^2). \end{aligned}$$

In the case of nonuniform mesh, we split the argument into two parts. For $(x_i, t_{j+1}) \in [\tau_1, d) \times (0, T]$ from Lemma 5, we obtain

$$|w_l^{-(j+1)}(x_i)| \leq C \exp^{-\theta_2 x_i} \leq C \exp^{-\theta_2 \tau_1} \leq CN^{-2}. \tag{17}$$

Also, from Lemma 9, we have

$$|W_l^{-(j+1)}(x_i)| \leq C \prod_{k=1}^{\frac{N}{8}} (1 + \theta_2 h_k)^{-1} \leq \left(1 + \frac{16 \ln N}{N}\right)^{-N/8} \leq CN^{-2}.$$

Hence, for all $(x_i, t_{j+1}) \in [\tau_1, d) \times (0, T]$, we have

$$|(W_l^{-(j+1)} - w_l^{-(j+1)})(x_i)| \leq |W_l^{-(j+1)}(x_i)| + |w_l^{-(j+1)}(x_i)| \leq CN^{-2}.$$

For $\sqrt{\alpha}\mu \leq \sqrt{\rho\epsilon}$, the truncation error for the left layer component in the inner region $(0,\tau_1)\times(0,T]$, is

$$|\mathcal{L}_c^{N,M}(W_l^{-(j+1)} - w_l^{-(j+1)})(x_i)| \leq C(N^{-2}(\ln N)^2 + \Delta t^2).$$

We choose the barrier function for the layer component as

$$\psi^{j+1}(x_i) = C_1((N^{-1}\ln N)^2 + \Delta t^2) \pm (W_l^{-(j+1)} - w_l^{-(j+1)})(x_i), \qquad (x_i, t_{j+1}) \in \Omega^{N-}.$$

For sufficiently large C, we have $\psi^{j+1}(x_0) \geq 0, \psi^{j+1}(x_{N/8}) \geq 0, \psi^0(x_i) \geq 0$ and $\mathcal{L}_c^{N,M}\psi^{j+1}(x_i) \leq 0$. Hence by the discrete maximum principle in [18], $\psi^{j+1}(x_i) \geq 0$. So, we can obtain the following bounds:

$$|(W_l^{-(j+1)} - w_l^{-(j+1)})(x_i)| \leq C(N^{-2}(\ln N)^2 + \Delta t^2), \quad \text{for all } (x_i, t_{j+1}) \in \Omega^{N-}.$$

For the case where $\sqrt{\alpha}\mu > \sqrt{\rho\epsilon}$ and when the mesh is uniform ($\tau_1 = \tau_2 = \frac{d}{4}$) in $(x_i, t_{j+1}) \in \Omega^{N-}$:

$$|\mathcal{L}_h^{N,M}(W_l^{-(j+1)} - w_l^{-(j+1)})(x_i)| \leq C(N^{-2}(\ln N)^3 + \Delta t^2).$$

The truncation error for the left layer component in the inner region $(0,\tau_1)\times(0,T]$, when mesh is nonuniform is

$$|\mathcal{L}_c^{N,M}(W_l^{-(j+1)} - w_l^{-(j+1)})(x_i)| \leq CN^{-2}(\ln N)^3 + C_1\Delta t^2,$$

$$\begin{aligned}|\mathcal{L}_m^{N,M}(W_l^{-(j+1)} - w_l^{-(j+1)})(x_i)| &\leq Ch_1\epsilon\|w_{l,xxx}^-\| + Ch_1^2\mu(\|w_{l,xxx}^-\| + \|w_{l,xx}^-\|) + C_1\Delta^2 \\ &\leq Ch_1^2\mu(\|w_l^{-(3)}\| + \|w_l^{-(2)}\|) + C_1\Delta t^2 \\ &\leq C(N^{-2}(\ln N)^3 + \Delta t^2),\end{aligned}$$

$$|\mathcal{L}_u^{N,M}(W_l^{-(j+1)} - w_l^{-(j+1)})(x_i)| \leq C(N^{-2}(\ln N)^3 + \Delta t^2).$$

We choose the barrier function for the layer component as

$$\psi^{j+1}(x_i) = C_1((N^{-1}\ln N)^3 + \Delta t^2) \pm (W_l^{-(j+1)} - w_l^{-(j+1)})(x_i),\ (x_i, t_{j+1})) \in \Omega^{N-}.$$

For sufficiently large C, we have $\psi^{j+1}(x_0) \geq 0, \psi^{j+1}(x_{N/8}) \geq 0, \psi^0(x_i) \geq 0$ and $\mathcal{L}_h^{N,M}\psi^{j+1}(x_i) \leq 0$. Hence by the discrete maximum principle in [18], $\psi^{j+1}(x_i) \geq 0$. So, we can obtain the following bounds in $(0,d)\times(0,T]$:

$$|(W_l^{-(j+1)} - w_l^{-(j+1)})(x_i)| \leq \begin{cases} C((N^{-1}\ln N)^2 + \Delta t^2), & \sqrt{\alpha}\mu \leq \sqrt{\rho\epsilon}, \\ C(N^{-2}(\ln N)^3 + \Delta t^2), & \sqrt{\alpha}\mu > \sqrt{\rho\epsilon}. \end{cases} \tag{18}$$

By a similar argument in the domain $(d,1)\times(0,T]$, we have

$$|(W_l^{+(j+1)} - w_l^{+(j+1)})(x_i)| \leq \begin{cases} C((N^{-1}\ln N)^2 + \Delta t^2), & \sqrt{\alpha}\mu \leq \sqrt{\rho\epsilon}, \\ C(N^{-2}(\ln N)^3 + \Delta t^2), & \sqrt{\alpha}\mu > \sqrt{\rho\epsilon}. \end{cases} \tag{19}$$

Combining the results (18) and (19), the desired result is obtained. □

Lemma 12. Let $W_r^-(x_i,t_{j+1}), w_r^-(x,t)$ be solution of the problem (13) and (4), respectively. The right singular component of the truncation error satisfies the following estimate:

$$\|W_r - w_r\|_{\Omega^{N-}\cup\Omega^{N+}} \leq \begin{cases} C((N^{-1}\ln N)^2 + \Delta t^2), & \sqrt{\alpha}\mu \leq \sqrt{\rho\epsilon}, \\ C(N^{-2}(\ln N)^3 + \Delta t^2), & \sqrt{\alpha}\mu > \sqrt{\rho\epsilon}. \end{cases} \tag{20}$$

Proof. The truncation error for $\sqrt{\alpha}\mu \leq \sqrt{\rho\epsilon}$, when the mesh is uniform ($\tau_1 = \tau_2 = \frac{d}{4}$) in $(x_i,t_{j+1}) \in \Omega^{N-}$ is

$$\begin{aligned} |\mathcal{L}_c^{N,M}(W_r^{-(j+1)} - w_r^{-(j+1)})(x_i)| &\leq CN^{-2}(\epsilon\|w_{l,xxxx}^-\| + \mu\|w_{r,xxx}^-\|) + C_1\Delta t^2 \\ &\leq C(N^{-2}(\ln N)^2 + \Delta t^2). \end{aligned}$$

In case of nonuniform mesh, we split the interval $(0,d)\times(0,T]$ into two parts $(0,d-\tau_2]\times(0,T]$ and $(d-\tau_2,d)\times(0,T]$.

In the domain $(x_i,t_{j+1}) \in (0,d-\tau_2]\times(0,T]$, we have

$$|w_r^{-(j+1)}(x_i)| \leq C\exp^{-\theta_1(d-x_i)} \leq C\exp^{-\theta_1\tau_2} \leq CN^{-2}. \tag{21}$$

Also, from Lemma 9,

$$|W_r^{-(j+1)}(x_i)| \leq C\prod_{k=1}^{\frac{N}{8}}(1+\theta_1 h_k)^{-1} \leq \left(1+\frac{16\ln N}{N}\right)^{-N/8} \leq CN^{-2}.$$

Hence, for all $(x_i,t_{j+1}) \in (0,d-\tau_2]\times(0,T]$, we have

$$|(W_r^{-(j+1)} - w_r^{-(j+1)})(x_i)| \leq |W_r^{-(j+1)}(x_i)| + |w_r^{-(j+1)}(x_i)| \leq CN^{-2}.$$

For $\sqrt{\alpha}\mu \leq \sqrt{\rho\epsilon}$, the truncation error for the left layer component in the inner region $(d-\tau_2, d)\times(0,T]$, is

$$|\mathcal{L}_c^{N,M}(W_r^{-(j+1)} - w_r^{-(j+1)})(x_i)| \leq C(N^{-2}(\ln N)^2 + \Delta t^2).$$

We choose the barrier function for the layer component as

$$\psi^{j+1}(x_i) = C_1((N^{-1}\ln N)^2+\Delta t^2)\pm(W_r^{-(j+1)}-w_r^{-(j+1)})(x_i), \qquad (x_i,t_{j+1})\in\Omega^{N-}.$$

For sufficiently large C_1, we have $\psi^{j+1}(x_{\frac{3N}{8}})\geq 0, \psi^{j+1}(x_{N/2})\geq 0, \psi^0(x_i)\geq 0$ and $\mathcal{L}_h^{N,M}\psi^{j+1}(x_i)\leq 0$. Hence by the discrete maximum principle in [18], $\psi^{j+1}(x_i)\geq 0$. So, by using the discrete maximum principle, we can obtain the following bounds:

$$|(W_r^{-(j+1)}-w_r^{-(j+1)})(x_i)| \leq C(N^{-2}(\ln N)^2+\Delta t^2), \quad \text{for all } (x_i,t_{j+1})\in\Omega^{N-}.$$

For $\sqrt{\alpha}\mu > \sqrt{\rho\epsilon}$, when mesh is uniform ($\tau_1=\tau_2=\frac{d}{4}$) in $(0,d)\times(0,T]$, we have

$$|\mathcal{L}_h^{N,M}(W_r^{-(j+1)} - w_r^{-(j+1)})(x_i)| \leq C(N^{-2}(\ln N)^3 + \Delta t^2).$$

In case of nonuniform mesh, the truncation error for the left layer component in the inner region $(d-\tau_2, d)\times(0,T]$, is

$$|\mathcal{L}_c^{N,M}(W_r^{-(j+1)} - w_r^{-(j+1)})(x_i)| \leq CN^{-2}(\ln N)^2 + C_1\Delta t^2.$$

We choose the barrier function for the layer component as

$$\psi^{j+1}(x_i) = C_1(N^{-2}\ln N^3+\Delta t^2)\pm(W_r^{-(j+1)}-w_r^{-(j+1)})(x_i), \qquad (x_i,t_{j+1})\in\Omega^{N-}.$$

For sufficiently large C_1, we have $\psi^{j+1}(x_{\frac{3N}{8}})\geq 0, \psi^{j+1}(x_{N/2})\geq 0, \psi^0(x_i)\geq 0$ and $\mathcal{L}_h^{N,M}\psi^{j+1}(x_i)\leq 0$. Hence by the discrete maximum principle in [18], $\psi^{j+1}(x_i)\geq 0$. So, by using the discrete maximum principle, we can obtain the following bounds in $(x_i,t_{j+1})\in\Omega^{N-}$:

$$|(W_r^{-(j+1)} - w_r^{-(j+1)})(x_i)| \leq \begin{cases} C((N^{-1}\ln N)^2 + \Delta t^2), & \sqrt{\alpha}\mu \leq \sqrt{\rho\epsilon}, \\ C(N^{-2}(\ln N)^3 + \Delta t^2), & \sqrt{\alpha}\mu > \sqrt{\rho\epsilon}. \end{cases} \quad (22)$$

By a similar argument in the domains $(x_i,t_{j+1})\in\Omega^{N+}$, we have

$$|(W_r^{+(j+1)} - w_r^{+(j+1)})(x_i)| \leq \begin{cases} C((N^{-1}\ln N)^2 + \Delta t^2), & \sqrt{\alpha}\mu \leq \sqrt{\rho\epsilon}, \\ C(N^{-2}(\ln N)^3 + \Delta t^2), & \sqrt{\alpha}\mu > \sqrt{\rho\epsilon}. \end{cases} \quad (23)$$

Combining the results (22) and (23), the desired result is obtained. □

Lemma 13. The error $e(x_{\frac{N}{2}}, t_{j+1})$ estimated at the point of discontinuity $(x_{\frac{N}{2}}, t_{j+1}) = (d, t_{j+1}), 0 \leq j \leq M-1$ satisfies the following estimates:

$$|\mathcal{L}_h^{N,M}(U-u)(x_{\frac{N}{2}}, t_{j+1})| \leq \begin{cases} \dfrac{Ch^2}{\epsilon^{3/2}}, & \sqrt{\alpha}\mu \leq \sqrt{\rho\epsilon}, \\ \dfrac{Ch^2\mu^3}{\epsilon^3}, & \sqrt{\alpha}\mu > \sqrt{\rho\epsilon}, \end{cases}$$

where $h = \max\{h_3, h_4\}$.

Proof. For $\sqrt{\alpha}\mu \leq \sqrt{\rho\epsilon}$, the truncation error at the point $x_{\frac{N}{2}} = d$ is

$$\begin{aligned} &\left| \mathcal{L}_h^{N,M} U^{j+1}(x_{N/2}) - \frac{2h_3^2 g_{N/2-1}^{j+1} h_4}{2\epsilon - h_3\mu a_{N/2-1}^{j+\frac{1}{2}}} - \frac{2h_4^2 g_{N/2+1}^{j+1} h_3}{2\epsilon + h_4\mu a_{N/2+1}^{j+\frac{1}{2}}} \right| \\ &\leq |\mathcal{L}_t^{N,M} U^{j+1}(x_{N/2})| \\ &\quad + \frac{2h_3^2 h_4}{2\epsilon - h_3\mu a_{N/2-1}^{j+\frac{1}{2}}} |\mathcal{L}_c^{N,M} u(x_{N/2-1}, t_{j+1}) - g_{N/2-1}^{j+1}| \\ &\quad + \frac{2h_4^2 h_3}{2\epsilon + h_4\mu a_{N/2+1}^{j+\frac{1}{2}}} |\mathcal{L}_c^{N,M} u(x_{N/2+1}, t_{j+1}) - g_{N/2+1}^{j+1}| \\ &\leq Ch^2 \|u_{xxx}\| \\ &\leq \frac{Ch^2}{\epsilon^{3/2}}. \end{aligned}$$

By using the bounds given in Lemma 4 for $\|u_{xxx}\|$, we get the above estimate. For $\sqrt{\alpha}\mu > \sqrt{\rho\epsilon}$, the truncation error at the point $x_{\frac{N}{2}} = \frac{d}{2}$ is

$$\begin{aligned} &\left| \mathcal{L}_h^{N,M} U^{j+1}(x_{N/2}) - \frac{2h_3^2 g_{N/2-1}^{j+\frac{1}{2}} h_4}{2\epsilon - h_3\mu a_{N/2-1}^{j+\frac{1}{2}}} - \frac{2h_4^2 g_{N/2+1}^{j+\frac{1}{2}} h_3}{2\epsilon + h_4\mu a_{N/2+1}^{j+\frac{1}{2}}} \right| \\ &\leq |\mathcal{L}_t^{N,M} U^{j+1}(x_{N/2})| \\ &\quad + \frac{2h_3^2 h_4}{2\epsilon - h_3\mu a_{N/2-1}^{j+\frac{1}{2}}} |\mathcal{L}_c^{N,M} u(x_{N/2-1}, t_{j+1}) - g_{N/2-1}^{j+1}| \end{aligned}$$

$$+\frac{2h_4^2h_3}{2\epsilon+h_4\mu a_{N/2+1}^{j+\frac{1}{2}}}|\mathcal{L}_c^{N,M}u(x_{N/2+1},t_{j+1})-g_{N/2+1}^{j+1}|$$
$$\leq Ch^2\|u_{xxx}\|$$
$$\leq \frac{Ch^2\mu^3}{\epsilon^3}$$

By using the bounds given in Lemma 4 for $\|u_{xxx}\|$, we get the above estimate. □

Theorem 3. Let $u(x,t)$ and $U(x,t)$ be the solutions of (1) and (8), respectively. Then

$$\|U-u\|_{\bar{\Omega}^{N,M}} \leq C(N^{-2}\ln N^3+\Delta t^2),$$

where C is a constant independent of ϵ,μ and discretization parameters N,M.

Proof. From Lemmas 10, 11 and 12, we have

$$\|U-u\|_{\Omega^{N-}\cup\Omega^{N+}} \leq \begin{cases} C(N^{-2}(\ln N)^2+\Delta t^2), & \sqrt{\alpha}\mu \leq \sqrt{\rho\epsilon}, \\ C(N^{-2}(\ln N)^3+\Delta t^2), & \sqrt{\alpha}\mu > \sqrt{\rho\epsilon}. \end{cases}$$

Let $\sqrt{\alpha}\mu \leq \sqrt{\rho\epsilon}$. To find error at the point of discontinuity $x_{\frac{N}{2}}=d$, we have considered the discrete barrier function $\phi_1(x_i,t_{j+1}) = \psi_1(x_i,t_{j+1}) \pm e(x_i,t_{j+1})$ defined in the interval $(d-\tau_2,d+\tau_3)\times(0,T]$, where

$$\psi_1(x_i,t_{j+1}) = C(N^{-2}(\ln N)^3+\Delta t^2) + \frac{Ch^2}{\epsilon^{3/2}}\begin{cases} x_i-(d+\tau_2), & (x_i,t_{j+1})\in\Omega^{N,M}\cap(d-\tau_2,d)\times(0,T], \\ d-\tau_3-x_i, & (x_i,t_{j+1})\in\Omega^{N,M}\cap(d,d+\tau_3)\times(0,T]. \end{cases}$$

For $x_i \in (d-\tau_2,d+\tau_3)$, $\phi_1(d-\tau_2,t_{j+1})$, $\phi_1(d+\tau_3,t_{j+1}),\phi_1(x_i,t_0)$ are nonnegative and $\mathcal{L}_h^{N,M}\phi_1(x_i,t_{j+1}) \leq 0$, for all $(x_i,t_{j+1}) \in (d-\tau_2,d+\tau_3)\times(0,T]$. Also, $\mathcal{L}_t^{N,M}\phi_1(x_{N/2},t_{j+1}) \leq 0$.

Hence, by applying the discrete minimum principle, we get $\phi_1(x_i,t_{j+1}) \geq 0$.

Therefore, for $(x_i,t_{j+1}) \in (d-\tau_2,d+\tau_3)\times(0,T]$, we have

$$|(U-u)(x_i,t_{j+1})| \leq C_1(N^{-2}(\ln N)^3+\Delta t^2)+\frac{C_2h^2\tau}{\epsilon^{3/2}} \leq C(N^{-2}(\ln N)^3+\Delta t^2). \tag{24}$$

For $\sqrt{\alpha}\mu > \sqrt{\rho\epsilon}$, consider the discrete barrier function $\phi_2(x_i, t_{j+1}) = \psi_2(x_i, t_{j+1}) \pm e(x_i, t_{j+1})$ defined in the interval $(d-\tau_2, d+\tau_3) \times (0, T]$, where

$$\psi_2(x_i, t_{j+1}) = C(N^{-2}(\ln N)^3 + \Delta t^2) + C_1 \frac{h^2\mu^3}{\epsilon^3} \begin{cases} x_i - (d+\tau_2), & (x_i, t_{j+1}) \in \Omega^{N,M} \cap (d-\tau_2, d) \times (0, T], \\ d - \tau_3 - x_i, & (x_i, t_{j+1}) \in \Omega^{N,M} \cap (d, d+\tau_3) \times (0, T]. \end{cases}$$

We have $\phi_2(d-\tau_2, t_{j+1})$ and $\phi_2(d+\tau_3, t_{j+1}), \phi_2(x_i, t_0),\ x_i \in (d-\tau_2, d+\tau_3)$ are nonnegative and $\mathcal{L}_h^{N,M}\phi_2(x_i, t_{j+1}) \leq 0$, for all $(x_i, t_{j+1}) \in (d-\tau_2, d+\tau_3) \times (0, T]$. Also, $\mathcal{L}_t^{N,M}\phi_2(x_i, t_{j+1}) \leq 0$.
Hence by applying the discrete minimum principle, we get $\phi_2(x_i, t_{j+1}) \geq 0$. Therefore, for $(x_i, t_{j+1}) \in (d-\tau_2, d+\tau_3) \times (0, T]$, we have

$$|(U-u)(x_i, t_{j+1})| \leq C(N^{-2}(\ln N)^3 + \Delta t^2). \tag{25}$$

By combining the result (24) and (25), we obtain the desired result. □

5 Numerical examples

To demonstrate the effectiveness of the proposed hybrid difference approach, we proposed a numerical method on three test problems with discontinuous convection coefficients and source terms. As the exact solutions to these problems are unknown, we used the double mesh method [7, 6] to evaluate the accuracy of the numerical approximations obtained. The double mesh difference is defined by

$$E^{N,M} = \max_j \left(\max_i |U_{2i,2j}^{2N,2M} - U_{i,j}^{N,M}| \right),$$

where $U_i^{N,M}$ and $U_{2i}^{2N,2M}$ are the solutions on the mesh $\bar{\Omega}^{N,M}$ and $\bar{\Omega}^{2N,2M}$, respectively. The order of convergence is given by

$$R^{N,M} = \log_2 \left(\frac{E^{N,M}}{E^{2N,2M}} \right).$$

We also compute the error and order of convergence by fixing ϵ and varying μ from a larger set, say $\mu \in S = \{10^{-j}, 2, 4, 6\}$ and name them as

$$E_{\epsilon}^{N,M} = \max_{\mu \in S} E^{N,M}.$$

The order of convergence is given by

$$R_{\epsilon}^{N,M} = \log_2 \left(\frac{E_{\epsilon}^{N,M}}{E_{\epsilon}^{2N,2M}} \right).$$

Example 1. Consider

$$(\epsilon u_{xx} + \mu a u_x - bu - u_t)(x,t) = f(x,t), \qquad (x,t) \in ((0,.5) \cup (0.5,1)) \times (0,1],$$

$$u(0,t) = u(1,t) = u(x,0) = 0,$$

with

$$a(x,t) = \begin{cases} -(1+x(1-x)), & 0 \le x \le 0.5, t \in (0,1], \\ 1+x(1-x), & 0.5 < x \le 1, t \in (0,1], \end{cases}$$

$$f(x,t) = \begin{cases} -2(1+x^2)t, & 0 \le x \le 0.5, t \in (0,1], \\ 2(1+x^2)t, & 0.5 < x \le 1, t \in (0,1], \end{cases}$$

and $b(x,t) = 1 + \exp(x)$.

Example 2. Consider

$$(\epsilon u_{xx} + \mu a u_x - bu - u_t)(x,t) = f(x,t), \qquad (x,t) \in ((0,0.5)) \cup (0.5,1)) \times (0,1],$$
$$u(0,t) = u(1,t) = u(x,0) = 0,$$

with

$$a(x,t) = \begin{cases} -(1+x(1-x)), & 0 \le x \le 0.5, t \in (0,1], \\ 1+x(1-x), & 0.5 < x \le 1, t \in (0,1], \end{cases}$$

$$f(x,t) = \begin{cases} -2(1+x^2)t, & 0 \le x \le 0.5, t \in (0,1], \\ 3(1+x^2)t, & 0.5 < x \le 1, t \in (0,1], \end{cases}$$

with and $b(x,t) = 1 + \exp(x)$.

Example 3. Consider

$$(\epsilon u_{xx} + \mu a u_x - bu - u_t)(x,t) = f(x,t), \qquad (x,t) \in ((0,.5) \cup (0.5,1)) \times (0,1],$$

$$u(0,t) = u(1,t) = u(x,0) = 0,$$

with

$$a(x,t) = \begin{cases} -(1+\exp(-xt)), & 0 \le x \le 0.5, t \in (0,1], \\ 2+x+t, & 0.5 < x \le 1, t \in (0,1], \end{cases}$$

$$f(x,t) = \begin{cases} (\exp(t^2)-1)(1+xt), & 0 \le x \le 0.5, t \in (0,1], \\ -(2+x)t^2, & 0.5 < x \le 1, t \in (0,1], \end{cases}$$

and $b(x,t) = 2+xt$.

6 Discussion

Table 1: Maximum point-wise error $E^{N,M}$ and approximate orders of convergence $R^{N,M}$ for Example 1 when $\epsilon = 2^{-6}$.

μ	Number of mesh points N					
	32	64	128	256	512	1024
2^{-8}	4.17e-03	1.41e-03	6.13e-04	2.94e-04	1.45e-04	6.99e-05
Order	1.5623	1.2055	1.0627	1.0181	1.0526	
2^{-12}	3.24e-03	8.65e-04	2.26e-04	5.92e-05	1.58e-05	3.98e-06
Order	1.9038	1.9330	1.9357	1.9099	1.9891	
2^{-16}	3.22e-03	8.54e-04	2.21e-04	5.62e-05	1.42e-05	3.55e-06
Order	1.9902	1.9972	1.9991	1.9995	1.9991	
2^{-20}	3.21e-03	8.53e-04	2.20e-04	5.60e-05	1.41e-05	3.53e-06
Order	1.9132	1.9532	1.9753	1.9867	1.9979	
2^{-24}	3.21e-03	8.53e-04	2.20e-04	5.60e-05	1.41e-05	3.53e-06
Order	1.9132	1.9532	1.9761	1.9879	1.9979	
⋮	⋮	⋮	⋮	⋮	⋮	⋮
2^{-60}	3.21e-03	8.53e-04	2.20e-04	5.60e-05	1.41e-05	3.53e-06
Order	1.9132	1.9532	1.9761	1.9880	1.9979	

Table 1 tabulates the error and numerical order of convergence for the Example 1 for different values of μ when $\epsilon = 2^{-6}$. Table 2 presents the numerical results for different values of ϵ and $\mu = 2^{-6}$ for Example 1. In Table 2, we note that the order of convergence for different values of ϵ increased till $N = 256$ and slightly decreased for $N = 512$. The numerical order of convergence for $N = 512$ is still two. This decrease in convergence is due

Table 2: Maximum point-wise error $E^{N,M}$ and approximate orders of convergence $R^{N,M}$ for Example 1 when $\mu = 2^{-6}$.

ϵ	Number of mesh points N				
	64	128	256	512	1024
2^{-16}	2.32e-01	9.41e-02	2.52e-02	5.63e-03	1.67e-03
Order	1.3012	1.9025	2.1589	1.7513	
2^{-20}	2.31e-01	1.03e-01	2.43e-02	5.57e-03	1.46e-03
Order	1.1738	2.0757	2.0678	1.9272	
2^{-24}	2.31e-01	1.03e-01	2.41e-02	5.48e-03	1.35e-03
Order	1.1643	2.0955	2.1374	2.0222	
2^{-28}	2.31e-01	1.03e-01	2.41e-02	5.47e-03	1.34e-03
Order	1.1638	2.0967	2.1415	2.0329	
2^{-32}	2.31e-01	1.03e-01	2.41e-02	5.46e-03	1.33e-03
Order	1.1637	2.0969	2.1424	2.0330	
2^{-36}	2.31e-01	1.03e-01	2.41e-02	5.46e-03	1.33e-3
Order	1.1637	2.0969	2.1424	2.0331	
2^{-40}	2.31e-01	1.03e-01	2.41e-02	5.47e-03	1.34e-03
Order	1.1637	2.0969	2.1424	2.0331	

Table 3: Maximum point-wise error $E_\epsilon^{N,M}$ and approximate orders of convergence $R_\epsilon^{N,M}$ for Example 1.

ϵ	Number of mesh points N				
	64	128	256	512	1024
10^{-2}	4.70e-03	2.25e-03	1.13e-03	5.69e-04	2.69e-04
Order	1.0632	0.9951	0.9871	1.0799	
10^{-4}	1.70e-01	6.05e-02	1.55e-02	3.88e-03	1.10e-03
Order	1.4912	1.9646	1.9982	1.8212	
10^{-6}	2.19e-01	9.18e-02	2.38e-02	5.79e-03	1.59e-03
Order	1.3947	2.0482	2.0955	2.0558	
10^{-8}	9.66e-02	4.22e-02	9.92e-03	2.37e-03	5.79e-04
Order	1.1963	2.0881	2.0654	2.0323	
10^{-10}	1.19e-01	4.22e-02	9.92e-03	2.37e-03	6.04e-04
Order	1.4918	2.0883	2.0667	2.0106	
10^{-12}	2.39e-01	8.09e-02	1.95e-02	4.58e-03	1.10e-03
Order	1.5658	2.0486	2.0938	2.0581	
10^{-14}	2.39e-01	8.09e-02	1.96e-02	4.58e-03	1.10e-03
Order	1.5658	2.0486	2.0938	2.0581	

Table 4: Maximum point-wise error $E^{N,M}$ and approximate orders of convergence $R^{N,M}$ for Example 2 when $\epsilon = 2^{-6}$.

μ	Number of mesh points N				
	64	128	256	512	1024
2^{-10}	4.67e-03	1.55e-03	6.01e-04	2.54e-04	1.20e-04
Order	1.5912	1.3668	1.2425	1.0818	
2^{-14}	4.00e-03	1.03e-03	2.62e-04	6.61e-05	1.66e-05
Order	1.9521	1.9751	1.9868	1.9939	
2^{-18}	4.00e-03	1.03e-03	2.63e-04	6.63e-05	1.66e-05
Order	1.9515	1.9750	1.9865	1.9938	
2^{-22}	4.00e-03	1.03e-03	2.63e-04	6.63e-05	1.67e-05
Order	1.9515	1.9750	1.9865	1.9938	
2^{-28}	4.00e-03	1.03e-03	2.63e-04	6.63e-05	1.67e-05
Order	1.9515	1.9750	1.9865	1.9938	
2^{-32}	4.00e-03	1.03e-03	2.63e-04	6.63e-05	1.67e-05
Order	1.9515	1.9750	1.9865	1.9938	
⋮	⋮	⋮	⋮	⋮	⋮
2^{-60}	4.00e-03	1.03e-03	2.63e-04	6.63e-05	1.67e-05
Order	1.9515	1.9750	1.9865	1.9938	

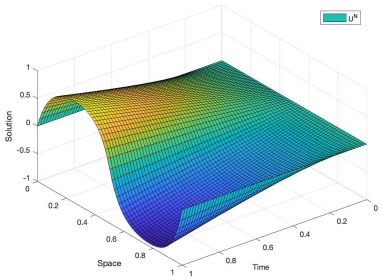


Figure 2: Numerical solution for $\epsilon = 2^{-6}, \mu = 2^{-32}$ when $N = 64$ for Example 1.

to the fact that in boundary layer region upwind scheme is replaced by the midpoint upwind scheme. The midpoint upwind scheme is not performing as well as the upwind scheme.

Figures 2 and 3 represent the surface plot of the numerical solution and error graph of Example 1, respectively, for $\mu = 2^{-32}$, $\epsilon = 2^{-6}$ and $N = 64$. We note that the maximum error is at the point of discontinuity. Figures 4 and

Table 5: Maximum point-wise error $E^{N,M}$ and approximate orders of convergence $R^{N,M}$ for Example 2 when $\mu = 2^{-6}$.

ϵ	Number of mesh points N				
	64	128	256	512	1024
2^{-18}	3.31e-01	1.51e-01	3.72e-02	9.95e-03	2.66e-03
Order	1.1321	2.0226	1.9015	1.9011	
2^{-22}	3.55e-01	1.55e-01	3.63e-02	9.75e-03	2.55e-03
Order	1.1982	2.0914	1.8940	1.9330	
2^{-26}	3.55e-01	1.55e-01	3.63e-02	9.75e-03	2.55e-03
Order	1.1982	2.0914	1.8940	1.9330	
2^{-30}	3.55e-01	1.55e-01	3.63e-02	9.75e-03	2.55e-03
Order	1.1982	2.0914	1.8940	1.9330	
2^{-34}	3.55e-01	1.55e-01	3.63e-02	9.75e-03	2.55e-03
Order	1.1982	2.0914	1.8940	1.9330	
2^{-38}	3.55e-01	1.55e-01	3.63e-02	9.75e-03	2.55e-03
Order	1.1982	2.0914	1.8940	1.9330	
2^{-42}	3.55e-01	1.55e-01	3.63e-02	9.75e-03	2.55e-03
Order	1.1982	2.0914	1.8940	1.9330	

Table 6: Maximum point-wise error $E^{N,M}$ and approximate orders of convergence $R^{N,M}$ for Example 3 when $\epsilon = 2^{-8}$.

μ	Number of mesh points N					
	32	64	128	256	512	1024
2^{-14}	7.74e-03	2.08e-03	5.31e-04	1.34e-04	4.79e-05	2.19e-05
Order	1.8947	1.9704	1.9834	1.4871	1.1292	
2^{-16}	7.83e-03	2.14e-03	5.62e-04	1.34e-04	3.36e-05	8.49e-06
Order	1.8692	1.9572	1.9654	1.9974	1.9858	
2^{-20}	7.86e-03	2.17e-03	5.69e-04	1.45e-04	3.85e-05	9.72e-06
Order	1.9108	1.9481	1.9657	1.9672	1.9868	
2^{-24}	7.86e-03	2.17e-03	5.69e-04	1.45e-04	3.85e-05	9.72e-06
Order	1.9902	1.9972	1.9991	1.9995	1.9872	
2^{-28}	7.86e-03	2.17e-03	5.69e-04	1.46e-04	3.73e-05	9.43e-06
Order	1.8592	1.9299	1.9605	1.9694	1.9845	
⋮	⋮	⋮	⋮	⋮	⋮	⋮
2^{-60}	7.86e-03	2.1e-03	5.69e-04	1.46e-04	3.73e-05	9.43e-06
Order	1.8592	1.9299	1.9605	1.9694	1.9845	

Table 7: Maximum point-wise error E^N and approximate orders of convergence R^N for Example 3 when $\mu = 2^{-6}$.

ϵ	Number of mesh points N				
	64	128	256	512	1024
2^{-18}	2.28e-02	9.24e-03	2.87e-03	7.47e-04	1.97e-04
Order	1.3030	1.8330	1.9301	1.9221	
2^{-22}	2.27e-02	9.25e-03	2.54e-03	6.61e-04	1.73e-04
Order	1.2951	1.8634	1.9434	1.9339	
2^{-26}	2.27e-02	9.25e-03	2.54e-03	6.61e-04	1.73e-04
Order	1.2951	1.8634	1.9434	1.9339	
2^{-30}	2.97e-02	9.26e-03	2.54e-03	6.61e-04	1.73e-04
Order	1.2951	1.8634	1.9434	1.9339	
2^{-34}	2.97e-02	9.2561e-03	2.54e-03	6.61e-04	1.73e-04
Order	1.2951	1.8634	1.9434	1.9339	
2^{-38}	2.97e-02	9.26e-03	2.54e-03	6.61e-04	1.73e-04
Order	1.2951	1.8634	1.9434	1.9339	
2^{-42}	2.97e-02	9.26e-03	2.54e-03	6.61e-04	1.73e-04
Order	1.2951	1.8634	1.9434	1.9339	

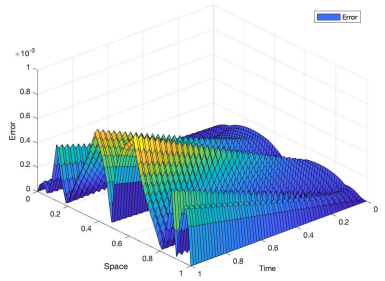


Figure 3: Error for $\epsilon = 2^{-6}, \mu = 2^{-32}$ when $N = 64$ for Example 1.

5 show the numerical solution and error graph for Example 1, respectively, when $\epsilon = 2^{-30}$, $\mu = 2^{-6}$, and $N = 128$. Here too the maximum error occurs at the point of discontinuity.

In Table 3, we present the maximum point-wise error and order of convergence for $\mu \in S = \{10^{-j}, j = 2, 4, 6.\}$ and for different ϵ. Tables 4 and 5 tabulate the value for different ϵ and μ for Example 2. In Figures 6 and 7, the numerical solution and error graph for Example 2, respectively, when

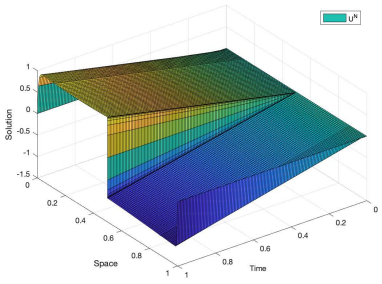


Figure 4: Numerical solution for $\epsilon = 2^{-30}, \mu = 2^{-6}$ when $N = 128$ for Example 1.

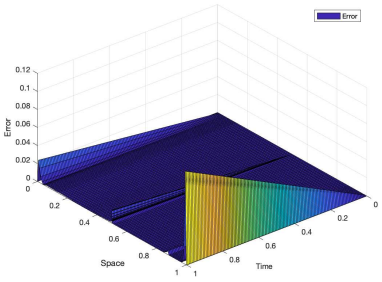


Figure 5: Error for $\epsilon = 2^{-30}, \mu = 2^{-6}$ when $N = 128$ for Example 1.

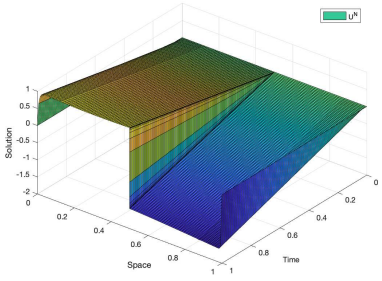


Figure 6: Numerical solution for $\epsilon = 2^{-22}, \mu = 2^{-6}$ when $N = 128$ for Example 2.

$\epsilon = 2^{-22}$, $\mu = 2^{-6}$, and $N = 128$ are presented. Here, the maximum error occurs at the boundaries near $x = 1$.

Similarly, Table 6 tabulates the error and numerical order of convergence for Example 3 for different values of μ when $\epsilon = 2^{-8}$. Table 7 tabulates the

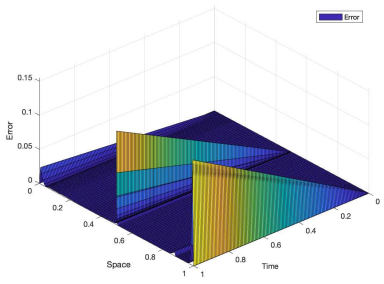


Figure 7: Error for $\epsilon = 2^{-22}, \mu = 2^{-6}$ when $N = 128$ for Example 2.

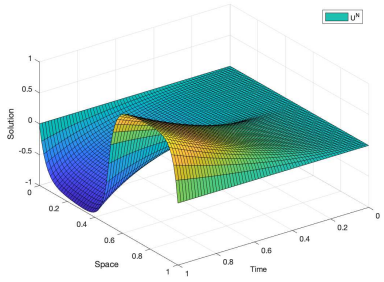


Figure 8: Numerical solution for $\epsilon = 2^{-8}, \mu = 2^{-36}$ when $N = 64$ for Example 3.

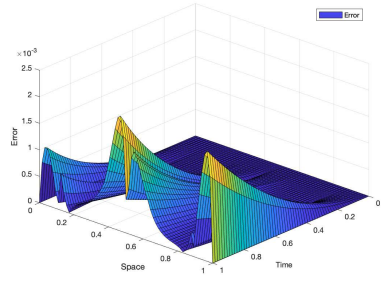


Figure 9: Error for $\epsilon = 2^{-8}, \mu = 2^{-36}$ when $N = 64$ for Example 3.

error and numerical order of convergence for Example 3 for different values of ϵ when $\mu = 2^{-6}$.

Figures 8 and 9 represent the surface plot of the numerical solution and error graph of Example 3, respectively, for $\epsilon = 2^{-8}$, $\mu = 2^{-36}$, and $N = 64$.

We note that the maximum error is at the point of discontinuity. Here, we also noted the same pattern in Tables 3, 5, and 7, as noted in Table 2, that is, the order of convergence is increased and then decreased for higher values of N.

7 Conclusions

In this article, we studied a singularly perturbed two-parameter parabolic problem characterized by a discontinuous convection coefficient and source term. The solution reveals boundary layers near the domain boundaries and internal layers at the point of discontinuity, induced by the discontinuities in the convection coefficient and source term. For the temporal discretization, we utilized the Crank–Nicolson method on a uniform mesh, achieving second-order accuracy in time. In the spatial direction, we implemented a hybrid difference scheme, which integrates the midpoint method, central difference method and upwind difference method on a suitably defined Shishkin mesh with a five-point formula at the discontinuity point, ensuring almost second-order accuracy in space. The theoretical analysis was corroborated by numerical results, demonstrating the efficacy and accuracy of the proposed approach.

Acknowledgements

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

References

[1] Bullo, T.A., Duressa, G.F. and Degla, G.A. *Higher order fitted operator finite difference method for two-parameter parabolic convection-diffusion problems*, Int. J. Eng. Technol. Manag. Appl. Sci. 11(4) (2019), 455–467.

[2] Cen, Z. *A hybrid difference scheme for a singularly perturbed convection-*

diffusion problem with discontinuous convection coefficient, Appl. Math. Comput. 169(1) (2005), 689–699.

[3] Chandru, M., Das, P. and Ramos, H. *Numerical treatment of two-parameter singularly perturbed parabolic convection diffusion problems with non-smooth data*, Math. Methods Appl. Sci. 41(14) (2018), 5359–5387.

[4] Chandru, M., Prabha, T., Das, P. and Shanthi, V. *A Numerical Method for Solving Boundary and Interior Layers Dominated Parabolic Problems with Discontinuous Convection Coefficient and Source Terms,* Differ. Equ. Dyn. Syst. 27 (2019), 91–112.

[5] Crank, J. and Nicolson, P. *A practical method for numerical evaluation of solutions of partial differential equations of the heat-conduction type*, Adv. Comput. Math. 7 (1996), 207–226.

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

[7] Das, P. and Mehrmann, V. *Numerical solution of singularly perturbed convection-diffusion-reaction problems with two small parameters*, BIT Numer. Math. 56(1) (2016), 51–76.

[8] Farrell, P.A., Hegarty, A.F., Miller, J.J.H., O'Riordan, E. and Shishkin, G.I. *Robust Computational Techniques for Boundary Layers*, vol. 1. Chapman & Hall/CRC,New York, 2000.

[9] Gracia, G.L. and O'Riordan, E. *Numerical approximation of solution derivatives in the case of singularly perturbed time dependent reaction–diffusion problems*, J. Comput. Appl. Math. 273 (2015), 13–24.

[10] Gupta, V., Kadalbajoo, M.K. and Dubey, R.K. *A parameter-uniform higher order inite difference scheme for singularly perturbed time-dependent parabolic problem with two small parameters*, Int. J. Comput. Math. 96(3) (2018), 1–29.

[11] Kadalbajoo, M.K. and Yadaw, A.S. *Parameter-uniform finite element method for two-parameter singularly perturbed parabolic reaction-diffusion problems*, Int. J. Comput. Methods 9(4) (2012), 1250047.

[12] Kumar, D. and Kumari, P. *Uniformly convergent scheme for two-parameter singularly perturbed problems with non-smooth data*, Numer. Methods Partial Differ.Equ. 37(1) (2021), 796–817.

[13] Markowich, P.A. *A finite difference method for the basic stationary semiconductor device equations*. numerical boundary value ODEs (Vancouver, B.C.,1984), Progr. Sci. Comput., Vol. 5, Birkhäuser Boston, Boston, MA, (1985), 285–301.

[14] Mekonnen, T.B. and Duressa, G.F. *Computational method for singularly perturbed two-parameter parabolic convection-diffusion problems*, IMA J. Numer. Anal. 7(1) (2020), 1829277.

[15] Mukherjee, K. and Natesan, S. *ϵ-uniform error estimate of hybrid numerical scheme for singularly perturbed parabolic problems with interior layers*, Numer. Algorithms, 58 (2011), 103–141.

[16] Munyakazi, J.B. *A robust finite difference method for two-parameter parabolic convection-diffusion problems*, Appl. Math. Inf. Sci. 9(6)(2015), 2877–2883.

[17] O'Malley, R.E. *Introduction to Singular Perturbations*, Academic Press, New York, London, 1974.

[18] O'Riordan, E., Pickett, M.L. and Shishkin, G.I. *Singularly perturbed problems modeling reaction-convection-diffusion processes*, Comput. Methods Appl. Math. 3(3) (2003), 424–442.

[19] O'Riordan, E., Shishkin, G.I. and Picket, M.L. *Parameter-uniform finite difference schemes for singularly perturbed parabolic diffusion-convection-reaction problems*, Math. Comput. 75(255) (2006), 1135–1154.

[20] Schlichting, H. *Boundary Layer Theory*, seventh ed., McGraw-Hill, New York, 1979.

[21] Singh, S. and Kumar, D. *Parameter uniform numerical method for a system of singularly perturbed parabolic convection–diffusion equations,* Math. Comput. Simul. 212 (2023), 360–381.

[22] Singh, S., Choudhary, R. and Kumar, D. *An efficient numerical technique for two-parameter singularly perturbed problems having discontinuity in convection coefficient and source term*, Comput. Appl. Math. 42(62) (2023), 42–62.

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

[24] Zahra, W.K., El-Azab, M.S. and El Mhlawy, A.M. *Spline difference scheme for two-parameter singularly perturbed partial differential equations*, Int. J. Appl. Math. 32(1-2) (2014), 185–201.

Iranian Journal of Numerical Analysis and Optimization
Vol. 15, No. 2, 2025, pp 770–822
https://doi.org/10.22067/ijnao.2025.92092.1596
https://ijnao.um.ac.ir/

Research Article

A numerical algorithm based on Jacobi polynomials for FIDEs with error estimation

K. Sadri*, D. Amilo and E. Hincal

Abstract

This study aims to address a specific class of mathematical problems known as fractional integro-differential equations. These equations are used to model various phenomena,, including heat conduction in materials with

*Corresponding author

Received 8 February 2025; revised 5 March 2025; accepted 14 March 2025

Khadijeh Sadri
Mathematics Research Center, Near East University TRNC, Mersin 10, Nicosia 99138, Turkey., Faculty of Art and Science, University of Kyrenia, Kyrenia, TRNC, Mersin 10, Turkey. e-mail: khadijeh.sadrikhatouni@neu.edu.tr

David Amilo
Mathematics Research Center, Near East University TRNC, Mersin 10, Nicosia 99138, Turkey., Faculty of Art and Science, University of Kyrenia, Kyrenia, TRNC, Mersin 10, Turkey. e-mail: amilodavid.ikechukwu@neu.edu.tr

Evren Hincal
Mathematics Research Center, Near East University TRNC, Mersin 10, Nicosia 99138, Turkey., Faculty of Art and Science, University of Kyrenia, Kyrenia, TRNC, Mersin 10, Turkey., Research Center of Applied Mathematics, Khazar University, Baku, Azerbaijan
e-mail: evren.hincal@neu.edu.tr

How to cite this article
Sadri, K., Amilo, D. and Hincal, E., A numerical algorithm based on Jacobi polynomials for FIDEs with error estimation. *Iran. J. Numer. Anal. Optim.*, 2025; 15(2): 770-822. https://doi.org/10.22067/ijnao.2025.92092.1596

memory, damping laws, diffusion processes, earthquake models, fluid dynamics, traffic flow, and continuum mechanics. This research focuses on problems where the fractional derivative operator is defined in the Caputo sense. Our proposed methodology employs an operational approach based on the use of shifted Jacobi polynomials. We derive operational matrices for fractional integration and product, which are then applied to approximate solutions for both linear and nonlinear problems. By using these matrices in conjunction with the collocation method, we transform the original problem into a system of algebraic equations. Notably, our approach is simpler and more cost-effective compared to established methods such as Adomian decomposition, Homotopy perturbation, Sinc-collocation, and Legendre wavelet techniques. We provide several illustrative examples to validate our method's effectiveness and reliability. Additionally, we present theorems concerning the existence of a unique solution and the convergence of our proposed approach.

AMS subject classifications (2020): Primary 33C45; Secondary 65D15, 34K37, 45J05.

Keywords: Fractional integro-differential equation; Caputo derivative operator; Jacobi polynomials; Operational matrices; Convergence.

1 Introduction

Fractional calculus has gained significant attention from mathematicians and engineers in recent decades due to the widespread use of fractional integral and derivative operators in various disciplines and engineering applications [22, 18, 19, 27, 4]. These operators are relevant in modeling numerous physical phenomena, including heat conduction in materials with memory, damping laws, diffusion processes, earthquake models, fluid dynamics, traffic flow, continuum mechanics, chemistry, acoustics, and psychology [22, 18, 19, 27, 4]. Functional equations of fractional order, such as fractional integro-differential equations (FIDEs), are crucial for accurately modeling these phenomena. The derivatives in these equations typically appear in the Riemann–Liouville and Caputo senses. The Caputo derivative is especially favored in practice due to its physically interpretable initial conditions, which resemble

those of integer-order differential equations. In contrast, initial conditions derived from the Riemann–Liouville derivative lack a clear physical interpretation [22, 18, 19, 27, 4]. FIDEs arise in various scientific and engineering fields, such as viscosity measurement in oil exploration, continuum and statistical mechanics, chemical kinetics, fluid dynamics, and biological models [12, 11, 23, 26]. Understanding the properties and physical nature of these equations is crucial, leading many researchers and mathematicians to develop or refine methods for solving integro-differential equations with fractional derivatives. However, solving these equations presents both analytical and numerical challenges, driving the search for effective methods. In this context, Huang et al. [22] extended the Taylor method to solve a generalization of fractional differential, Fredholm integral, and Volterra integral equations. Authors in [16, 15] developed the Homotopy perturbation method to solve both linear and nonlinear FIDEs. The Homotopy analysis method has been applied to solve linear FIDEs in [1], while in [24], the Chebyshev pseudo-spectral method was used to solve systems of FIDEs of the Volterra type. In [25], the authors presented three numerical schemes for solving linear FIDEs. Mokhtary [29] applied a discrete Galerkin method to solve linear FIDEs. Laguerre polynomials and a collocation method were developed in [49] to solve fractional linear Volterra integro-differential equations. Nazari, Shahmorad, and Jahanshahi [31] proposed a quadrature method for solving nonlinear FIDEs of the Hammerstein type. Babaei, Jafari, and Banihashemi [5] introduced a collocation approach based on sixth-kind Chebyshev polynomials to reduce variable-order FIDEs to a system of algebraic equations and determined an approximate solution. Darweesh, Al-Khaled, and Al-Yaqeen [9] used the Laplace Haar wavelet method to solve systems of linear fractional Fredholm integro-differential equations and evaluated the rate of convergence.

Among the many numerical methods developed, spectral methods are particularly notable for their high accuracy and ease of implementation. These methods, including Galerkin, collocation, and tau methods, involve expressing the solution to a functional equation as a linear combination of basis functions, which transforms the original equation into a discrete algebraic form. In this context, Rahimkhani, Ordokhani, and Babolian [34]

employed a Bernoulli wavelet operational approach to solve fractional delay differential equations. In [35], the authors introduced generalized fractional-order Bernoulli wavelet functions to obtain numerical solutions for fractional-order pantograph differential equations. In [37], a new technique based on Muntz–Legendre functions and their operational matrices was proposed to solve fractional-order pantograph equations. Bernoulli wavelet functions were also applied to solve delay fractional-order optimal control problems in [33]. Rahimkhani, Ordokhani, and Babolian [36] utilized fractional-order Bernoulli functions to achieve numerical solutions for fractional Fredholm–Volterra integro-differential equations. Rahimkhani and Ordokhani [32] used a Bernoulli wavelet collocation method to solve fractional-order partial differential equations. The Bernstein collocation method was developed in [41] to solve Sylvester matrix differential equations. A method combining Bell polynomials and the Galerkin approach was proposed to address fractional optimal control problems [43]. Sadek and Bataineh [40] presented generalized Bernstein functions and the collocation method to approximate solutions for χ-fractional differential equations. Jacobi orthogonal polynomials are widely used as basis functions in spectral methods. For example, shifted Jacobi polynomials have been employed alongside spectral tau and collocation methods to solve linear and nonlinear multi-term fractional differential equations [13]. Similarly, the shifted Chebyshev spectral tau method has been used to construct numerical solutions for linear multi-order fractional differential equations [6]. A new numerical approach for solving fractional-order pantograph partial differential equations was introduced in [52] by utilizing two-variable Gegenbauer polynomials. The Lucas wavelets, combined with the Legendre–Gauss quadrature rule and modified operational matrices for integration and pseudo-operational fractional derivatives, were used to study the solution of fractional Fredholm–Volterra integro-differential equations [10]. Hosseininia, Heydari, and Avazzadeh [21] proposed the use of orthonormal shifted discrete Legendre polynomials and the collocation method for solving the variable-order fractional extended Fisher–Kolmogorov equation. A finite class of Romanovski–Jacobi polynomials was used as basis functions in the spectral tau method to approximate solutions to time-fractional partial differential equations on a semi-infinite interval [3]. Shifted Jacobi polynomials

were employed by Heydari, Zhagharian, and Razzaghi [20] to find numerical solutions to one- and two-dimensional stochastic multi-order fractional diffusion-wave equations. In [2], shifted Jacobi polynomials and a fractional-order shifted Jacobi–Gauss collocation method were applied to solve FIDEs with weakly singular kernels. In [51], shifted sixth-kind Chebyshev polynomials, combined with the collocation method, were used to convert systems of FIDEs with weakly singular kernels into algebraic systems, allowing for approximate solutions. Authors in [7] employed shifted Jacobi polynomials with an operational collocation method to obtain approximate solutions to a class of weakly singular FIDEs. Ebrahimi and Sadri [14] solved fractional-order pantograph differential equations by introducing fractional-order Jacobi functions based on Jacobi polynomials, presenting a relationship between the basis functions with delay and the original Jacobi polynomials. These examples highlight the versatility and effectiveness of spectral methods in solving FIDEs and related problems.

In this paper, we extend the use of classical orthogonal Jacobi polynomials to solve both linear and nonlinear integro-differential equations with fractional orders of arbitrary nature. Specifically, we focus on three distinct types of these equations:

$$D^{\nu}y(t)+h(t)\,y(t)+\int_0^t k(t,s)\,y(s)\,ds+\int_0^t \tilde{k}(t,s)\,D^{\gamma}y(s)\,ds=f(t),\quad 0\leqslant t\leqslant 1,\quad 0<\gamma<\nu\leqslant m,$$

$$y''(t)+h(t)\,y'(t)+g(t)\,D^{\nu}y(t)+\int_0^t k(t,s)\,y(s)\,ds=f(t),\qquad 0\leqslant t\leqslant 1,\quad 0<\nu<2,$$

$$D^{\nu}y(t)+h(t)\,y(t)+\lambda_1\int_0^t k(t,s)\,y^2(s)\,ds+\lambda_2\int_0^1 \tilde{k}(t,s)\,y^2(s)\,ds=f(t),\quad 0\leqslant t\leqslant 1.$$

Under appropriate initial conditions, let $\nu,\gamma\in\mathbb{R}$, $m\in\mathbb{Z}^+$, and f,g,h,k, and $\tilde{k}$ be known and continuous real-valued functions. The function y is unknown, and λ_1 and λ_2 are real numbers. As the first step in our solution approach, we derive the fractional operational matrices for integration and multiplication through straightforward algebraic computations, thus reducing computational costs. These resulting matrices and approximations are then substituted into the original problem, transforming it into a corresponding system of algebraic equations. This system is subsequently collocated at the $N+1$ roots of the $N+1$th shifted Jacobi polynomials, $P_{N+1}^{(\alpha,\beta)}(t)$, where

$t \in [0, 1]$. Notably, Jacobi polynomials are parameterized by α and β, both of which are greater than -1. By varying these parameters, different variants of this family of orthogonal polynomials can be obtained. Solving the resulting system gives an approximate solution, denoted as $y_N(t)$. For handling nonlinear systems, we employ the well-known Newton's iteration method. Additionally, we provide a convergence analysis of our proposed method within a Jacobi-weighted Sobolev space.

The contributions of this paper can be summarized as follows:

- Enhanced precision compared to existing methods.
- Flexibility in adjusting parameters α and β, enabling investigation into their impact on obtained solutions.
- Computation of error bounds for resulting approximate solutions within a Jacobi-weighted Sobolev space.
- Estimation of errors in approximate solutions in scenarios where exact solutions are unknown, allowing comparison with absolute errors.

The novelty of this research lies in its innovative application of the Jacobi operational method to solve a wide range of integro-differential equations (IDEs) featuring fractional derivatives of arbitrary order ν. By harnessing the versatility of shifted Jacobi polynomials on the interval $[0, 1]$, the study introduces a novel approach to transforming complex IDEs—both linear and nonlinear—into more manageable algebraic equations. This method not only simplifies the solution process but also offers a systematic framework for determining coefficients and constructing operational matrices for fractional integration and product. Furthermore, the study delves into investigating the existence, uniqueness, and convergence of solutions, shedding light on the method's theoretical underpinnings and practical efficacy. Through illustrative examples and comparative analyses with existing methods, the research showcases the superior accuracy and computational efficiency of the proposed approach, positioning it as a promising tool for tackling a broad spectrum of functional equations. The organization of this paper is as follows:

- Section 1 introduces the topic.

- Section 2 provides preliminary definitions and concepts of fractional calculus.
- Section 3 delves into the existence and uniqueness of solutions for the equations under consideration.
- Section 4 discusses shifted Jacobi polynomials and their properties, along with deriving their operational matrices of fractional integration and product.
- Section 5 presents and proves theorems regarding the convergence of the proposed method.
- Section 6 offers several examples to illustrate the simplicity and efficiency of our proposed method.
- Finally, section 7 contains discussions and conclusions.

2 Preliminaries

Despite the various definitions of fractional-order derivative and integral operators, such as the Θ-conformable fractional derivatives [39], Caputo cotangent fractional derivatives [38], Hilfer cotangent fractional derivatives [42], and q-trigonometric derivatives [44], the Caputo derivative and Riemann–Liouville integral operators remain popular and widely used. Some definitions and properties of fractional operators used in this paper are recalled [48].

Definition 1. The Riemann–Liouville fractional integral operator of order $\nu > 0$ of a function $u(t)$, is defined as

$$\begin{aligned} &I^{\nu}u(t) = \frac{1}{\Gamma(\nu)}\int_0^t (t-s)^{\nu-1}\,u(s)\,ds, \qquad \nu > 0, \quad t > 0,\\ &I^{0}u(t) = u(t), \end{aligned} \tag{1}$$

where $\Gamma(\alpha)$ is the Gamma function defined as

$$\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1}\,\exp(-t)\,dt, \quad Re(\alpha) > 0.$$

Definition 2. The fractional derivative operator in the Caputo sense of order $\nu > 0$ of a function $u(t)$, is given as

$$D^{\nu}u(t) = I^{m-\nu}D^{m}u(t) = \frac{1}{\Gamma(m-\nu)}\int_0^t (t-s)^{m-\nu-1}\frac{d^m}{ds^m}u(s)\,ds,\ m-1<\nu\leqslant m,\ t>0,$$
$$D^{0}u(t) = u(t). \tag{2}$$

Some properties of the operators I^{ν} and D^{ν} are recalled as follows:

1. $I^{\nu_1}I^{\nu_2}u(t) = I^{\nu_1+\nu_2}u(t)$,
2. $I^{\nu}(\lambda_1\,u_1(t)+\lambda_2\,u_2(t)) = \lambda_1\,I^{\nu}u_1(t)+\lambda_2\,I^{\nu}u_2(t)$,
3. $I^{\nu}t^{\gamma} = \frac{\Gamma(\gamma+1)}{\Gamma(\nu+\gamma+1)}t^{\nu+\gamma}, \quad \gamma>-1$,
4. $D^{\nu}I^{\nu}u(t) = u(t)$,
5. $I^{\nu}D^{\nu}u(t) = u(t) - \sum_{k=0}^{m-1}\frac{u^{(k)}(0^+)}{k!}t^k, \quad m-1<\nu\leqslant m$,
6. $D^{\nu}t^{\gamma} = \begin{cases} 0, & \nu>\gamma, \\ \frac{\Gamma(\gamma+1)}{\Gamma(\gamma-\nu+1)}t^{\gamma-\nu}, & \text{otherwise}, \end{cases}$
7. $D^{\nu}\lambda = 0, \quad \lambda\in\mathbb{R}$,

where $\nu,\nu_1,\nu_2,\gamma,\lambda,\lambda_1,\lambda_2\in\mathbb{R}$.

3 Existence and uniqueness of solution

In this section, we explore various forms of integro-differential equations involving fractional derivatives of arbitrary order. To investigate the existence of unique solutions and the convergence of our proposed methodology, we analyze the following FIDE:

$$D^{\nu}y(t)+h(t)\,y(t)+\int_0^t k(t,s)\,y(s)\,ds+\int_0^t \tilde{k}(t,s)\,D^{\gamma}y(s)\,ds = f(t), \qquad 0\leqslant t\leqslant 1, \tag{3}$$

where $0<\gamma<\nu$, $f(t)$, $h(t)$, $k(t,s)$, and $\tilde{k}(t,s)$, $t,s\in[0,1]$, are continuous real functions, and the proper initial conditions $y^{(i)}(0)=d_i$, $i=0,1,\ldots,m-1$ exist, where $m-1<\nu\leqslant m$.

In this section, the existence of a solution of (3) is established using the fixed point theorem. Let Y be a Banach space and let $C(J,Y)$ be the

Banach space of continuous function $y(t) \in Y$, $t \in J = [0,1]$, with the norm $\|y\| = \max_{t\in J} |y(t)|$. Moreover, suppose that $B_r(y, Y)$ is a closed ball with center at y and radius r in Y. By applying the fractional integral operator, (3) will be converted to the following integral equation:

$$\begin{aligned} y(t) = F(t) &- \frac{1}{\Gamma(\nu)} \int_0^t (t-s)^{\nu-1} h(s)\, y(s)\ ds - \frac{1}{\Gamma(\nu)} \int_0^t (t-s)^{\nu-1} \int_0^s k(s,s')\, y(s')\, ds'\, ds \\ &- \frac{1}{\Gamma(\nu)} \int_0^t (t-s)^{\nu-1} \int_0^s \tilde{k}(s,s')\, D^{\gamma} y(s')\, ds'\, ds, \end{aligned} \tag{4}$$

where

$$F(t) = \sum_{i=0}^{m-1} \frac{d_i\, t^i}{\Gamma(i+1)} + \frac{1}{\Gamma(\nu)} \int_0^t (t-s)^{\nu-1} f(s)\, ds.$$

Now, suppose that for any given function $y(t) \in Y$ and kernels $k(t,s), \tilde{k}(t,s)$, there exist positive constants $L, M, \widetilde{M} \in \mathbb{R}$ such that the following conditions are satisfied:

$$\begin{aligned} &\|D^{\gamma} y(t)\| \leqslant L\ \|y(t)\|, \quad \text{for all } t \in J, \\ &\|k(t,s)\| \leqslant M, \qquad \|\tilde{k}(t,s)\| \leqslant \widetilde{M}, \quad \text{for all } (t,s) \in J \times J. \end{aligned} \tag{5}$$

Following theorem states the existence of a solution of (3).

Theorem 1. Suppose that $\nu > 0 \in \mathbb{R}$, that conditions (5) hold, and that $\|h\|/\Gamma(\nu+1) + (M + \widetilde{M}L)/\Gamma(\nu+2) < 1/2$, where $\|h\| = \max_{t\in J} |h(t)|$. Then, fractional-order integro-differential equation (3) has a unique solution.

Proof. Let $W = C(J, Y)$, and define a mapping $\Psi y(t) : W \longrightarrow W$, as follows:

$$\begin{aligned} \Psi y(t) = F(t) &- \frac{1}{\Gamma(\nu)} \int_0^t (t-s)^{\nu-1} h(s)\, y(s)\ ds - \frac{1}{\Gamma(\nu)} \int_0^t (t-s)^{\nu-1} \int_0^s k(s,s')\, y(s')\, ds'\, ds \\ &- \frac{1}{\Gamma(\nu)} \int_0^t (t-s)^{\nu-1} \int_0^s \tilde{k}(s,s')\, D^{\gamma} y(s')\, ds'\, ds. \end{aligned}$$

It should be shown that Ψ has a fixed point and this fixed point is a solution of (3). Set $r \geqslant 2\|F\|$, where $\|F\| = \max_{t\in J} |F(t)|$. Then, it can be shown that $\Psi B_r \subseteq B_r$, where $B_r = \{y(t) \in W |\, \|y\| \leqslant r\}$. So, one has

$$\begin{aligned} ||\Psi y(t)|| \leqslant ||F|| &+ \frac{||h||}{\Gamma(\nu)} \int_0^t (t-s)^{\nu-1} ||y(s)||\, ds \\ &+ \frac{1}{\Gamma(\nu)} \int_0^t (t-s)^{\nu-1} \int_0^s ||k(s,s')||\, ||y(s')||\, ds'\, ds \end{aligned}$$

$$
\begin{aligned}
&+\frac{1}{\Gamma(\nu)}\int_0^t (t-s)^{\nu-1}\int_0^s ||\tilde{k}(s,s')||\,||D^{\gamma}y(s')||\,ds'\,ds\\
&\leqslant ||F||+\frac{||h||\,r}{\nu\,\Gamma(\nu)}+\frac{M\,r}{\nu(\nu+1)\,\Gamma(\nu)}+\frac{\widetilde{M}\,L\,r}{\nu(\nu+1)\,\Gamma(\nu)}\\
&=||F||+\frac{||h||\,r}{\Gamma(\nu+1)}+\frac{M\,r}{\Gamma(\nu+2)}+\frac{\widetilde{M}\,L\,r}{\Gamma(\nu+2)}\\
&\leqslant r.
\end{aligned}
$$

Thus, Ψ maps B_r into itself. Now, for $y_1(t), y_2(t) \in W$, one has

$$
\begin{aligned}
||\Psi y_1(t)-\Psi y_2(t)|| &\leqslant \frac{||h||}{\Gamma(\nu)}\int_0^t (t-s)^{\nu-1}||y_1(s)-y_2(s)||\,ds\\
&\quad+\frac{M}{\Gamma(\nu)}\int_0^t (t-s)^{\nu-1}\int_0^s ||y_1(s')-y_2(s')||\,ds'\,ds\\
&\quad+\frac{\widetilde{M}}{\Gamma(\nu)}\int_0^t (t-s)^{\nu-1}\int_0^s ||D^{\gamma}(y_1(s')-y_2(s'))||\,ds'\,ds\\
&\leqslant\left(\frac{||h||}{\Gamma(\nu+1)}+\frac{M+\widetilde{M}L}{\Gamma(\nu+2)}\right)||y_1(t)-y_2(t)||.
\end{aligned}
$$

Since $||h||/\Gamma(\nu+1)+(M+\widetilde{M}L)/\Gamma(\nu+2)<1/2$, the mapping Ψ is a contraction one, and therefore a unique fixed point $y(t)\in B_r$ exists such that $\Psi y(t)=y(t)$. □

4 Shifted Jacobi polynomials and their operational matrices

The shifted Jacobi polynomials are defined on the interval $[0,1]$, with the weight function $w^{(\alpha,\beta)}(t)=t^{\beta}(1-t)^{\alpha}$. These polynomials can be determined by the following recursive relation [47]:

$$
P_{i+1}^{(\alpha,\beta)}(t)=A(\alpha,\beta,i)\,P_i^{(\alpha,\beta)}(t)+(2t-1)\,B(\alpha,\beta,i)\,P_i^{(\alpha,\beta)}(t)-C(\alpha,\beta,i)\,P_{i-1}^{(\alpha,\beta)}(t),\ i=1,2,\ldots, \tag{6}
$$

where

$$
\begin{aligned}
A(\alpha,\beta,i)&=\frac{(2i+\alpha+\beta+1)(\alpha^2-\beta^2)}{2(i+1)(i+\alpha+\beta+1)(2i+\alpha+\beta)},\\
B(\alpha,\beta,i)&=\frac{(2i+\alpha+\beta+2)(2i+\alpha+\beta+1)}{2(i+1)(i+\alpha+\beta+1)},
\end{aligned}
$$

$$C(\alpha,\beta,i)=\frac{(i+\alpha)(i+\beta)(2i+\alpha+\beta+2)}{(i+1)(i+\alpha+\beta+1)(2i+\alpha+\beta)},$$

and

$$P_0^{(\alpha,\beta)}(t)=1,\quad P_1^{(\alpha,\beta)}(t)=\frac{\alpha+\beta+2}{2}(2t-1)+\frac{\alpha-\beta}{2}.$$

In addition, these shifted polynomials are orthogonal, that is,

$$\int_0^1 P_i^{(\alpha,\beta)}(t)\,P_j^{(\alpha,\beta)}(t)\,w^{(\alpha,\beta)}(t)\,dt=h_i\;\delta_{ij},$$

where

$$h_i=\frac{\Gamma(i+\alpha+1)\Gamma(i+\beta+1)}{(2i+\alpha+\beta+1)\,i!\,\Gamma(i+\alpha+\beta+1)},$$

and δ_{ij} denotes the Kronecker delta function. Moreover, the shifted Jacobi polynomials have the following power series presentation, which will be used throughout this research:

$$P_i^{(\alpha,\beta)}(t)=\sum_{k=0}^{i}\frac{(-1)^{i-k}\Gamma(i+\beta+1)\Gamma(i+k+\alpha+\beta+1)\,t^k}{\Gamma(k+\beta+1)\Gamma(i+\alpha+\beta+1)\,(i-k)!\,k!},\qquad i=0,1,\ldots. \tag{7}$$

A square integrable function $u(t)$, in the interval $[0,1]$, can be expressed in terms of shifted Jacobi polynomials as the following equation:

$$u(t)=\sum_{j=0}^{\infty}\hat{u}_j\,P_j^{(\alpha,\beta)}(t), \tag{8}$$

where the coefficients $\hat{u}_j$ are given by

$$\hat{u}_j=\frac{1}{h_j}\int_0^1 u(t)\,P_j^{(\alpha,\beta)}(t)\,w^{(\alpha,\beta)}(t)\,dt,\qquad j=0,1,\ldots.$$

In practice, only the first $(N+1)$ terms of the shifted Jacobi polynomials are considered. Therefore, we have

$$u(t)\approx u_N(t)=\sum_{j=0}^{N}\hat{u}_j\,P_j^{(\alpha,\beta)}(t)=\Phi^T(t)\,\hat{U}=\hat{U}^T\,\Phi(t), \tag{9}$$

where the vectors $\hat{U}$ and $\Phi(t)$ are given by

$$\hat{U}=[\hat{u}_0,\hat{u}_1,\ldots,\hat{u}_N]^T,\qquad \Phi(t)=[P_0^{(\alpha,\beta)}(t),P_1^{(\alpha,\beta)}(t),\ldots,P_N^{(\alpha,\beta)}(t)]^T. \tag{10}$$

Similarly, any continuous two-variable function, say $g(t,s)$, defined on $[0,1]\times[0,1]$, can be expanded as follows in terms of the double-shifted Jacobi polynomials:

$$g_N(t,s)=\sum_{i=0}^{N}\sum_{j=0}^{N} g_{ij}\,P_i^{(\alpha,\beta)}(t)\,P_j^{(\alpha,\beta)}(s)=\Phi^T(t)\,G\,\Phi(s),$$

where G is a $(N+1)\times(N+1)$ matrix and its entries are given by

$$g_{ij}=\frac{1}{h_i\,h_j}\int_0^1\int_0^1 g(t,s)\,P_i^{(\alpha,\beta)}(t)\,P_j^{(\alpha,\beta)}(s)\,w^{(\alpha,\beta)}(t)\,w^{(\alpha,\beta)}(s)\,dt\,ds,\ \ i,j=0,1,\ldots,N. \quad (11)$$

The following relations hold for the shifted Jacobi polynomials:

(i) $P_i^{(\alpha,\beta)}(0)=(-1)^i\begin{pmatrix} i+\beta \\ i \end{pmatrix}$,

(ii) $\frac{d^i P_n^{(\alpha,\beta)}(t)}{dt^i}=\frac{\Gamma(n+i+\alpha+\beta+1)}{\Gamma(n+\alpha+\beta+1)}P_{n-i}^{(\alpha+i,\beta+i)}(t),\quad i=0,1,\ldots.$

4.1 The Jacobi operational matrices

In implementing operations on the Jacobi basis, we frequently encounter the integration of the vector $\Phi(t)$ defined in (10), as well as the need to evaluate the product of two vectors, $\Phi(t)$ and $\Phi^T(t)$ (the product matrix). To address this, we will derive the corresponding operational matrices. To proceed, some lemmas regarding the shifted Jacobi polynomials are necessary. These are as follows:

Lemma 1. The shifted Jacobi polynomial $P_i^{(\alpha,\beta)}(t)$, $t\in[0,1]$, can be presented in the form

$$P_i^{(\alpha,\beta)}(t)=\sum_{j=0}^{i}\gamma_j^{(i)}\,t^j,\quad i=0,1,\ldots,$$

where the coefficients $\gamma_j^{(i)}$ are calculated as

$$\gamma_j^{(i)}=(-1)^{i-j}\begin{pmatrix} i+j+\alpha+\beta \\ j \end{pmatrix}\begin{pmatrix} i+\beta \\ i-j \end{pmatrix}.$$

Proof. The coefficients $\gamma_j^{(i)}$ can be obtained as

$$\gamma_j^{(i)} = \frac{1}{j!}\frac{d^j}{dt^j}\, P_i^{(\alpha,\beta)}(t)\Big|_{t=0}.$$

Now, using the relations (i) and (ii) stated in the above, the lemma can be proved. □

Lemma 2. If $i \in \mathbb{N}$ and $l \geqslant i$, then

$$\int_0^1 t^l\, P_i^{(\alpha,\beta)}(t)\, w^{(\alpha,\beta)}(t)\, dt$$
$$= \sum_{k=0}^{i} \frac{(-1)^{i-k}\,\Gamma(i+\beta+1)\,\Gamma(i+k+\alpha+\beta+1)\,\Gamma(l+k+\beta+1)\,\Gamma(\alpha+1)}{\Gamma(k+\beta+1)\,\Gamma(i+\alpha+\beta+1)\,\Gamma(l+k+\alpha+\beta+2)\,(i-k)!\,k!}.$$

Proof. The lemma can be easily proved by integrating of (7). □

In FIDEs, such as (3), one of the derivative orders, either ν or γ, may be an integer. Therefore, both the operational matrices for the integration of integer and fractional orders need to be applied to approximate the main equation. Specifically, the integral operational matrix of integer order is used to approximate the integral component of the given equation. As a result, Jacobi operational matrices for integration are derived for both cases. The entries of these matrices are computed using the following theorems.

Theorem 2. Let $\Phi(t)$ be the Jacobi vector in (10) and $\nu \in \mathbb{R}$. Then, one has

$$I^\nu \Phi(t) \approx \mathbf{P}^{(\nu)}\, \Phi(t),$$

where I^ν is the Riemann–Liouville fractional integral operator of order ν and $\mathbf{P}^{(\nu)}$ is the $(N+1)\times(N+1)$ fractional operational matrix of integration and is defined by

$$\mathbf{P}^{(\nu)} = \begin{bmatrix} \theta(0,0) & \theta(0,1) & \dots & \theta(0,N) \\ \theta(1,0) & \theta(1,1) & \dots & \theta(1,N) \\ \vdots & \vdots & \ddots & \vdots \\ \theta(N,0) & \theta(N,1) & \dots & \theta(N,N) \end{bmatrix},$$

where

$$\theta(i,j) = \sum_{k=0}^{i} \omega'_{ijk}, \qquad i = 0,1,\dots,N, \quad j = 1,2,\dots,N, \tag{12}$$

and ω'_{ijk} are given by

$$\omega'_{ijk} = \frac{(-1)^{i-k}\ \Gamma(j+\beta+1)\ \Gamma(i+\beta+1)\ \Gamma(i+k+\alpha+\beta+1)\ \Gamma(\alpha+1)}{h_j\ \Gamma(j+\alpha+\beta+1)\ \Gamma(k+\beta+1)\ \Gamma(i+\alpha+\beta+1)\ \Gamma(k+\nu+1)\ (i-k)!}$$
$$\times \sum_{l=0}^{j} \frac{(-1)^{j-l}\ \Gamma(j+l+\alpha+\beta+1)\ \Gamma(l+k+\nu+\beta+1)}{\Gamma(l+\beta+1)\ \Gamma(l+k+\nu+\alpha+\beta+2)\ l!(j-l)!}, \quad i,j=0,1,\ldots,N.$$

Proof. Applying fractional integral operator (1) to relation (7) leads to

$$I^{\nu} P_i^{(\alpha,\beta)}(t) = \sum_{k=0}^{i} \frac{(-1)^{i-k}\ \Gamma(i+\beta+1)\ \Gamma(i+k+\alpha+\beta+1)\ t^{k+\nu}}{\Gamma(k+\beta+1)\ \Gamma(i+\alpha+\beta+1)\ \Gamma(k+\nu+1)\ (i-k)!}. \tag{13}$$

Moreover, $t^{k+\nu}$ can be approximated in terms of the shifted Jacobi polynomials as the following form:

$$t^{k+\nu} \approx \sum_{j=0}^{N} a_{k,j}\ P_j^{(\alpha,\beta)}(t),$$

where

$$a_{k,j} = \frac{1}{h_j} \int_0^1 t^{k+\nu}\ P_j^{(\alpha,\beta)}(t)\ w^{(\alpha,\beta)}(t)\ dt.$$

According to Lemma 2, (13) can be rewritten as

$$\begin{aligned} &I^{\nu} P_i^{(\alpha,\beta)}(t) \\ &\approx \sum_{j=0}^{N} \left\{ \sum_{k=0}^{i} \frac{(-1)^{i-k}\ \Gamma(j+\beta+1)\ \Gamma(i+\beta+1)\ \Gamma(i+k+\alpha+\beta+1)\ \Gamma(\alpha+1)}{h_j\ \Gamma(j+\alpha+\beta+1)\ \Gamma(k+\beta+1)\ \Gamma(i+\alpha+\beta+1)\ \Gamma(k+\nu+1)\ (i-k)!} \right. \\ &\left. \times \sum_{l=0}^{j} \frac{(-1)^{j-l}\ \Gamma(j+l+\alpha+\beta+1)\ \Gamma(l+k+\nu+\beta+1)}{\Gamma(l+\beta+1)\ \Gamma(l+k+\nu+\alpha+\beta+2)\ l!(j-l)!} \right\} P_j^{(\alpha,\beta)}(t) \\ &= \sum_{j=0}^{N} \theta(i,j)\ P_j^{(\alpha,\beta)}(t), \end{aligned}$$

where $\theta(i,j)$ are given in (12). Rewriting the last relation as a vector form gives

$$I^{\nu} P_i^{(\alpha,\beta)}(t) = [\theta(i,0), \theta(i,1), \ldots, \theta(i,N)]\ \Phi(t), \qquad i = 0,1,\ldots,N.$$

This leads to the desired result. □

Theorem 3. Let $\Phi(t)$ be the Jacobi vector in (10). Then,

$$\int_0^t \Phi(t')\ dt' \approx \mathbf{P}\ \Phi(t),$$

where $\mathbf{P}$ is the $(N+1)\times(N+1)$ operational matrix of integration (of integer order) and is defined by

$$\mathbf{P} = \begin{bmatrix} \pi(0,0) & \pi(0,1) & \dots & \pi(0,N) \\ \pi(1,0) & \pi(1,1) & \dots & \pi(1,N) \\ \vdots & \vdots & \ddots & \vdots \\ \pi(N,0) & \pi(N,1) & \dots & \pi(N,N) \end{bmatrix},$$

where

$$\pi(i,j) = \sum_{k=0}^{i} \omega_{ijk}, \qquad i,j = 0,1,\dots,N, \tag{14}$$

and ω_{ijk} are given by

$$\omega_{ijk} = \frac{(-1)^{i-k}\ \Gamma(j+\beta+1)\ \Gamma(i+\beta+1)\ \Gamma(i+k+\alpha+\beta+1)\ \Gamma(\alpha+1)}{h_j\ \Gamma(j+\alpha+\beta+1)\ \Gamma(k+\beta+1)\ \Gamma(i+\alpha+\beta+1)\ (k+1)!\ (i-k)!} \times \sum_{l=0}^{j} \frac{(-1)^{j-l}\ \Gamma(j+l+\alpha+\beta+1)\ \Gamma(l+k+\beta+2)}{\Gamma(l+\beta+1)\ \Gamma(l+k+\alpha+\beta+3)\ l!(j-l)!}, \quad i,j = 0,1,\dots,N.$$

Proof. The proof is almost the same as that presented in Theorem 2. By setting $\nu = 1$, in the proof of Theorem 2, the matrix $\mathbf{P}$ will be obtained. □

In the following, some useful and applicable lemmas are presented to obtain the Jacobi operational matrix of the product.

Lemma 3. If $P_j^{(\alpha,\beta)}(t)$ and $P_k^{(\alpha,\beta)}(t)$ are, respectively, jth and kth shifted Jacobi polynomials, then the product of $P_j^{(\alpha,\beta)}(t)$ and $P_k^{(\alpha,\beta)}(t)$ can be written as

$$Q_{j+k}^{(\alpha,\beta)}(t) = P_j^{(\alpha,\beta)}(t) P_k^{(\alpha,\beta)}(t) = \sum_{r=0}^{j+k} \lambda_r^{(j,k)}\ t^r,$$

where the coefficients $\lambda_r^{(j,k)}$ are determined as follows:

The quantities $\gamma_l^{(k)}$ and $\gamma_{r-l}^{(j)}$ are introduced, respectively, for $P_k^{(\alpha,\beta)}(t)$ and $P_j^{(\alpha,\beta)}(t)$ based on Lemma 1.

Proof. See [45, p. 496, Lemma 3] . □

Lemma 4. If $P_i^{(\alpha,\beta)}(t)$, $P_j^{(\alpha,\beta)}(t)$, and $P_k^{(\alpha,\beta)}(t)$ are, respectively, ith, jth, and kth shifted Jacobi polynomials, then

$$q_{ijk} = \int_0^1 P_i^{(\alpha,\beta)}(t)\ P_j^{(\alpha,\beta)}(t)\ P_k^{(\alpha,\beta)}(t)\ w^{(\alpha,\beta)}(t)\ dt$$

If $j \geq k$:

$r = 0, 1, \ldots, j + k,$

if $r > j$ then

$\lambda_r^{(j,k)} = \sum_{l=r-j}^{k} \gamma_{r-l}^{(j)} \gamma_l^{(k)},$

else

$r_1 = \min\{r, k\},$

$\lambda_r^{(j,k)} = \sum_{l=0}^{r_1} \gamma_{r-l}^{(j)} \gamma_l^{(k)},$

end.

If $j < k$:

$r = 0, 1, \ldots, j + k,$

if $r \leq j$ then

$r_1 = \min\{r, j\},$

$\lambda_r^{(j,k)} = \sum_{l=0}^{r_1} \gamma_{r-l}^{(j)} \gamma_l^{(k)},$

else

$r_2 = \min\{r, k\},$

$\lambda_r^{(j,k)} = \sum_{l=r-j}^{r_2} \gamma_{r-l}^{(j)} \gamma_l^{(k)},$

end.

$$= \sum_{l=0}^{j+k} \sum_{m=0}^{i} \frac{(-1)^{i-m} \lambda_l^{(j,k)} \Gamma(i+\beta+1) \Gamma(i+m+\alpha+\beta+1) \Gamma(l+m+\beta+1) \Gamma(\alpha+1)}{\Gamma(m+\beta+1) \Gamma(i+\alpha+\beta+1) \Gamma(l+m+\alpha+\beta+2) (i-m)! \, m!},$$

where $\lambda_l^{(j,k)}$ is computed by Lemma 3.

Proof. See [46, p. 12, Lemma 5]. □

The following theorem presents a general formula for finding the $(N+1) \times (N+1)$ operational matrix of product $\widetilde{V}$ whenever

$$\Phi(t) \, \Phi^T(t) \, V \approx \widetilde{V} \, \Phi(t), \tag{15}$$

and V is a given $(N+1)$ vector.

Theorem 4. The entries of the matrix $\widetilde{V}$ in (15) are computed as follows:

$$\widetilde{V}_{jk} = \frac{1}{h_k} \sum_{i=0}^{N} V_i \, q_{ijk}, \qquad j, k = 0, 1, \ldots, N,$$

where q_{ijk} is introduced by Lemma 4 and V_i is the components of the vector V in (15).

Proof. See [46, p. 12, Theorem 2]. □

Remark 1. If $u(t) \approx U^T\ \Phi(t) = \Phi^T(t)\ U$ and $v(t) \approx V^T\ \Phi(t) = \Phi^T(t)\ V$, where U and V are $(N+1)$ vectors, using the operational matrix of product, it is easily shown that

$$\begin{aligned} u^2(t) &\approx U^T\ \widetilde{U}\ \Phi(t), \\ u^3(t) &\approx F^T\ \widetilde{U}\ \Phi(t), \qquad F = \widetilde{U}^T\ U, \\ u(t)\ v(t) &\approx U^T\ \widetilde{V}\ \Phi(t). \end{aligned}$$

The following remark is useful to approximate the integral parts of equations under study.

Remark 2. Let the vector $\Phi(t)$ be the shifted Jacobi vector in (10). Then

$$\mathbf{M} = \int_0^1 \Phi(t)\ \Phi^T(t)\ dt,$$

where

$$\mathbf{M}_{i,j} = \sum_{r=0}^{i+j} \frac{\lambda_r^{(i,j)}}{r+1}, \quad i,j = 0,1,\ldots,N,$$

and $\lambda_r^{(j,k)}$ is introduced by Lemma 3.

4.2 Methodology

To continue with the implementation of the proposed method, consider the following cases.

4.2.1 Case I

$$\begin{aligned} D^\nu y(t) + h(t)\ y(t) + \int_0^t k(t,s)\ y(s)\ ds \\ + \int_0^t \tilde{k}(t,s)\ D^\gamma y(s)\ ds = f(t), \qquad 0 \leqslant t \leqslant 1, \quad 0 < \gamma < \nu \leqslant m, \end{aligned} \tag{16}$$

with initial conditions,

$$y^{(j)}(0) = d_j, \qquad j = 0, 1, \ldots, m-1,$$

$h(t)$, $k(t,s)$, and $\tilde{k}(t,s)$ are known continuous functions. Since $D^\nu y(t)$ has the highest order of derivative in (16), it can be approximated as

$$D^\nu y(t) \approx \Phi^T(t)\ C, \tag{17}$$

where the vectors $\Phi(t)$ is defined by (10) and $C = [c_0, c_1, \ldots, c_N]$ is the vector of unknown coefficients. By applying the fractional integral operator of order ν to (17), an approximation of unknown function $y(t)$ is resulted as follows:

$$\begin{aligned} y(t) &\approx \Phi^T(t)\ \mathbf{P}^{(\nu)\mathbf{T}}\ C + \sum_{j=0}^{m-1} \frac{d_j\ t^j}{j!} \\ &\approx \Phi^T(t)\ \mathbf{P}^{(\nu)\mathbf{T}}\ C + \Phi^T(t)\ F_1, \end{aligned} \tag{18}$$

where F_1 is a known $(N+1)$ vector and computed as

$$\sum_{j=0}^{m-1} \frac{d_j\ t^j}{j!} \approx \Phi^T(t) F_1,$$

$$(F_1)_k = \frac{1}{h_k} \sum_{j=0}^{m-1} \frac{d_j}{j!} \int_0^1 t^j\ P_k^{(\alpha,\beta)}(t) w^{(\alpha,\beta)}(t) dt, \qquad k = 0, 1, \ldots, N.$$

Now, using approximation (17), the following steps can be pursued to approximate the term $D^\gamma y(t)$, $\gamma < \nu$:

$$D^\nu y(t) = D^{\nu-\gamma} D^\gamma\ y(t) \approx \Phi^T(t)\ C.$$

Applying the fractional integral operator of order $\nu - \gamma$ on both sides of the last approximation yields the following approximation:

$$\begin{aligned} D^\gamma\ y(t) &\approx \Phi^T(t)\ \mathbf{P}^{(\nu-\gamma)\mathbf{T}}\ C + \sum_{j=\lceil\gamma\rceil}^{m-1} \frac{d_j\ t^{j-\gamma}}{\Gamma(j-\gamma+1)} \\ &\approx \Phi^T(t)\ \mathbf{P}^{(\nu-\gamma)\mathbf{T}}\ C + \Phi^T(t)\ F_2, \end{aligned} \tag{19}$$

where

$$\sum_{j=\lceil\gamma\rceil}^{m-1} \frac{d_j\ t^{j-\gamma}}{\Gamma(j-\gamma+1)} \approx \Phi^T(t) F_2,$$

$$(F_2)_k = \frac{1}{h_k} \sum_{j=\lceil\gamma\rceil}^{m-1} \frac{d_j}{\Gamma(j-\gamma+1)} \int_0^1 t^{j-\gamma}\ P_k^{(\alpha,\beta)}(t) w^{(\alpha,\beta)}(t) dt, \qquad k = 0, 1, \ldots, N.$$

The kernels of integral parts, $k(t,s)$ and $\tilde{k}(t,s)$, are approximated as follows by means of (11):

$$k(t,s) \approx \Phi^T(t)\ K\ \Phi(s), \qquad \tilde{k}(t,s) \approx \Phi^T(t)\ \widetilde{K}\ \Phi(s),$$

where K and $\widetilde{K}$ are the known $(N+1)\times(N+1)$ matrices. Using the introduced matrices, the integral parts of (16) are approximated as follows:

$$\begin{aligned}
\int_0^t k(t,s)\ y(s)\ ds &\approx \int_0^t \Phi^T(t)\ K\ \Phi(s)\ \Phi^T(s)\ U_1\ ds \\
&\approx \Phi^T(t)\ K \int_0^t \Phi(s)\ \Phi^T(s)\ U_1\ ds \\
&= \Phi^T(t)\ K\ \widetilde{U}_1\ \mathbf{P}\ \Phi(t), \qquad U_1 = \mathbf{P}^{(\nu)\mathbf{T}}\ C + F_1, \\
\int_0^t \tilde{k}(t,s)\ D^\gamma\ y(s)\ ds &\approx \Phi^T(t)\ \widetilde{K}\ \widetilde{U}_2\ \mathbf{P}\ \Phi(t), \qquad U_2 = \mathbf{P}^{(\nu-\gamma)\mathbf{T}}\ C + F_2,
\end{aligned} \tag{20}$$

where $\widetilde{U}_1, \widetilde{U}_2$ are operational matrices of product, corresponding to the vectors U_1, U_2, respectively. In this way, the unknown function and its derivatives are approximated after substituting approximations (17)–(20) into (16), the main equation is converted into the following algebraic equation:

$$\Phi^T(t)\ C + h(t)\ \Phi^T(t)\ U_1 + \Phi^T(t)\ K\ \widetilde{U}_1\ \mathbf{P}\ \Phi(t) + \Phi^T(t)\ \widetilde{K}\ \widetilde{U}_2\ \mathbf{P}\ \Phi(t) - f(t) \approx 0. \tag{21}$$

Lemma 5. Let $y_N(t)$ be the approximate solution obtained from the proposed scheme, and let $y(t)$ be the exact solution to (16). Suppose that $0 < \|h\|\Gamma(\nu+1) + \frac{M+\widetilde{M}}{\Gamma(\nu+2)} < 1$, where $\tilde{y}(t)$ represents the error of $y_N(t)$, and $\mathcal{H}_N(t)$ is the perturbation term. Then, the following equation holds:

$$\|\tilde{y}\| \leq \Lambda^*\ \|\mathcal{H}_n\|,$$

where Λ^* is a positive constant.

Proof. As can be seen, applying the Riemann–Liouville integral operator to (16) results in (3). it is clear that $y_N(t)$ and $y_N(t)+\tilde{y}(t)$ satisfy the following equations:

$$\begin{aligned} y_N(t) = F(t) + \mathcal{H}_N(t) &- \frac{1}{\Gamma(\nu)} \int_0^t (t-s)^{\nu-1} h(s) y_N(s)\, ds \\ &- \frac{1}{\Gamma(\nu)} \int_0^t (t-s)^{\nu-1} \int_0^s k(s,r) y_N(r)\, dr\, ds \\ &- \frac{1}{\Gamma(\nu)} \int_0^t (t-s)^{\nu-1} \int_0^s \tilde{k}(s,r) D^{\gamma} y_N(r)\, dr\, ds, \end{aligned} \tag{22}$$

$$\begin{aligned} y_N(t) + \tilde{y}(t) = F(t) &- \frac{1}{\Gamma(\nu)} \int_0^t (t-s)^{\nu-1} h(s)(y_N(s)+\tilde{y}(s))\, ds \\ &- \frac{1}{\Gamma(\nu)} \int_0^t (t-s)^{\nu-1} \int_0^s k(s,r)(y_N(r)+\tilde{y}(r))\, dr\, ds \\ &- \frac{1}{\Gamma(\nu)} \int_0^t (t-s)^{\nu-1} \int_0^s \tilde{k}(s,r) D^{\gamma}(y_N(r)+\tilde{y}(r))\, dr\, ds, \end{aligned} \tag{23}$$

where

$$F(t) = \sum_{i=0}^{m-1} \frac{d_i\, t^i}{\Gamma(i+1)} + \frac{1}{\Gamma(\nu)} \int_0^t (t-s)^{\nu-1} f(s)\, ds.$$

Subtracting (22) from (23) leads to the following equation:

$$\begin{aligned} \tilde{y}(t) = &- \frac{1}{\Gamma(\nu)} \int_0^t (t-s)^{\nu-1} h(s)(y_N(s)+\tilde{y}(s))\, ds \\ &- \frac{1}{\Gamma(\nu)} \int_0^t (t-s)^{\nu-1} \int_0^s k(s,r)(y_N(r)+\tilde{y}(r))\, dr\, ds \\ &- \frac{1}{\Gamma(\nu)} \int_0^t (t-s)^{\nu-1} \int_0^s \tilde{k}(s,r) D^{\gamma}(y_N(r)+\tilde{y}(r))\, dr\, ds. \end{aligned} \tag{24}$$

Taking the norm from (24) yields the following inequality:

$$\begin{aligned} \|\tilde{y}\| &\le \|\mathcal{H}_N\| + \frac{\|h\|\,\|\tilde{y}\|}{\Gamma(\nu)} \int_0^t (t-s)^{\nu-1} ds + \frac{M\,\|\tilde{y}\|}{\Gamma(\nu)} \int_0^t (t-s)^{\nu-1} ds + \frac{\widetilde{M}\,L\,\|\tilde{y}\|}{\Gamma(\nu)} \int_0^t (t-s)^{\nu-1} ds \\ &\le \mathcal{H}_N\| + \frac{\|h\|\,\|\tilde{y}\|}{\Gamma(\nu+1)} + fracM\,\|\tilde{y}\|\Gamma(\nu+2) + \frac{\widetilde{M}\,L\,\|\tilde{y}\|}{\Gamma(\nu+2)}. \end{aligned}$$

Since $0 < \|h\|\Gamma(\nu+1) + (M+\widetilde{M})/\Gamma(\nu+2) < 1$, so one has

$$\|\tilde{y}\| \le \frac{1}{1 - \frac{\|h\|}{\Gamma(\nu+1)} - \frac{M+\widetilde{M}}{\Gamma(\nu+2)}} \|\mathcal{H}_N\|,$$

and the desired result is acquired.

Thus, the variation of the approximate solution is bounded. Since the existence and uniqueness of the given equation have already been established in section 3, it follows that (16) is well-posed. □

4.2.2 Case II

$$y''(t) + h(t)\ y'(t) + g(t)\ D^{\nu}y(t) + \int_0^t k(t,s)\ y(s)\ ds = f(t),\ 0 \leqslant t \leqslant 1,\ 0 < \nu < 2, \quad (25)$$

subject to initial conditions

$$y^{(j)}(0) = d_j, \qquad j = 0, 1,$$

where $h(t)$, $g(t)$, $k(t,s)$, and $\tilde{k}(t,s)$ are known continuous functions. Since $y''(t)$ has the highest order of derivative in (25), successive integrating leads to the following approximations:

$$\begin{aligned} &y''(t) \approx \Phi^T(t)\ C, \\ &y'(t) \approx \Phi^T(t)\ \mathbf{P}^T\ C + \Phi^T(t)\ F_1, \qquad d_1 \approx \Phi^T(t)\ F_1, \\ &y(t) \approx \Phi^T(t)\ (\mathbf{P}^T)^2\ C + \Phi^T(t)\ \mathbf{P}^T\ F_1 + \Phi^T(t)\ F_2, \qquad d_0 \approx \Phi^T(t)\ F_2, \end{aligned} \quad (26)$$

where $\mathbf{P}$ is the operational matrix of integration of integer order, as introduced in Theorem 3. To approximate $D^{\nu}y(t)$, where $\nu \in \mathbb{R}$ is a noninteger value, the first approximation in (26) can be used as follows:

$$D^2 y(t) = D^{2-\nu} D^{\nu} y(t) \approx \Phi^T(t)\ C.$$

So,

$$\begin{aligned} D^{\nu}y(t) &\approx \Phi^T(t)\ \mathbf{P}^{(\mathbf{2}-\nu)\mathbf{T}}\ C + \sum_{i=\lceil \nu \rceil}^{m-1} \frac{d_i\ t^{i-\nu}}{\Gamma(i-\nu+1)} \\ &\approx \Phi^T(t)\ \mathbf{P}^{(\mathbf{2}-\nu)\mathbf{T}}\ C + \Phi^T(t)\ F_3, \qquad m-1 < \nu \leqslant m, \end{aligned} \quad (27)$$

where $\mathbf{P}^{(\mathbf{2}-\nu)}$ is the operational matrix of fractional integration related to the fractional integral operator of order $2-\nu$. The integral part of (25) is converted to the following algebraic approximation:

$$\int_0^t k(t,s)\ y(s)\ ds \approx \Phi^T(t)\ K\ \widetilde{U}\ \mathbf{P}\ \Phi(t), \qquad U = (\mathbf{P}^T)^2\ C + \mathbf{P}^T\ F_1 + F_2, \tag{28}$$

where $\widetilde{U}$ is the operational matrix of product, corresponding to the vector U. Substituting the approximations (26)–(28) into (25) yields the following algebraic equation:

$$\Phi^T(t)\ C + h(t)\ \Phi^T(t)\{\mathbf{P}^T\ C + F_1\} + g(t)\ \Phi^T(t)\{\mathbf{P}^{(2-\nu)}T\ C + F_3\} + \Phi^T(t)\ K\ \widetilde{U}\ \mathbf{P}\ \Phi(t) - f(t) \approx 0. \tag{29}$$

4.2.3 Case III

$$D^\nu y(t) + h(t)\ y(t) + \lambda_1 \int_0^t k(t,s)\ y^2(s)\ ds + \lambda_2 \int_0^1 \tilde{k}(t,s)\ y^2(s)\ ds = f(t), \quad 0 \leqslant t \leqslant 1, \quad \nu \in \mathbb{R}, \tag{30}$$

subject to initial conditions

$$y^{(j)}(0) = d_j, \qquad j = 0, 1, \ldots, m-1,$$

where $\lambda_1, \lambda_2 \in \mathbb{R}$ and $h(t)$, $k(t,s)$, and $\tilde{k}(t,s)$ are known continuous functions. Following the process the same as for the last two cases, leads to the approximations below:

$$\begin{aligned}
&D^\nu y(t) \approx \Phi^T(t)\ C, \quad y(t) \approx \Phi^T(t)\ \mathbf{P}^{(\nu)\mathbf{T}}\ C + \Phi^T(t)\ F_1 = \Phi^T(t)\ U,\\
&y^2(t) \approx U^T\ \widetilde{U}\ \Phi(t) = \Phi^T(t)\ F_2, \quad \int_0^t k(t,s)\ y^2(s)\ ds \approx \Phi^T(t)\ K\ \widetilde{F}_2\ \mathbf{P}\ \Phi(t),\\
&\int_0^1 \widetilde{k}(t,s)\ y^2(s)\ ds \approx \Phi^T(t)\ \widetilde{K}\ \mathbf{M}\ \widetilde{F}_2,
\end{aligned} \tag{31}$$

where $\mathbf{M}$ is the matrix introduced by Remark 2. Substituting the approximations (31) into (30) yields the following algebraic equation:

$$\Phi^T(t)\ C + h(t)\ \Phi^T(t)\ U + \lambda_1 \Phi^T(t)\ K\ \widetilde{F}_2\ \mathbf{P}\ \Phi(t) + \lambda_2 \Phi^T(t)\ \widetilde{K}\ \mathbf{M}\ F_2 - f(t) \approx 0. \tag{32}$$

The resulting algebraic equations (21), (29), and (32) are collocated at the $(N+1)$ roots of the $(N+1)$th shifted Jacobi polynomials on the interval $[0,1]$. By solving the systems of generated algebraic equations, the unknown vector C can be determined. Thus, by selecting appropriate values for the parameters α and β, an approximate solution can be obtained using the approximation in (18), the third approximation in (26), and the second ap-

proximation in (31). The resulting nonlinear systems can be solved using the Newton iteration method.

Remark 3. Based on the discussions in this section and the statement made in Remark 2, the nonlinear terms and integral components in the equations under study are approximated in vector form. Therefore, the initial conditions of the equations must be approximated in terms of the Jacobi basis.

5 Convergence analysis and error bounds

In this section, some upper bounds are presented for the residual function and the proposed algorithm. It can be seen that the upper bounds decrease when the number of terms in the series solution increases. For this purpose, some useful definitions and theorems are stated.

The set of all algebraic polynomials of degree at most N is denoted by $\mathbb{P}_N$. The orthogonal projection $\mathcal{P}_{N,\alpha,\beta} : L^2_{w^{(\alpha,\beta)}}(J) \to \mathbb{P}_N$, $J = [0,1]$, is considered for $u(t) \in L^2_{w^{(\alpha,\beta)}}(J)$ and defined by

$$(\mathcal{P}_{N,\alpha,\beta}u - u,\ v) = 0, \qquad \text{for all } v \in \mathbb{P}_N.$$

The Jacobi-weighted Sobolev space is introduced as

$$\mathcal{J}^r_{w^{(\alpha,\beta)}}(J) = \{u|\ \text{u is measurable and } \|u\|_{r,w^{(\alpha,\beta)}} < \infty\}, \qquad r \in \mathbb{N},$$

equipped with the following norm and semi-norm:

$$\|u\|_{r,w^{(\alpha,\beta)}} = \left(\sum_{k=0}^{r} \| \frac{d^k u}{dt^k} \|^2_{w^{(\alpha+k,\beta+k)}}\right)^{\frac{1}{2}}, \qquad |u|_{r,w^{(\alpha,\beta)}} = \|\frac{d^r u}{dt^r}\|_{w^{(\alpha+r,\beta+r)}}.$$

Theorem 5. For any $u \in \mathcal{J}^r_{w^{(\alpha,\beta)}}(J)$, $r \in \mathbb{N}$, $0 \leqslant \mu \leqslant r$, one has

$$\|u - \mathcal{P}_{N,\alpha,\beta}u\|_{\mu,w^{(\alpha,\beta)}} \leqslant c(N(N+\alpha+\beta))^{\frac{\mu-r}{2}} |u|_{r,w^{(\alpha,\beta)}},$$

where c is a positive constant independent of N, α, and β.

Proof. See [17, p. 5, Theorem 2.1] . □

Corollary 1. If $k(t,s)$ is a continuous function on $J \times J$ and $k_N(t,s) = \mathcal{P}_{N,\alpha,\beta}k(t,s)$ is an approximation to $k(t,s)$ in the Jacobi-weighted Sobolev

space, then the error can be bounded as follows:

$$\|k(t,s)-k_N(t,s)\|_{\mu,W^{(\alpha,\beta)}} \leqslant c_0(N(N+\alpha+\beta))^{\mu-r}|k(t,s)|_{r,W^{(\alpha,\beta)}}, \quad (33)$$

where $W^{(\alpha,\beta)}(t,s) = w^{(\alpha,\beta)}(t)w^{(\alpha,\beta)}(s)$. For more details, the interested reader is referred to . 502-5041[45].

Lemma 6. If $n \in \mathbb{N}$, $u \in \mathcal{J}^r_{w^{(\alpha,\beta)}}(J)$, and $D^n u(t)$ is the nth derivative of $u(t)$, then

$$\begin{aligned}&\|D^n u - D^n \mathcal{P}_{N,\alpha,\beta}u\|_{\mu,w^{(\alpha+n,\beta+n)}}\\ &\leqslant c^*\Big((N+n)(N+n+\alpha+\beta)\Big)^{\frac{\mu-r}{2}}|u^{(n)}|_{r,w^{(\alpha+n,\beta+n)}}, \quad 0\leqslant\mu\leqslant r. \quad (34)\end{aligned}$$

Proof. By differentiating the Jacobi series solution and applying Theorem 5, the desired result is acquired. □

Lemma 7. [Generalized Hardy's inequality [8, p. 679, Lemma 3.8]] For all measurable functions $u \geqslant 0$, the following generalized Hardy's inequality:

$$\left(\int_a^b |(Tu)(t)|^q\ w_1(t)\ dt\right)^{1/q} \leqslant \rho\left(\int_a^b |u(t)|^p\ w_2(t)\ dt\right)^{1/p},$$

holds if and only if

$$\sup_{a<t<b}\left(\int_t^b w_1(t)\ dt\right)^{1/q}\left(\int_a^t w_2^{1-p'}(t)\ dt\right)^{1/p'} < \infty,$$

where $p' = \frac{p}{p-1}$, $1 < p \leqslant q < \infty$, w_1, w_2 are weight functions, and T is an integral operator of the following form:

$$(Tu)(t) = \int_a^t \bar{k}(t,s)\ u(s)\ ds,$$

where $\bar{k}(t,s)$ is a given kernel.

Theorem 6. If $\nu, \mu \in \mathbb{R}$, $\nu > 0$, $0 \leqslant \mu \leqslant r$, and $D^\nu u(t)$ denotes the Caputo fractional derivative of $u(t) \in \mathcal{J}^r_{w^{(\alpha,\beta)}}(J)$ of order ν, then

$$\begin{aligned}&\|D^\nu u - D^\nu \mathcal{P}_{N,\alpha,\beta}u\|_{\mu,w^{(\alpha+m,\beta+m)}}\\ &\leqslant \frac{\rho\ c'}{\Gamma(m-\nu)}\Big((N+m)(N+m+\alpha+\beta)\Big)^{\frac{\mu-r}{2}}|u^{(m)}|_{r,w^{(\alpha+m,\beta+m)}},\end{aligned}$$

where $m-1<\nu\leqslant m$.

Proof. Using the properties of the Riemann–Liouville fractional integral operator leads to

$$D^{\nu}u(t)=I^{m-\nu}D^{m}u(t)=\frac{1}{\Gamma(m-\nu)}\int_{0}^{t}(t-s)^{m-\nu-1}D^{m}u(s)ds.$$

So, by means of Lemma 7 and then Lemma 6, respectively, one has

$$\begin{aligned}
&\|D^{\nu}u-D^{\nu}\mathcal{P}_{N,\alpha,\beta}u\|_{\mu,w^{(\alpha+m,\beta+m)}}\\
&=\|I^{m-\nu}(D^{m}u-D^{m}\mathcal{P}_{N,\alpha,\beta}u)\|_{\mu,w^{(\alpha+m,\beta+m)}}\\
&=\frac{1}{\Gamma(m-\nu)}\left\|\int_{0}^{t}(t-s)^{m-\nu-1}(D^{m}u(s)-D^{m}\mathcal{P}_{N,\alpha,\beta}u(s))\;ds\right\|_{\mu,w^{(\alpha+m,\beta+m)}}\\
&\leqslant\frac{\rho}{\Gamma(m-\nu)}\|D^{m}u-D^{m}\mathcal{P}_{N,\alpha,\beta}u\|_{\mu,w^{(\alpha+m,\beta+m)}}\\
&\leqslant\frac{\rho\;c'}{\Gamma(m-\nu)}\Big((N+m)(N+m+\alpha+\beta)\Big)^{\frac{\mu-r}{2}}|u^{(m)}|_{r,w^{(\alpha+m,\beta+m)}}.
\end{aligned}$$

□

Now, consider (16) again. If $y_N(t)$ is an approximate solution to $y(t)$, then the residual function is as follows:

$$H_N(t)=D^{\nu}y_N(t)+h(t)\;y_N(t)+\int_0^t k_N(t,s)\;y_N(s)\;ds+\int_0^t\widetilde{k}_N(t,s)\;D^{\gamma}y_N(s)\;ds-f(t). \tag{35}$$

The next lemma estimates a bound for residual function (35), which tends to zero when N tends to infinity.

Lemma 8. Let $r\in\mathbb{N}$, $\mu\in\mathbb{R}$, and $0\leqslant\mu\leqslant r$. Let us consider the residual function given in (35). When $N\to\infty$ one has $H_N(t)\to 0$.

Proof. Subtracting (35) from (16) and setting $e_N(t)=y(t)-y_N(t)$, as error function, leads to the following equation:

$$\begin{aligned}
H_N(t)=&-D^{\nu}(y(t)-y_N(t))-h(t)\;e_N(t)\\
&+\int_0^t k(t,s)\;y(s)\;ds-\int_0^t k_N(t,s)\;y_N(s)\;ds\\
&+\int_0^t\widetilde{k}(t,s)\;D^{\gamma}y(s)\;ds-\int_0^t\widetilde{k}_N(t,s)\;D^{\gamma}y_N(s)\;ds.
\end{aligned} \tag{36}$$

After some manipulation, (36) leads to the following equation:

$$H_N(t) = -D^{\nu}(y(t) - y_N(t)) - h(t)\ e_N(t) + \int_0^t k(t,s)\ e_N(s)\ ds + \int_0^t (k(t,s) - k_N(t,s))\ y_N(s)\ ds + \int_0^t \tilde{k}(t,s) D^{\gamma}(y(s) - y_N(s))\ ds + \int_0^t (\tilde{k}(t,s) - \tilde{k}_N(t,s)) D^{\gamma} y_N(s)\ ds. \tag{37}$$

Now, a bound is computed for residual function (37) so that

$$\|H_N\|_{\mu,w^{(\alpha,\beta)}} \leqslant G_1 + G_2 + G_3 + G_4 + G_5 + G_6, \tag{38}$$

$$\|H_N\|_{\mu,w^{(\alpha,\beta)}} \leqslant G_1 + G_2 + G_3 + G_4 + G_5 + G_6, \tag{39}$$

where

$$G_1 = \|D^{\nu}(y - y_N)\|_{\mu,w^{(\alpha,\beta)}}, \qquad G_2 = \|h(t)\ e_N(t)\|_{\mu,w^{(\alpha,\beta)}},$$
$$G_3 = \left\| \int_0^t k(t,s)\ e_N(s)\ ds \right\|_{\mu,w^{(\alpha,\beta)}},$$
$$G_4 = \left\| \int_0^t (k(t,s) - k_N(t,s))\ y_N(s)\ ds \right\|_{\mu,w^{(\alpha,\beta)}},$$
$$G_5 = \left\| \int_0^t \tilde{k}(t,s) D^{\gamma}(y(s) - y_N(s))\ ds \right\|_{\mu,w^{(\alpha,\beta)}},$$
$$G_6 = \left\| \int_0^t (\tilde{k}(t,s) - \tilde{k}_N(t,s)) D^{\gamma} y_N(s)\ ds \right\|_{\mu,w^{(\alpha,\beta)}}.$$

Using Theorem 6 leads to the following bound for G_1:

$$G_1 = \|D^{\nu}(y - y_N)\|_{\mu,w^{(\alpha,\beta)}} \leqslant \frac{c_0}{\Gamma(m-\nu)} \Big((N+m)(N+m+\alpha+\beta) \Big)^{\frac{\mu-r}{2}} |y^{(m)}|_{r,w^{(\alpha+m,\beta+m)}},$$

where $m = \lceil \nu \rceil$. According to continuity of function $h(t)$ over $[0,1]$, there exists $M_h > 0$ such that $||h|| \leqslant M_h$. So, using Theorem 5 one has

$$G_2 = \|h(t)\ e_N(t)\|_{\mu,w^{(\alpha,\beta)}} \leqslant c_1\ M_h \left(N(N+\alpha+\beta) \right)^{\frac{\mu-r}{2}} |y|_{r,w^{(\alpha,\beta)}}.$$

By applying Lemma 7 and Theorem 5, a bound can be computed for G_3 as follows:

$$G_3 = \left\| \int_0^t k(t,s)\ e_N(s)\ ds \right\|_{\mu,w^{(\alpha,\beta)}} \leqslant \rho_1\ \|e_N(t)\|_{\mu,w^{(\alpha,\beta)}}$$

$$\leqslant \rho_1\ c_1\Big(N(N+\alpha+\beta)\Big)^{\frac{\mu-r}{2}}\ |y|_{r,w^{(\alpha,\beta)}}.$$

Now, define $e(k_N) = k(t,s) - k_N(t,s)$. Employing Lemma 7 and Corollary 1, respectively, provide the following bound for G_4:

$$\begin{aligned} G_4 &= \left\| \int_0^t (k(t,s) - k_N(t,s))\ y_N(s)\ ds \right\|_{\mu,w^{(\alpha,\beta)}} \leqslant \rho_2 \|e(k_N)\|_{\mu,w^{(\alpha,\beta)}} \\ &\leqslant \rho_2\ c_2\ \Big(N(N+\alpha+\beta)\Big)^{\mu-r}\ |k(t,s)|_{r,W^{(\alpha,\beta)}}. \end{aligned}$$

Again, based on Lemma 7 and Theorem 6, the following bound is obtained for G_5:

$$\begin{aligned} G_5 &= \left\| \int_0^t \tilde{k}(t,s) D^{\gamma}(y(s) - y_N(s))\ ds \right\|_{\mu,w^{(\alpha,\beta)}} \leqslant \rho_3\ \|D^{\gamma} e_N(s)\|_{\mu,w^{(\alpha,\beta)}} \\ &\leqslant \frac{\rho_3\ c_3}{\Gamma(m_1-\gamma)}\ \Big((N+m_1)(N+m_1+\alpha+\beta)\Big)^{\frac{\mu-r}{2}}\ |y^{(m_1)}|_{r,w^{(\alpha+m_1,\beta+m_1)}}, \end{aligned}$$

where $m_1 = \lceil \gamma \rceil$. Defining $e(\tilde{k}_N) = \tilde{k}(t,s) - \tilde{k}_N(t,s)$ and using Lemma 7 and Corollary 1 lead to a bound for G_6 as follows:

$$\begin{aligned} G_6 &= \left\| \int_0^t D^{\gamma} y_N(s)(\tilde{k}(t,s) - \tilde{k}_N(t,s))\ ds \right\|_{\mu,w^{(\alpha,\beta)}} \leqslant \rho_4\ \|e(\tilde{k}_N)\|_{\mu,w^{(\alpha,\beta)}} \\ &\leqslant \rho_4\ c_4\ \Big(N(N+\alpha+\beta)\Big)^{\mu-r}\ |\tilde{k}(t,s)|_{r,W^{(\alpha,\beta)}}. \end{aligned}$$

Therefore, using the above bounds, inequality (38) will be as follows:

$$\begin{aligned} \|H_N\|_{\mu,w^{(\alpha,\beta)}} \leqslant\ & \frac{c_0}{\Gamma(m-\nu)}\Big((N+m)(N+m+\alpha+\beta)\Big)^{\frac{\mu-r}{2}}\ |y^{(m)}|_{r,w^{(\alpha+m,\beta+m)}} \\ &+ (M_h+\rho_1)\ c_1\Big(N(N+\alpha+\beta)\Big)^{\frac{\mu-r}{2}} |y|_{r,w^{(\alpha,\beta)}} \\ &+ (\rho_2\ c_2|k(t,s)|_{r,W^{(\alpha,\beta)}} + \rho_4\ c_4|\tilde{k}(t,s)|_{r,W^{(\alpha,\beta)}})\Big(N(N+\alpha+\beta)\Big)^{\mu-r} \\ &+ \frac{\rho_3\ c_3}{\Gamma(m_1-\gamma)}\ \Big((N+m_1)(N+m_1+\alpha+\beta)\Big)^{\frac{\mu-r}{2}}\ |y^{(m_1)}|_{r,w^{(\alpha+m_1,\beta+m_1)}}. \end{aligned}$$

According to the continuity of the functions $y(t)$ and its derivatives, $k(t,s)$, and $\tilde{k}(t,s)$ over the compact intervals $[0,1]$ and $[0,1]\times[0,1]$, the residual function $H_N(t)$ becomes smaller as N is sufficiently large. Hence, the desired result is achieved. □

The following theorem presents an error bound for the proposed method.

Theorem 7. If $y_N(t)$ is the Jacobi approximate solution to $y(t)$, $e_N(t) = y(t) - y_N(t)$, $e(k_N) = k(t,s) - k_N(t,s)$, and $e(\tilde{k}_N) = \tilde{k}(t,s) - \tilde{k}_N(t,s)$ are the error functions, and $0 \leqslant \mu \leqslant r$, then $\|e_N\|_{\mu,w^{(\alpha,\beta)}} \to 0$ as $N \to \infty$.

Proof. It is clear that the function $y_N(t)$ is an approximate solution to problem (4), that is,

$$\begin{aligned} y_N(t) = F(t) + H_N(t) &- \frac{1}{\Gamma(\nu)} \int_0^t (t-s)^{\nu-1} h(s)\ y_N(s)\ ds \\ &- \frac{1}{\Gamma(\nu)} \int_0^t (t-s)^{\nu-1} \int_0^s k_N(s,s')\ y_N(s')\ ds'\ ds \\ &- \frac{1}{\Gamma(\nu)} \int_0^t (t-s)^{\nu-1} \int_0^s \tilde{k}_N(s,s')\ D^{\gamma} y_N(s')\ ds'\ ds. \end{aligned} \tag{40}$$

So, by subtracting (39) from (4) and after some manipulations, the following error equation is obtained:

$$\begin{aligned} e_N(t) = -H_N(t) &- \frac{1}{\Gamma(\nu)} \int_0^t (t-s)^{\nu-1} h(s)\ e_N(s)\ ds \\ &- \frac{1}{\Gamma(\nu)} \int_0^t (t-s)^{\nu-1} \int_0^s k_N(s,s')\ e_N(s')\ ds'\ ds \\ &- \frac{1}{\Gamma(\nu)} \int_0^t (t-s)^{\nu-1} \int_0^s y_N(s') e(k_N)\ ds'\ ds \\ &- \frac{1}{\Gamma(\nu)} \int_0^t (t-s)^{\nu-1} \int_0^s \tilde{k}(s,s') D^{\gamma} e_N(s')\ ds'\ ds \\ &- \frac{1}{\Gamma(\nu)} \int_0^t (t-s)^{\nu-1} \int_0^s D^{\gamma} y_N(s') e(\tilde{k}_N)\ ds'\ ds. \end{aligned} \tag{41}$$

A bound for (40) is calculated as follows:

$$\|e_N\|_{\mu,w^{(\alpha,\beta)}} \leqslant \|H_N\|_{\mu,w^{(\alpha,\beta)}} + Q_1 + Q_2 + Q_3 + Q_4 + Q_5, \tag{42}$$

where

$$\begin{aligned} Q_1 &= \frac{1}{\Gamma(\nu)} \left\| \int_0^t (t-s)^{\nu-1} h(s)\ e_N(s)\ ds \right\|_{\mu,w^{(\alpha,\beta)}}, \\ Q_2 &= \frac{1}{\Gamma(\nu)} \left\| \int_0^t (t-s)^{\nu-1} \int_0^s k_N(s,s')\ e_N(s')\ ds'\ ds \right\|_{\mu,w^{(\alpha,\beta)}}, \\ Q_3 &= \frac{1}{\Gamma(\nu)} \left\| \int_0^t (t-s)^{\nu-1} \int_0^s y_N(s') e(k_N)\ ds'\ ds \right\|_{\mu,w^{(\alpha,\beta)}}, \\ Q_4 &= \frac{1}{\Gamma(\nu)} \left\| \int_0^t (t-s)^{\nu-1} \int_0^s \tilde{k}(s,s') D^{\gamma} e_N(s')\ ds'\ ds \right\|_{\mu,w^{(\alpha,\beta)}}, \end{aligned}$$

$$Q_5 = \frac{1}{\Gamma(\nu)} \left\| \int_0^t (t-s)^{\nu-1} \int_0^s D^{\gamma} y_N(s') e(\tilde{k}_N)\ ds'\ ds \right\|_{\mu,w^{(\alpha,\beta)}}.$$

According to the continuity of the function $h(t)$, there exists M_h such that $||h|| \leqslant M_h$. Using Lemma 7 and Theorem 5, the following bounds are obtained for Q_1 and Q_2:

$$\begin{aligned} Q_1 &\leqslant \frac{1}{\Gamma(\nu)} \left\| \int_0^t (t-s)^{\nu-1} h(s)\ e_N(s)\ ds \right\|_{\mu,w^{(\alpha,\beta)}} \leqslant \frac{\rho_1}{\Gamma(\nu)} \|h(t)\ e_N(t)\|_{\mu,w^{(\alpha,\beta)}} \\ &\leqslant \frac{\rho_1\ M_h}{\Gamma(\nu)} \|e_N(t)\|_{\mu,w^{(\alpha,\beta)}} \leqslant \frac{\rho_1\ M_h\ c_0}{\Gamma(\nu)} \Big(N(N+\alpha+\beta)\Big)^{\frac{\mu-r}{2}} |y|_{r,w^{(\alpha,\beta)}}, \end{aligned}$$

$$\begin{aligned} Q_2 &\leqslant \frac{1}{\Gamma(\nu)} \left\| \int_0^t (t-s)^{\nu-1} \int_0^s k_N(s,s')\ e_N(s')\ ds'\ ds \right\|_{\mu,w^{(\alpha,\beta)}} \\ &\leqslant \frac{\rho_2}{\Gamma(\nu)} \left\| \int_0^s k_N(s,s')\ e_N(s')\ ds' \right\|_{\mu,w^{(\alpha,\beta)}} \leqslant \frac{\rho_2\ \rho_2'}{\Gamma(\nu)} \|e_N(t)\|_{\mu,w^{(\alpha,\beta)}} \\ &\leqslant \frac{\rho_2\ \rho_2'\ c_0}{\Gamma(\nu)} \Big(N(N+\alpha+\beta)\Big)^{\frac{\mu-r}{2}} |y|_{r,w^{(\alpha,\beta)}}. \end{aligned}$$

Applying Lemma 7 and Corollary 1 lead to the following bounds for Q_3 and Q_5:

$$\begin{aligned} Q_3 &= \frac{1}{\Gamma(\nu)} \left\| \int_0^t (t-s)^{\nu-1} \int_0^s y_N(s') e(k_N)\ ds'\ ds \right\|_{\mu,w^{(\alpha,\beta)}} \\ &\leqslant \frac{\rho_3}{\Gamma(\nu)} \left\| \int_0^s y_N(s') e(k_N)\ ds' \right\|_{\mu,w^{(\alpha,\beta)}} \leqslant \frac{\rho_3\ \rho_3'}{\Gamma(\nu)} \|e(k_N)\|_{\mu,w^{(\alpha,\beta)}} \\ &\leqslant \frac{\rho_3\ \rho_3'\ c_1}{\Gamma(\nu)} \Big(N(N+\alpha+\beta)\Big)^{\mu-r} |k(t,s)|_{r,W^{(\alpha,\beta)}}, \end{aligned}$$

$$\begin{aligned} Q_5 &= \frac{1}{\Gamma(\nu)} \left\| \int_0^t (t-s)^{\nu-1} \int_0^s D^{\gamma} y_N(s') e(\tilde{k}_N)\ ds'\ ds \right\|_{\mu,w^{(\alpha,\beta)}} \\ &\leqslant \frac{\rho_5}{\Gamma(\nu)} \left\| \int_0^s D^{\gamma} y_N(s') e(\tilde{k}_N)\ ds'\ ds' \right\|_{\mu,w^{(\alpha,\beta)}} \leqslant \frac{\rho_5\ \rho_5'}{\Gamma(\nu)} \|e(\tilde{k}_N)\|_{\mu,w^{(\alpha,\beta)}} \\ &\leqslant \frac{\rho_5\ \rho_5'\ c_3}{\Gamma(\nu)} \Big(N(N+\alpha+\beta)\Big)^{\mu-r} |\tilde{k}(t,s)|_{r,W^{(\alpha,\beta)}}. \end{aligned}$$

In order to compute a bound for Q_4, Lemma 7 and Theorem 6 are employed. So, one has

$$Q_4 = \frac{1}{\Gamma(\nu)} \left\| \int_0^t (t-s)^{\nu-1} \int_0^s \tilde{k}(s,s') D^\gamma e_N(s')\ ds'\ ds \right\|_{\mu,w^{(\alpha,\beta)}}$$
$$\leqslant \frac{\rho_4}{\Gamma(\nu)} \left\| \int_0^s \tilde{k}(s,s') D^\gamma e_N(s')\ ds' \right\|_{\mu,w^{(\alpha,\beta)}} \leqslant \frac{\rho_4\ \rho_4'}{\Gamma(\nu)} \|D^\gamma e_N(t)\|_{\mu,w^{(\alpha,\beta)}}$$
$$\leqslant \frac{\rho_4\ \rho_4'\ c_2}{\Gamma(m_1-\gamma)\ \Gamma(\nu)} \Big((N+m_1)(N+m_1+\alpha+\beta) \Big)^{\frac{\mu-r}{2}} |y^{(m_1)}|_{r,w^{(\alpha+m_1,\beta+m_1)}},$$

where $m_1 = \lceil \gamma \rceil$. Therefore, a bound will be computed for $\|e_N\|_{\mu,w^{(\alpha,\beta)}}$ as follows:

$$\|e_N(t)\|_{\mu,w^{(\alpha,\beta)}} \leqslant \|H_N(t)\|_{\mu,w^{(\alpha,\beta)}} + \frac{1}{\Gamma(\nu)} \Big\{ (\rho_1\ M_h + \rho_2\ \rho_2')\ c_0 \Big(N(N+\alpha+\beta) \Big)^{\frac{\mu-r}{2}} |y|_{r,w^{(\alpha,\beta)}}$$
$$+ (\rho_3\ \rho_3'\ c_1\ |k(t,s)|_{r,W^{(\alpha,\beta)}} + \rho_5\ \rho_5'\ c_3\ |\tilde{k}(t,s)|_{r,W^{(\alpha,\beta)}}) \Big(N(N+\alpha+\beta) \Big)^{\mu-r}$$
$$+ \frac{\rho_4\ \rho_4'\ c_2}{\Gamma(m_1-\gamma)} \Big((N+m_1)(N+m_1+\alpha+\beta) \Big)^{\frac{\mu-r}{2}} |y^{(m_1)}|_{r,w^{(\alpha+m_1,\beta+m_1)}} \Big\}.$$

Hence, the desired result is acquired. □

Remark 4. The convergence of methods for (25) and (30) can be proved in a similar way. In order to approximate the nonlinear terms in (30), the following way is proposed:

$$y^2(t) - y_N^2(t) = (y(t) - y_N(t))(y(t) + y_N(t)) = e_N(t)(e_N(t) + 2y_N(t)).$$

Remark 5. In practice, to evaluate the effectiveness of the proposed method, the absolute errors of the approximate solutions obtained will be computed, provided that the exact solutions are available. For problems where exact solutions do not exist, the error equation, such as (40), will be solved using the same procedure suggested for the main problem, and an estimate of the absolute error will be obtained.

6 Numerical examples

In this section, the method presented in the previous section is applied to solve several examples taken from [16, 15, 53, 30, 28, 50] for comparison purposes. The proposed method is contrasted with the Homotopy Perturbation, Sinc-Collocation, Adomian Decomposition, Jacobi-Gauss integration, and Legendre and Bernoulli Wavelet methods, as discussed in [16, 15, 53, 30, 28, 50]. A

comparison of the results is then provided. All computations were performed using Maple 13 software on a laptop equipped with an Intel R processor running at 2.20 GHz and 4.00 GB of RAM.

Example 1. As the first example, consider the following Volterra integro-differential equation of an arbitrary fractional order ν [53]:

$$D^{\nu}y(t) - t(1+2t)\int_0^t e^{s(t-s)}\ y(s)\ ds - 1 - 2t = 0, \qquad 0 \leqslant t \leqslant 1, \quad 0 < \nu \leqslant 1, \tag{43}$$

with initial condition $y(0) = 1$. If $\nu = 1$, then the exact solution is $y(t) = \exp(t^2)$. According to what expressed in section 4, the following approximations can be considered:

$$D^{\nu}y(t) \approx \Phi^T(t)\ C, \quad y(t) \approx \Phi^T(t)\ \mathbf{P}^{(\nu)\mathbf{T}}\ C + \Phi^T(t)\ F = \Phi^T(t)\ U,$$
$$e^{s(t-s)} \approx \Phi^T(t)\ K\ \Phi(s), \quad \int_0^t e^{s(t-s)}\ y(s)\ ds \approx \Phi(t)\ K\ \widetilde{U}\ \mathbf{P}\ \Phi(t),$$

where $y(0)$ is approximated as $\Phi^T(t)\ F$ and $\widetilde{U}$ is the operational matrix of product, corresponding to the vector $U = \mathbf{P}^{(\nu)\mathbf{T}}\ C + F$. Substituting the above approximations into (42) leads to the following algebraic equation:

$$\Phi^T(t)\ C - t(1+2t)\Phi(t)\ K\ \widetilde{U}\ \mathbf{P}\ \Phi(t) - 1 - 2t \approx 0. \tag{44}$$

By choosing $N = 10$, (43) is collocated at the roots of $P_{11}^{(\alpha,\beta)}(t)$. By arbitrarily selecting the values of the parameters α and β, the unknown vector C is determined by solving the resulting system. The maximum absolute errors for various values of α, β, $N = 10$, and $\nu = 1$ are listed in Table 1. Additionally, the estimated absolute errors, $Error_{Est}$, are computed and presented in Table 1. As shown, the estimated errors are in agreement with the absolute errors. (42) is solved in [53] using the Jacobi–Gauss integration method for $N = 2 : 2 : 20$ and $\nu = 1$. By choosing $N = 2 : 2 : 20$ and setting $\alpha = \beta = 0$, (43) is collocated at the roots of $P_{N+1}^{(0,0)}(t)$. By solving the resulting algebraic systems, the unknown vector C is determined, and the approximate solution can be obtained for any value of N. In Figure 1, the maximum absolute errors are plotted against N for $\alpha = \beta = 0$. It can be observed that the absolute errors decrease as N increases from 2 to 18,

but for $N = 20$, the absolute error increases, which agrees with the results reported in [53]. The numerical results are shown in Figure 2 for values of $\nu = 0.25, 0.5, 0.75, 1$, $\alpha = 1/2$, $\beta = -1/2$, and $N = 10$. The approximate solutions for various values of ν and the exact solution are presented in part (a) of Figure 2. It can be observed that the numerical solutions converge to the analytic solution as ν approaches 1. Part (b) provides a graphical comparison between the exact and approximate solutions for $\nu = 1$.

Table 1: Maximum absolute and estimated errors for various values of α, β, $N = 10$, and $\nu = 1$ for Example 1

(α, β)	$Error_{Abs}$	$Error_{Est}$	(α, β)	$Error_{Abs}$	$Error_{Est}$
$(0, 0)$	3.5146×10^{-8}	3.5187×10^{-8}	$(\frac{1}{2}, \frac{1}{2})$	7.1293×10^{-8}	7.0560×10^{-8}
$(2, -\frac{1}{2})$	1.0013×10^{-6}	1.0012×10^{-6}	$(1, 1)$	1.1774×10^{-7}	1.1786×10^{-7}
$(-\frac{1}{2}, \frac{1}{2})$	1.5371×10^{-7}	1.5389×10^{-7}	$(\frac{1}{2}, -\frac{1}{2})$	1.2362×10^{-7}	1.3083×10^{-7}
$(-\frac{1}{3}, \frac{1}{4})$	8.1979×10^{-8}	8.2075×10^{-8}	$(-\frac{1}{4}, -\frac{1}{5})$	2.5541×10^{-8}	2.5572×10^{-8}

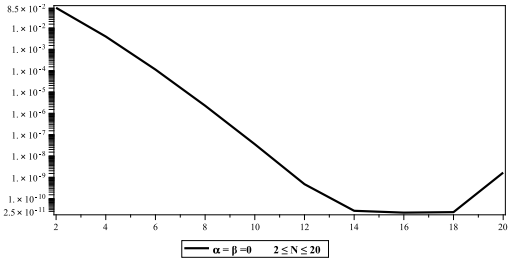


Figure 1: Maximum absolute error for different numbers of collocation points and $\alpha = \beta = 0$ for Example 1

Example 2. Consider the following linear FIDE [30]:

$$D^{\nu}y(t) - y(t) + \int_0^t y(s)\, ds - t(1+e^t) - 3e^t = 0, \quad 0 \leqslant t \leqslant 1, \quad 3 < \nu \leqslant 4, \tag{45}$$

with initial conditions $y(0) = y'(0) = 1$, $y''(0) = 2$, and $y'''(0) = 3$. If $\nu = 4$ the exact solution is $y(t) = 1 + t\exp(t)$. Substituting the approximations given in the previous example into (44) leads to the following algebraic equation:

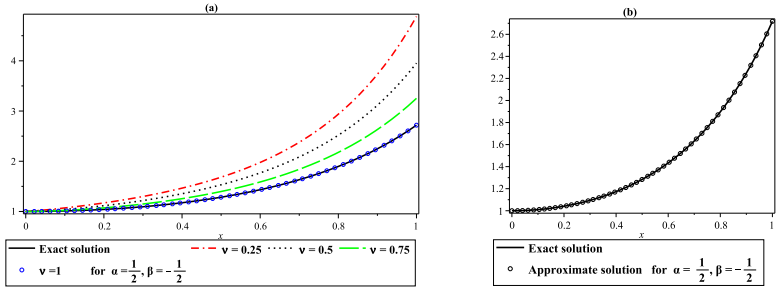


Figure 2: (a) Approximation solutions with different orders, ν, and exact solution, (b) Approximate and exact solutions for $\nu = 1$, $N = 10$, $\alpha = \frac{1}{2}$, and $\beta = -\frac{1}{2}$ for Example 1

$$\Phi^T(t)\, C - \left(\Phi^T(t)\, \mathbf{P}^{(\nu)\mathbf{T}}\, C + \Phi^T(t)\, F\right) + \Phi(t)\, K\, \widetilde{U}\, \mathbf{P}\, \Phi(t) - t(1+e^t) - 3e^t \approx 0, \tag{46}$$

where $1 + t + t^2 + t^3/2$ obtained from the initial conditions, is approximated as $\Phi^T(t)\, F$. By choosing $N = 10$ and setting $\nu = 4$ (the classical fourth-order integro-differential equation), (45) is collocated at the roots of $P_{11}^{(\alpha,\beta)}(t)$. By determining the values of the parameters α, β arbitrarily and solving the resultant algebraic system, the unknown vector C can be determined. Table 2 displays the maximum absolute errors for various values of α, β, and $N = 10$.

In Table 3, the values of the approximate and exact solutions are compared at the points $t_i = 0.2i, i = 0, 1, \ldots, 5$, for $\alpha = -1/4$, $\beta = -1/5$, and various values of ν. As shown in Table 3, the numerical values approach the exact values as $\nu \to 4$. A graphical comparison between the exact and approximate solutions is presented in Figure 3 for $\alpha = 1/3$, $\beta = 1/4$, and $N = 10$. It is evident from the figure that the numerical results converge to the exact solution as ν approaches 4.

Example 3. Consider the following nonlinear integro-differential equation with a fractional derivative of order ν [15]:

$$D^\nu y(t) + \int_0^t y^2(s)\, ds + \frac{t}{2} - \sinh(t) - \frac{1}{4}\sinh(2t) = 0, \qquad 0 \leqslant t \leqslant 1, \quad 0 < \nu \leqslant 2, \tag{47}$$

with initial conditions $y(0) = 0$ and $y'(0) = 1$. If $\nu = 2$ the exact solution is $y(t) = \sinh(t)$. By substituting the approximations given in Example 1 into (47), the following algebraic equation is obtained:

Table 2: Maximum absolute errors for various values of α, β, $N = 10$, and $\nu = 4$ for Example 2

(α, β)	$Error_{Abs}$	(α, β)	$Error_{Abs}$
$(0, 0)$	6.7806×10^{-13}	$(\frac{1}{2}, \frac{1}{2})$	1.3583×10^{-12}
$(-\frac{1}{5}, -\frac{1}{3})$	1.5960×10^{-12}	$(1, 1)$	2.2981×10^{-12}
$(-\frac{1}{2}, \frac{1}{2})$	2.6885×10^{-12}	$(\frac{1}{2}, -\frac{1}{2})$	2.5708×10^{-12}
$(\frac{1}{3}, \frac{1}{4})$	1.1678×10^{-12}	$(-\frac{1}{4}, -\frac{1}{5})$	4.8693×10^{-13}

Table 3: Approximate and exact solutions for different values of ν, $\alpha = -\frac{1}{4}$, $\beta = -\frac{1}{5}$, and $N = 10$ in Example 2

t_i	$\nu = 3.25$	$\nu = 3.50$	$\nu = 3.75$	$\nu = 3.90$	$\nu = 4$	Exact values
0.0	0.999999	0.999999	0.999999	0.999999	1.000000	1.000000
0.2	1.246742	1.245301	1.244608	1.244383	1.244280	1.244280
0.4	1.619754	1.607612	1.600653	1.598032	1.596729	1.596729
0.6	2.178540	2.136546	2.109904	2.098986	2.093271	2.093271
0.8	2.996977	2.895698	2.826706	2.796715	2.780432	2.780432
1.0	4.166138	3.965176	3.820669	3.754964	3.718281	3.718281

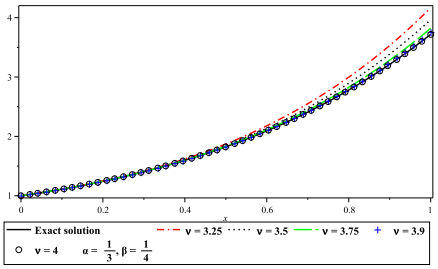


Figure 3: Approximation solutions with different values of ν and exact solution for $\alpha = \frac{1}{3}$, $\beta = \frac{1}{4}$, and $N = 10$ for Example 2

$$\Phi^T(t)\ C + \left(\Phi^T(t)\ \mathbf{P}^{(\nu)\mathbf{T}}\ C + \Phi^T(t)\ F\right) + \Phi(t)\ K\ \widetilde{U}\ \mathbf{P}\ \Phi(t) + \frac{t}{2} - \sinh(t) - \frac{1}{4}\sinh(2t) \approx 0, \tag{48}$$

where $\widetilde{U}$ is the operational matrix of the product corresponding to the vector $U = \mathbf{P}^{(\nu)\mathbf{T}}C + F$.

Table 4 displays the maximum absolute and estimated errors for various values of α, β, $\nu = 2$, and $N = 10$. The table shows that the numerical results are in good agreement with the exact solution. From Table 4, it can be observed that for cases where $\alpha = \beta$, the absolute errors decrease as the common values of α and β decrease. For $\alpha \neq \beta$, the absolute errors are smaller when $\alpha > \beta$. The estimated errors are also listed in Table 4. (47) is solved in [15] using the Homotopy Perturbation Method (HPM) and Variational Iteration Method (VIM) for $N = 2$. The results reported by [15] are depicted in part (a) of Figure 4. The plot of the approximate solution obtained by the Jacobi operational method is shown in part (b) of Figure 4 for $\alpha = \beta = 0$, $N = 2$, and $\nu = 2$. Based on Figure 4, it can be concluded that the result obtained by the Jacobi operational method is more precise than those obtained by VIM. In Table 5, the exact and approximate solutions are computed at the points $t_i = 0.2i, i = 0, 1, \ldots, 5$, for $\alpha = 1$, $\beta = -1/2$, and various values of ν. From Table 5, it can be seen that the numerical values approach the exact values as $\nu \to 2$. A graphical comparison between the exact and approximate solutions is presented in part (a) of Figure 5 for $\alpha = 1$, $\beta = -1/2$, and $N = 10$. Part (b) of Figure 5 plots the absolute error functions (i.e., the difference between the approximate results and the exact

solution for $\nu = 2$). It can be observed that the absolute errors decrease as ν approaches 2.

Table 4: Maximum absolute and estimated errors for various values of α and β, $N = 10$, and $\nu = 2$ for Example 3

(α, β)	$Error_{Abs}$	$Error_{Est}$	(α, β)	$Error_{Abs}$	$Error_{Est}$
$(0, 0)$	4.0258×10^{-14}	5.1161×10^{-12}	$(\frac{1}{2}, \frac{1}{2})$	2.3641×10^{-13}	7.2078×10^{-12}
$(1, 1)$	6.1463×10^{-13}	9.4707×10^{-12}	$(2, 2)$	1.9956×10^{-12}	1.4367×10^{-11}
$(-\frac{1}{2}, \frac{1}{2})$	3.5233×10^{-13}	1.3004×10^{-12}	$(\frac{1}{2}, -\frac{1}{2})$	1.3970×10^{-13}	1.5617×10^{-12}
$(\frac{1}{3}, \frac{1}{3})$	1.5246×10^{-13}	6.4781×10^{-12}	$(-\frac{1}{4}, -\frac{1}{3})$	4.4786×10^{-14}	3.5623×10^{-12}
$(-\frac{1}{2}, 1)$	2.7671×10^{-12}	2.2636×10^{-11}	$(1, -\frac{1}{2})$	2.1062×10^{-13}	1.3716×10^{-12}

Table 5: Approximate and exact solutions for various values of ν, $\alpha = 1$, $\beta = -\frac{1}{2}$, and $N = 10$ in Example 3

t_i	$\nu = 1.25$	$\nu = 1.50$	$\nu = 1.75$	$\nu = 1.90$	$\nu = 2$	Exact values
0.0	1.8129×10^{-6}	8.7268×10^{-7}	2.1825×10^{-7}	5.1992×10^{-8}	4.6979×10^{-15}	0.000000
0.2	0.210520	0.205396	0.202710	0.201776	0.201336	0.201336
0.4	0.450364	0.430729	0.418353	0.413346	0.410752	0.410752
0.6	0.726395	0.685527	0.656555	0.643696	0.636653	0.636653
0.8	1.043130	0.977677	0.926485	0.902061	0.888105	0.888105
1.0	1.403192	1.314333	1.237566	1.198417	1.175201	1.175201

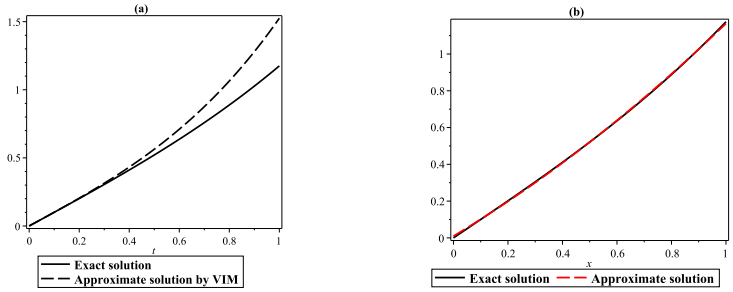


Figure 4: (a) Exact and approximate solutions obtained by VIM in Ref. [15], (b) Exact and approximate solutions obtained by Jacobi operational method for $\alpha = \beta = 0$, $N = 2$, and $\nu = 2$ for Example 3

Example 4. Consider the following mixed Volterra-Fredholm FIDE [28]:

$$D^{\nu+1}y(t) - \int_0^t (e^s + 1)\, y^2(s)\, ds - \int_0^1 ts\, y^2(s)\, ds - g(t) = 0, \qquad 0 \leqslant t \leqslant 1,\ 0 < \nu \leqslant 1, \quad (49)$$

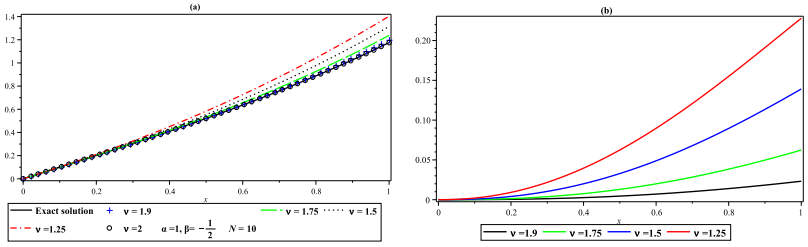


Figure 5: (a) Approximation solutions for different values of ν and exact solution, (b) Absolute error functions for $\alpha = 1$, $\beta = -\frac{1}{2}$, and $N = 10$ in Example 3

where $g(t) = \exp(t) - ((\exp(t) - t - 1)^3/3) - t(\exp(2)/4 - 2\exp(1) + 11/3)$. The initial conditions of problem are $y(0) = y'(0) = 0$ and its exact solution is $y(t) = \exp(t) - t - 1$ for $\nu = 1$. By applying the fundamental matrices, presented in section 4 for (49), the following algebraic equation is achieved:

$$\Phi^T(t)\ C - \Phi(t)\ K_1\ \widetilde{F}_2\ \mathbf{P}\ \Phi(t) - \Phi(t)\ K_2\ \mathbf{M}\ F_2 - g(t) \approx 0, \qquad (50)$$

where $\widetilde{F}_2$ is the operational matrix of product, corresponding to the vector $F_2 = \widetilde{F}_1^T\ F_1$ where $F_1 = \mathbf{P}^{(\nu+1)\mathbf{T}}\ C$. Moreover, the matrix $\mathbf{M}$ is introduced in Remark 2. By choosing $N = 14$, (50) is collocated at the roots of $P_{15}^{(\alpha,\beta)}(t)$. By solving the resulting algebraic systems, the unknown vector C can be determined for $\alpha = \beta = 1$. In Table 6, the values of the exact and approximate solutions are compared at the points $t_i = 0.2i, i = 0, 1, \ldots, 5$, for various values of $\nu = 0.7, 0.8, 0.9, 0.99, 1$. It can be observed from Table 6 that the numerical values approach the exact values as ν approaches 1. Moreover, (49) is solved in [28] using the Legendre Wavelet method (LWM), and the numerical results are shown in part (a) of Figure 6. Additionally, a graphical comparison between the exact solution and the Jacobi approximate solution is provided in part (b) of Figure 6 for $\alpha = \beta = 1$, $N = 14$, and $\nu = 0.7, 0.8, 0.9, 0.99, 1$. Based on Figure 6, no significant difference is observed between the approximate solutions obtained using the Jacobi operational method and the LWM.

Example 5. Consider the following fractional-order integro-differential equation:

Table 6: Values of approximate and exact solutions at selected points for different values of ν, $\alpha = \beta = 1$, and $N = 14$ for Example 4

t_i	$\nu = 0.7$	$\nu = 0.8$	$\nu = 0.9$	$\nu = 0.99$	$\nu = 1$	Exact values
0.0	-6.8186×10^{-5}	-2.3299×10^{-5}	-7.5562×10^{-6}	-4.9978×10^{-7}	7.1005×10^{-13}	0.000000
0.2	0.045519	0.035517	0.027619	0.021959	0.021403	0.021402
0.4	0.160712	0.133726	0.110976	0.093602	0.091834	0.091824
0.6	0.349039	0.300858	0.258855	0.225663	0.222222	0.222118
0.8	0.624058	0.550097	0.484580	0.431823	0.426292	0.425540
1.0	1.010148	0.902478	0.807252	0.730208	0.722096	0.718281

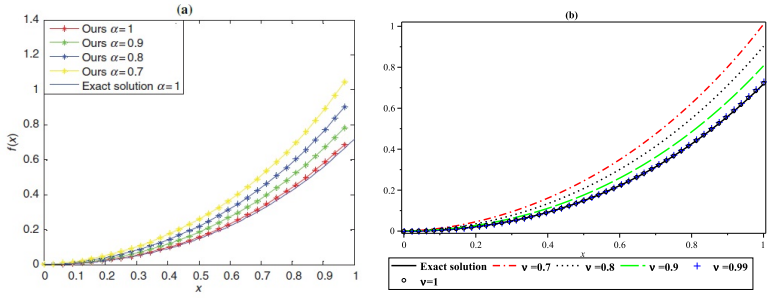


Figure 6: (a) Exact and approximate solutions obtained by LWM in Ref. [28], (b) Exact and approximate solutions obtained by Jacobi operational method for different values of ν, $\alpha = \beta = 1$, and $N = 14$ for Example 4

$$D^{\frac{1}{2}}y(t)+t^{\frac{3}{2}}\int_0^t (t+s)\,y(s)\,ds+\int_0^t ts\,D^{\frac{3}{8}}y(s)\,ds-g(t)=0, \qquad 0\leqslant t\leqslant 1, \quad (51)$$

where

$$g(t)=-\frac{2871928019}{3181474923}t^{\frac{5}{2}}+\frac{2850446155}{1841978638}t^{\frac{7}{2}}+\frac{3}{4}t^{\frac{11}{2}}-\frac{1}{2}t^{\frac{9}{2}}+\frac{11}{40}t^6-\frac{9}{40}x^5-\frac{1500440031}{8846634281}t^{\frac{45}{8}}+\frac{759388855}{3289964417}t^{\frac{53}{8}},$$

with initial condition $y(0) = 0$ and exact solution $y(t) = 3/4t^4 - 1/2t^3$.

By applying the fundamental matrices, presented in section 4 for (50), the following algebraic equation is resulted:

$$\Phi^T(t)\;C+t^{\frac{3}{2}}\Phi(t)\;K_1\;\widetilde{F}_1\;\mathbf{P}\;\Phi(t)+\Phi(t)\;K_2\;\widetilde{F}_2\;\mathbf{P}\;\Phi(t)-g(t)\approx 0, \qquad (52)$$

where the matrices $\widetilde{F}_1$ and $\widetilde{F}_2$ are the operational matrices of the product corresponding to the vectors $F_1 = \mathbf{P}^{(\frac{1}{2})\mathbf{T}}C$ and $F_2 = \mathbf{P}^{(\frac{1}{8})\mathbf{T}}C$, respectively. Table

7 presents the maximum absolute errors for various values of the parameters α and β with $N = 10$. The data in this table shows good agreement between the numerical results and the exact solution. In Table 8, the approximate solutions are calculated at equally spaced points $t_i = 0.1i, i = 0, 1, \ldots, 10$, for $N = 10$, $\alpha = -1/5$, $\beta = -1/4$, as well as for $\alpha = \beta = 2$. As observed, the error values change slowly over the interval $[0, 1]$, confirming the stability of the proposed method. A graphical comparison between the exact and approximate solutions, along with the plot of the absolute error function, is shown in parts (a) and (b) of Figure 7 for $\alpha = 1$, $\beta = 3$, and $N = 10$. Figure 7 demonstrates that the exact and approximate solutions are in good agreement. Additionally, the maximum absolute errors are plotted in Figure 8 for different numbers of collocation points, with $\alpha = 1/2$, $\beta = -1/2$, and various values of N (i.e., $N = 8 : 2 : 18$).

Table 7: Maximum absolute errors for various values of α and β and $N = 10$ for Example 5

(α, β)	$Error_{Abs}$	(α, β)	$Error_{Abs}$
$(0, 0)$	7.8229×10^{-7}	$(\frac{1}{2}, \frac{1}{2})$	1.5760×10^{-6}
$(1, 1)$	2.5249×10^{-6}	$(2, 2)$	4.6806×10^{-6}
$(-\frac{1}{2}, \frac{1}{2})$	1.9848×10^{-6}	$(\frac{1}{2}, -\frac{1}{2})$	1.9419×10^{-6}
$(-\frac{1}{2}, \frac{3}{4})$	1.8039×10^{-6}	$(1, \frac{1}{2})$	1.5760×10^{-6}
$(3, 1)$	7.0270×10^{-6}	$(1, 3)$	6.4222×10^{-6}
$(-\frac{1}{5}, -\frac{1}{4})$	3.9286×10^{-7}	$(\frac{1}{4}, \frac{1}{5})$	9.7123×10^{-4}

Example 6. Consider the following integro-differential equation with fractional derivative:

$$y''(t) + \frac{1}{t^2}\, y'(t) + \frac{1}{t}\, D^{0.7} y(t) = \int_0^t (6t^2 + 1)\, \cos(s)\, y(s)\, ds + g(t), \qquad 0 \leqslant t \leqslant 1, \tag{53}$$

where

$$\begin{aligned} g(t) = {} & (6t^6 - 71t^4 - 6t^7 + 132t^2 - 120t - 700t^3 + 24 + 119t^5) \sin(t) \\ & + (24t^5 - 140t^3 - 660t^2 + 355t^4 - 24t - 30t^6 - 120) \cos(t) \end{aligned}$$

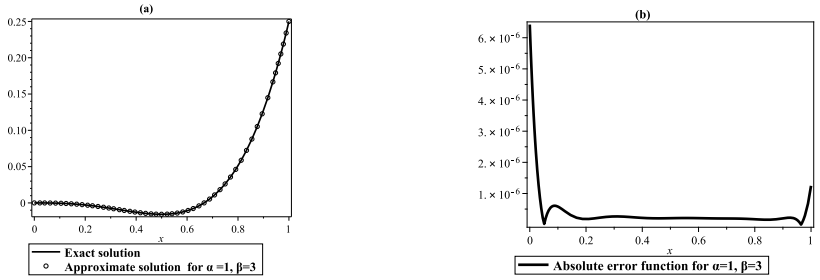


Figure 7: (a) Exact and approximate solutions, (b) Absolute error function for $\alpha = 1$, $\beta = 3$, and $N = 10$ for Example 5

Table 8: Values of absolute errors at some selected points for $N = 10$ of Example 5

t_i	Exact solution	$(\alpha, \beta) = (-\frac{1}{5}, -\frac{1}{4})$	$Error_{Abs}$	$(\alpha, \beta) = (2, 2)$	$Error_{Abs}$
0.0	0.000000000	-3.9286×10^{-7}	3.9286×10^{-7}	-4.6806×10^{-6}	4.6806×10^{-6}
0.1	-0.000425000	-0.000425109	1.0886×10^{-7}	-0.000425838	8.3841×10^{-7}
0.2	-0.002800000	-0.002799878	1.2197×10^{-7}	-0.002800614	6.1398×10^{-7}
0.3	-0.007425000	-0.007424974	2.6238×10^{-8}	-0.007425426	4.2625×10^{-7}
0.4	-0.012800000	-0.012800069	6.8770×10^{-8}	-0.012800338	3.3816×10^{-7}
0.5	-0.015625000	-0.015624900	1.0027×10^{-7}	-0.015625221	2.2117×10^{-7}
0.6	-0.010800000	-0.010800008	7.7315×10^{-9}	-0.010800155	1.5486×10^{-7}
0.7	0.008575000	0.008574946	5.4109×10^{-8}	0.008574919	8.0539×10^{-8}
0.8	0.051200000	0.051200089	8.8618×10^{-8}	0.051199977	2.2952×10^{-8}
0.9	0.127575000	0.127574902	9.7517×10^{-8}	0.127575027	2.6598×10^{-8}
1.0	0.250000000	0.249999746	2.5450×10^{-7}	0.249999155	8.4477×10^{-7}

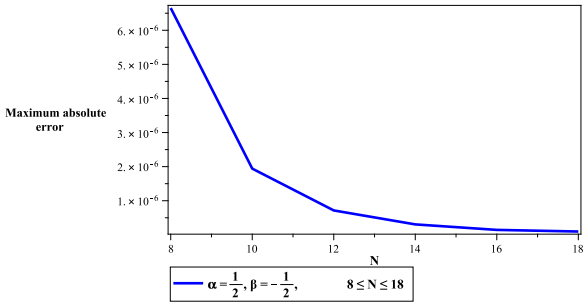


Figure 8: Maximum absolute error for different numbers of collocation points, $\alpha = \frac{1}{2}$, and $\beta = -\frac{1}{2}$ for Example 5

$$+20t^3+713t^2-4t+\frac{5995979123}{1902622932}t^{\frac{33}{10}}-\frac{6662541851}{2458295656}t^{\frac{23}{10}}+120.$$

The initial conditions are $y(0)=y'(0)=0$ and the exact solution does not exist. Following the process presented in case II in section 4 for (53) leads to the following algebraic equation:

$$\Phi^T(t)\,C+\frac{1}{t^2}\,\Phi(t)\,\mathbf{P}\,C+\frac{1}{t}\,\Phi(t)\,\mathbf{P^{(1.3)T}}\,C-\Phi(t)\,K\,\widetilde{F}\,\mathbf{P}\,\Phi(t)-g(t)\approx 0, \quad (54)$$

where the matrix $\widetilde{F}$ is the operational matrix of the product corresponding to the vector $F=(\mathbf{P^T})^{\mathbf{2}}C$. By choosing $N=10$, (54) is collocated at the roots of $P_{11}^{(\alpha,\beta)}(t)$. By determining the values of the parameters α and β and solving the resulting algebraic system, the unknown vector C is determined. Since the exact solution of (53) is not available, the corresponding error equation is solved to estimate the absolute errors. Table 9 presents the maximum estimated errors for various values of the parameters α and β with $N=10$. The maximum estimated errors are plotted in Figure 9 for different numbers of collocation points, where $\alpha=\beta=0$ and various values of N (i.e., $N=6:2:16$). In Table 10, the values of the approximate solutions are computed at the selected points $t_i=0.2i, i=1,2,\ldots,5$, for $N=6,10,14$ and $\alpha=\beta=0$. The maximum estimated errors decrease as N increases, and the values of the approximate solutions at the points t_i approach certain values with increasing N.

Table 9: Maximum estimated errors for various values of α and β and $N=10$ for Example 6

(α,β)	$Error_{Est}$	(α,β)	$Error_{Est}$
$(0,0)$	2.0157×10^{-9}	$(\frac{1}{2},\frac{1}{2})$	2.3112×10^{-9}
$(1,1)$	2.8243×10^{-9}	$(1,2)$	1.9696×10^{-9}
$(-\frac{1}{2},\frac{1}{2})$	2.0820×10^{-9}	$(\frac{1}{2},-\frac{1}{2})$	1.1915×10^{-8}
$(-\frac{1}{3},-\frac{1}{4})$	3.4068×10^{-9}	$(\frac{1}{4},\frac{1}{5})$	1.5846×10^{-3}
$(\frac{1}{5},-\frac{1}{3})$	1.7922×10^{-9}	$(2,1)$	3.0742×10^{-8}

Example 7. Consider the following nonlinear integro-differential equation with fractional derivative of order ν [16]:

Table 10: Values of approximate solutions for $N = 6, 10, 14$, $\alpha = \beta = 0$ for Example 6

t_i	$N = 6$	$N = 10$	$N = 14$
0.2	-0.0012800925	-0.0012799986	-0.0012800001
0.4	-0.0153599943	-0.0153600011	-0.0153600000
0.6	-0.0518398699	-0.0518399999	-0.5184000000
0.8	-0.0819199727	-0.0819200001	-0.0819200000
1.0	6.3832×10^{-8}	1.4580×10^{-10}	3.4336×10^{-11}
$Error_{Est}$	1.5511×10^{-7}	2.0157×10^{-9}	1.8897×10^{-10}

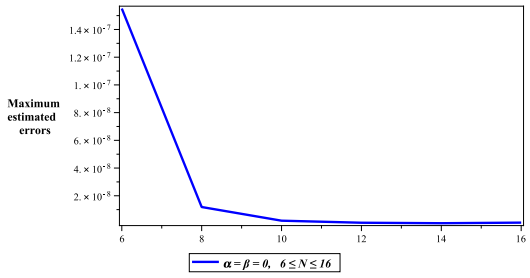


Figure 9: Maximum estimated error for various values of N and $\alpha = \beta = 0$ for Example 6

$$D^{\nu}y(t) = 1 + \int_0^t y(s)\ D^{\nu}y(s)\ ds, \qquad 0 \leqslant t \leqslant 1, \quad 0 < \nu \leqslant 1. \tag{55}$$

The initial condition for this example is $y(0) = 0$ and the exact solution is $y(t) = \sqrt{2}\tan(\sqrt{2}/2t)$ if $\nu = 1$. Following the process presented for nonlinear problems in section 4 for (55) leads to the following nonlinear algebraic equation:

$$\Phi^T(t)\ C - \Phi(t)\ K\ \widetilde{F}_2\ \mathbf{P}\ \Phi(t) - 1 \approx 0, \tag{56}$$

where the matrix $\widetilde{F}_2$ is the operational matrix of the product corresponding to the vector $F_2 = \widetilde{C}^T F_1$, where $F_1 = \mathbf{P}^{(\nu)\mathbf{T}} C$. Table 11 displays the maximum absolute errors for different values of α, β, and $N = 14$. The table demonstrates good consistency between the numerical results and the analytical solution. A graphical comparison between the exact and approximate solutions is shown in Figure 10 for values of $\nu = 0.5, 0.6, 0.7, 0.8, 0.9, 1$. (55)

has been solved using the Homotopy perturbation method (HPM), with the results reported in Ref. [16]. The absolute error functions of the approximate solutions can be seen in parts (a) (obtained by HPM) and (b) (obtained by the Jacobi operational method for $\alpha = 1/3, \beta = -1/4$) of Figure 11 for $N = 6$. It is evident that the error obtained by HPM is smaller than that obtained using the Jacobi operational method.

Table 11: Maximum absolute errors for various values of α and β, and $N = 14$ for Example 7

(α, β)	$Error_{Abs}$	(α, β)	$Error_{Abs}$
$(0, 0)$	3.2695×10^{-12}	$(\frac{1}{2}, \frac{1}{2})$	7.4506×10^{-12}
$(1, 1)$	1.3958×10^{-11}	$(2, 2)$	7.0196×10^{-11}
$(-\frac{1}{2}, \frac{1}{2})$	1.2529×10^{-12}	$(\frac{1}{2}, -\frac{1}{2})$	1.2842×10^{-11}
$(\frac{1}{3}, -\frac{1}{4})$	8.1919×10^{-12}	$(2, 1)$	5.5847×10^{-11}
$(-\frac{1}{3}, \frac{1}{4})$	1.0624×10^{-11}	$(1, 2)$	7.3824×10^{-11}

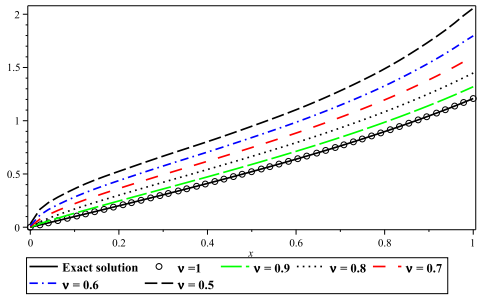


Figure 10: Approximate solutions for various values of ν, $N = 14$, $\alpha = \frac{1}{3}$, and $\beta = -\frac{1}{4}$ for Example 7

Example 8. Consider the following nonlinear system of integro-differential equations with fractional derivatives of order ν and γ [50]:

$$\begin{cases} D^{\nu}u(t) = u^2(t) + v^2(t) - \int_0^t u(s)\ ds, \\ D^{\gamma}v(t) = -\frac{1}{2}v^2(t) - u(t) - \int_0^t u(s)\ v(s)\ ds + \frac{1}{2}, \end{cases} \tag{57}$$

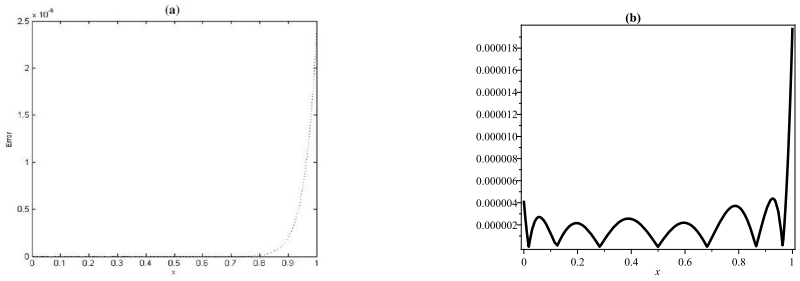


Figure 11: Absolute error function of approximate solution by :(a) HPM, (b) Jacobi operational method for $\alpha = \frac{1}{3}$, $\beta = -\frac{1}{4}$, and $N = 6$ in Example 7

where $0 < \nu, \gamma \leqslant 1$. The initial conditions for this example are $u(0) = 0$, $v(0) = 1$, and the exact solutions are $u(t) = \sin(t)$ and $v(t) = \cos(t)$ if $\nu = \gamma = 1$. Following the process presented for nonlinear problems in section 4 for system (57) leads to the following nonlinear system of algebraic equations:

$$\begin{cases} \Phi^T(t)\ C_1 - (\Phi^T(t)\ F_1)^2 - (\Phi^T(t)\ F_2)^2 + \Phi(t)\ K\ \widetilde{F}_1\ \mathbf{P}\ \Phi(t) \approx 0, \\ \Phi^T(t)\ C_2 + \frac{1}{2}(\Phi^T(t)\ F_2)^2 + \Phi^T(t)\ F_1 + \Phi(t)\ K\ \widetilde{U}\ \mathbf{P}\ \Phi(t) - \frac{1}{2} \approx 0, \end{cases} \tag{58}$$

where

$$\begin{aligned} &u(t) \approx \Phi^T(t)\ \mathbf{P}^{(\nu)T}\ C_1 = \Phi^T(t)\ F_1, \\ &v(t) \approx \Phi^T(t)\ \mathbf{P}^{(\gamma)T}\ C_2 + 1 \approx \Phi^T(t)\ \mathbf{P}^{(\gamma)T}\ C_2 + \Phi^T(t)\ V = \Phi^T(t)\ F_2, \\ &U = \widetilde{F}_2\ F_1. \end{aligned}$$

Table 12 displays the maximum absolute errors for different values of α, β, $\nu = \gamma = 1$, and $N = 10$. The table demonstrates good agreement between the numerical results and the analytical solutions. System (57) is solved using the Bernoulli wavelet method in [50]. The absolute errors of the approximate solutions, obtained by both the Jacobi collocation and Bernoulli wavelet methods, are calculated at the points $t_i = (2i-1)/16$, for $i = 1, 2, \ldots, 16$, and are shown in Tables 13 and 14. As seen, the results from the proposed method are more accurate (for $\alpha = -1/5$, $\beta = 1$, $\nu = \gamma = 1$, and $N = 10$). The approximate solutions obtained for $\nu = \gamma = 0.65, 0.75, 0.85, 0.95, 1$, $N = 10$, $\alpha = -1/3$, and $\beta = -1/4$ are depicted in Figure 12. It is evident that the approximate solutions approach the exact solutions as ν and γ approach 1.

Table 12: Maximum absolute errors of $u_{10}(t)$ and $v_{10}(t)$ for various values of α and β, and $\nu = \gamma = 1$ for Example 8

(α, β)	$Error_u$	$Error_v$	(α, β)	$Error_u$	$Error_v$
$(0, 0)$	3.0826×10^{-14}	3.5991×10^{-14}	$(1, 1)$	4.7157×10^{-13}	1.8738×10^{-13}
$(-\frac{1}{5}, 1)$	1.6600×10^{-12}	1.7294×10^{-12}	$(1, -\frac{1}{5})$	2.4327×10^{-13}	5.9886×10^{-13}
$(-\frac{1}{3}, -\frac{1}{4})$	5.5345×10^{-14}	3.8509×10^{-14}	$(\frac{1}{2}, -\frac{1}{2})$	1.0828×10^{-13}	1.3202×10^{-13}
$(-\frac{1}{2}, \frac{1}{2})$	2.3938×10^{-13}	1.3725×10^{-13}	$(\frac{1}{2}, \frac{1}{2})$	1.8085×10^{-13}	7.0934×10^{-14}

Table 13: Values of absolute errors of $u_{10}(t)$ at some points for $\alpha = -\frac{1}{5}, \beta = 1$ in Example 8

t_i	$Error_u$	$Error_u$ in [50]	t_i	$Error_u$	$Error_u$ in [50]
0.03125	8.7968×10^{-14}	4.51×10^{-5}	0.53125	8.1234×10^{-13}	6.87×10^{-4}
0.15625	2.3332×10^{-13}	2.20×10^{-4}	0.65625	1.0164×10^{-13}	8.24×10^{-4}
0.28125	4.2781×10^{-13}	3.68×10^{-4}	0.78125	1.2319×10^{-13}	9.55×10^{-4}
0.40625	6.1337×10^{-13}	5.41×10^{-4}	0.90625	1.4680×10^{-13}	1.08×10^{-3}

Table 14: Absolute errors of the approximate solution $v_{10}(t)$ for $\alpha = -\frac{1}{5}$ and $\beta = 1$ at selected points in Example 8

t_i	$Error_v$	$Error_v$ in [50]	t_i	$Error_v$	$Error_v$ in [50]
0.03125	7.8345×10^{-13}	4.77×10^{-4}	0.53125	5.8795×10^{-13}	1.05×10^{-4}
0.15625	7.3162×10^{-13}	4.19×10^{-4}	0.65625	4.7715×10^{-13}	4.59×10^{-5}
0.28125	6.7339×10^{-13}	3.38×10^{-4}	0.78125	3.5498×10^{-13}	2.18×10^{-4}
0.40625	6.3567×10^{-13}	2.33×10^{-4}	0.90625	2.1013×10^{-13}	4.11×10^{-4}

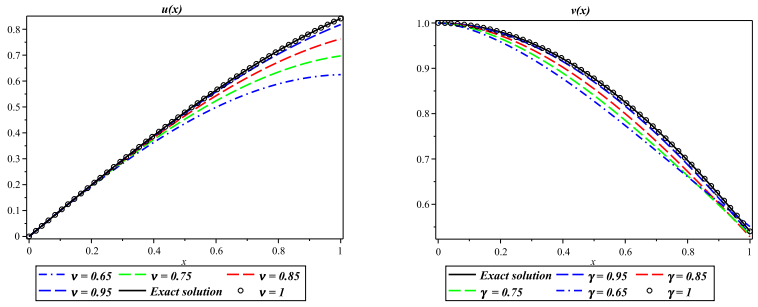


Figure 12: Appeoximate solutions for various values of ν and γ, $\alpha = -\frac{1}{3}, \beta = -\frac{1}{4}$, and $N = 10$ for Example 8

7 Conclusion

In this study, the Jacobi operational method has been successfully applied to solve both linear and nonlinear integro-differential equations involving fractional derivatives of arbitrary order ν. This method utilized shifted Jacobi polynomials defined on the interval $[0, 1]$, transforming the given problem, whether linear or nonlinear, into a set of more manageable algebraic equations. The resulting algebraic system is easier to solve compared to the original problem, highlighting a significant advantage of this approach. An additional strength of the proposed method is the straightforward determination of coefficients for the Jacobi series solution through the algorithms provided. Operational matrices for fractional integration and product are efficiently constructed using these algorithms. The study also investigated the existence and uniqueness of solutions for the equations and analyzed the convergence of the numerical approach. Illustrative examples provided in Examples 1–4, 7, and 8 showed that exact solutions exist only for integer values of ν. However, the numerical solutions obtained demonstrated strong agreement with the analytic solutions in these cases. For fractional values of ν within the range $m-1 < \nu < m$, where $m \in \mathbb{N}$, the numerical solutions gradually converge to the exact solutions as ν approaches m. In Example 5, where an exact solution is provided, the resulting absolute errors demonstrate the effectiveness of the proposed algorithm in solving such equations. Remarkably, Remark 5 offered a reliable method to estimate the absolute errors for the presented examples, which is particularly useful in scenarios like Example 6, where an exact solution is unavailable. The precision of the Jacobi operational method is further demonstrated by the consistently small and nearly uniform errors across the analyzed interval, affirming the validity of the Jacobi collocation method. A comparative analysis with other established methods, such as HPM, VIM, Sinc-collocation, Legendre wavelets, and Bernoulli wavelet methods, showed that the proposed technique produces more accurate results. Additionally, this method avoided the computational complexities of existing methods, such as Homotopy perturbation and Adomian decomposition, especially in determining Adomian polynomials. Based on the promising results obtained, it is expected that the Jacobi colloca-

tion method will become a robust tool for solving both linear and nonlinear functional equations. Future work aims to extend this method to address singular integro-differential equations with variable-order derivatives, although some modifications will be required. However, the method can be extended to higher-dimensional fractional equations, such as fractional integro-partial differential equations and partial differential equations with fractional orders. While the operational collocation method with Jacobi polynomials is effective for solving FIDEs in one dimension, applying it to higher dimensions presents challenges such as handling fractional derivatives in multiple dimensions, managing complex boundary conditions, ensuring numerical stability, and controlling computational costs. These challenges can be addressed by adopting a pseudo-operational approach.

CRediT authorship contribution statement

Khadijeh Sadri: Writing – review & editing, Writing-original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization, Supervision. David Amilo: Writing-original draft, Validation, Formal analysis, Data curation. Evren Hincal: Visualization, Resources, Project administration.

Declaration of competing interest

The authors declare no conflicts of interest regarding the publication of this manuscript.

Data availability

Due to the nature of the research, supporting data is not available

Funding

This study receives no funding.

Acknowledgements

Authors are very grateful to anonymous referees for their careful reading and valuable comments which led to the improvement of this paper.

References

[1] Abbasbandy, S. *On convergence of homotopy analysis method and its application to fractional integro-differential equations*, Quaes. Math. 36 (2013), 93–105.

[2] Abdelkawy, M.A., Lopes, A.M. and Babatin, M.M. *Shifted fractional Jacobi collocation method for solving fractional functional differential equations of variable order*, Chaos. Soliton. Frac. 134 (2020), 109721.

[3] Abo-Gabal, H., Zaky, M.A. and Doha, E.H. *Fractional Romanovski–Jacobi tau method for time-fractional partial differential equations with nonsmooth solutions*, Appl. Num. Math. 182 (2022), 214–234.

[4] Ahmad, W.M. and El–Khazali, R. *Fractional-order dynamical models of love*, Chaos. Solit. Frac. 33 (2007), 1367–1375.

[5] Babaei, A., Jafari, H. and Banihashemi, S. *Numerical solution of variable order fractional nonlinear quadratic integro-differential equations based on the sixth-kind Chebyshev collocation method*, J. Comput. Appl. Math. 377 (2020), 112908.

[6] Bhrawy, A.H., Tharwat, M.M. and Yildirim, A. *A new formula for fractional integrals of Chebyshev polynomials: Application for solving multi-term fractional differential equations*, App. Math. Modelling. 37 (2013), 4245–4252.

[7] Biazar, J. and Sadri, K. *Solution of weakly singular fractional integro-differential equations by using a new operational approach*, J. Comput. Appl. Math. 352 (2019), 453–477.

[8] Chen, Y. and Tang, T. *Convergence analysis of the Jacobi spectral collocation methods for Volterra integral equations with a weakly singular kernel*, Math. Comput. 79 (2010), 147–167.

[9] Darweesh, A., Al-Khaled, K. and Al-Yaqeen, O.A. *Haar wavelets method for solving class of coupled systems of linear fractional Fredholm integro-differential equations*, Heliyon. 9(9) (2023), e19717.

[10] Dehestani, H., Ordokhani, Y. and Razzaghi, M. *Combination of Lucas wavelets with Legendre–Gauss quadrature for fractional Fredholm–Volterra integro-differential equations*, J. Comput. Appl. Math. 382 (2021), 113070.

[11] Dehghan, M. and Fakhar-Izadi, F. *The spectral collocation method with three different bases for solving a nonlinear partial differential equation arising in modeling of nonlinear waves*, Math. Comput. Model. 53 (2011), 1865–1877.

[12] Diethelm, A. and Free, D. *On the solution of nonlinear fractional-order differential equations used in the modeling of viscoplasticity, in: Scientific Computing in Chemical Engineering II: Computational Fluid Dynamics, Reaction Engineering, and Molecular Properties* (Eds. F. Keil, W. Mackens, H. Vob and J. Werther), Springer Verlag, Heidelberg, Pp. 217–224, 1999.

[13] Doha, E.H., Bhrawy, A.H. and Ezz-Eldien, S.S. *A new Jacobi operational matrix: An application for solving fractional differential equations*, App. Math. Modelling. 36 (2012), 4931–4943.

[14] Ebrahimi, H. and Sadri, K. *An operational approach for solving fractional pantograph differential equation*, Iranian J. Numer. Anal. Opt. 9(1) (2019), 37–68.

[15] Elbeleze, A.A., Kilicman, A. and Taib, B.M. *Approximate solution of integro-differential equation of fractional (arbitrary) order*, J. King Saud. Univ. Sci. 28(1) (2016), 61–68.

[16] Ghazanfari, B., Ghazanfari, A.G. and Veisi, F. *Homotopy Perturbation method for the nonlinear fractional integro-differential equations*, Aust. J. Basic. App. Sci. 4(12) (2010), 5823–5829.

[17] Guo, B.Y. and Wang, L.L. *Jacobi approximations in non-uniformly Jacobi-weighted Sobolev spaces*, J. Approx. Theory. 128 (2004), 1–41.

[18] He, J.H. *Nonlinear oscillation with fractional derivative and its applications*, In: International Conference on Vibrating Engineering, Dalian, China, pp. 288–291, 1998.

[19] He, J.H. *Some applications of nonlinear fractional differential equations and their approximations*. Bull. Sci. Technol. 15 (1999), 86–90.

[20] Heydari, M.H., Zhagharian, Sh. and Razzaghi, M. *Jacobi polynomials for the numerical solution of multi-dimensional stochastic multi-order time fractional diffusion-wave equations*, Comput. Math. Appl. 152 (2023), 91–101.

[21] Hosseininia, M., Heydari, M.H. and Avazzadeh, Z. *Orthonormal shifted discrete Legendre polynomials for the variable-order fractional extended Fisher–Kolmogorov equation*, Chaos. Soliton. Frac. 155 (2022), 111729.

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

[23] Keskin, Y., Karaoglu, O., Servi, S. and Oturan, G. *The approximate solution of high order linear fractional differential equations with variable coefficients in term of generalized Taylor series*, Math. Comput. App. 16 (2011), 617–629.

[24] Khader, M.M. and Sweilam, N.H. *On the approximate solutions for system of fractional integro-differential equations using Chebyshev pseudo-spectral method*, Appl. Math. Modelling. 37(24) (2013), 9819–9828.

[25] Kumar, K., Pandey, R.K. and Sharma, S. *Comparative study of three numerical schemes for fractional integro-Differential equations*, J. Comput. App. Math. 315 (2017), 287–302.

[26] Li, X.J. and Xu, C.J. *A space-time spectral method for the time fractional diffusion equation*, SIAM. J. Numer. Anal. 47(3) (2009), 2108–2131.

[27] Mainardi, F. *Fractals and Fractional Calculus Continuum Mechanics*, Springer Verlag, Pp, 291–348, 1997.

[28] Meng, Z., Wang, L., Li, H. and Zhang, W. *Legendre wavelets method for solving fractional integro-differential equations*, Int. J. Comput. Math. 92(6) (2015), 1275–1291.

[29] Mokhtary, P. *Discrete Galerkin method for fractional integro-differential equations*, Acta. Math. Sci. 36B(2) (2016), 560–578.

[30] Momani, S. and Noor, M.A. *Numerical methods for fourth-order fractional integro-differential equations*, App. Math. Comput. 182 (2006), 754–760.

[31] Nazari Susahab, D., Shahmorad, S. and Jahanshahi, M. *Efficient quadrature rules for solving nonlinear fractional integro-differential equations of the Hammerstein type*, App. Math. Modelling. 39(18) (2015), 5452–5458.

[32] Rahimkhani, P. and Ordokhani, Y. *A numerical scheme based on Bernoulli wavelets and collocation method for solving fractional partial differential equations with Dirichlet boundary conditions.* Numer. Methods Partial Differ. Equ. 35(1) (2019), 34–59.

[33] Rahimkhani, P., Ordokhani, Y. and Babolian, E. *An efficient approximate method for solving delay fractional optimal control problems*, Nonlinear. Dyn. 86 (2016), 1649—1661.

[34] Rahimkhani, P., Ordokhani, Y. and Babolian, E. *A new operational matrix based on Bernoulli wavelets for solving fractional delay differential equations*, Numer. Algor. 74 (2017), 223—245.

[35] Rahimkhani, P., Ordokhani, Y. and Babolian, E. *Numerical solution of fractional pantograph differential equations by using generalized fractional-order Bernoulli wavelet*, J. Comput. Appl. Math. 309 (2017), 493–510.

[36] Rahimkhani, P., Ordokhani, Y. and Babolian, E. *Fractional-order Bernoulli functions and their applications in solving fractional Fredholem–Volterra integro-differential equations*. Appl. Numer. Math. 122 (2017), 66–81.

[37] Rahimkhani, P., Ordokhani, Y. and Babolian, E. *Muntz–Legendre wavelet operational matrix of fractional-order integration and its applications for solving the fractional pantograph differential equations*, Numer. Algor. 77 (2018), 1283—1305.

[38] Sadek, L. *A cotangent fractional derivative with the application*, Fractal Fract. 7(6) (2023), 444.

[39] Sadek, L., Baleanu, D., Abdo, M.S. and Shatanawi, W. *Introducing novel Θ-fractional operators: Advances in fractional calculus*, J. King Saud University - Sci. 36(9) (2024), 103352.

[40] Sadek, L. and Bataineh, A.S. *The general Bernstein function: Application to χ-fractional differential equations*, Math. Method. Appl. Sci. 47(7), (2024) 6117–6142.

[41] Sadek, L., Bataineh, A.S., Isik, O.R., Alaoui, H.T. and Hashim, I. *A numerical approach based on Bernstein collocation method: Application to differential Lyapunov and Sylvester matrix equations*, Math. Comput. Simul. 212 (2023), 475–488.

[42] Sadek, L. and Lazar, T.A. *On Hilfer cotangent fractional derivative and a particular class of fractional problems*, AIMS Mathematics. 8(12) (2023), 28334–28352.

[43] Sadek, L., Ounamane, S., Abouzaid, B. and Sadek, E.M. *The Galerkin Bell method to solve the fractional optimal control problems with inequality constraints*, J. Comput. Sci. 77 (2024), 102244.

[44] Sadek, L. and Algefary, A. *On quantum trigonometric fractional calculus*, Alex. Eng. J. 120 (2025), 371–377.

[45] Sadri, K., Amini, A. and Cheng, C. *Low cost numerical solution for three-dimensional linear and nonlinear integral equations via three-dimensional Jacobi polynomials*, J. Comput. App. Math. 319 (2017), 493–513.

[46] Sadri, K., Amini, A. and Cheng, C. *A new numerical method for delay and advanced integro-differential equations*, Numer. Algor. 77 (2018), 381–412.

[47] Szego, G. *Orthogonal polynomials*, American Mathematical Society, Providence, Rhode Island, 1939.

[48] Tavares, D., Almeida, R. and Torres, D.F.M. *Caputo derivatives of fractional variable order: Numerical approximations*, Commun. Nonlinear Sci. Numer. Simul. 35 (2016), 69–87.

[49] Varol Bayram, D. and Daşcıoğlu, A. *A method for fractional Volterra integro-differential equations by Laguerre polynomials*, Adv. Differ. Equ. 2018 (2018), 466.

[50] 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.

[51] Yaghoubi, S., Aminikhah, H. and Sadri, K. *A new efficient method for solving yystem of weakly singular fractional integro-differential equations by shifted sixth-kind Chebyshev polynomials*, J. Math. 2022 (2022), 9087359.

[52] Yaghoubi, S., Aminikhah, H. and Sadri, K. *A spectral shifted Gegenbauer collocation method for fractional pantograph partial differential equations and its error analysis*, Sadhana. 48 (2023), 213.

[53] Yang, Y. *Convergence Analysis of the Jacobi Spectral-Collocation Method for Fractional Integro-Differential Equations*, Acta Math. Sci. 34(3) (2014), 673–690.

Iranian Journal of Numerical Analysis and Optimization
Vol. 15, No. 2, 2025, pp 823–851
https://doi.org/10.22067/ijnao.2025.91108.1558
https://ijnao.um.ac.ir/

Research Article

Mathematical modeling, analysis, and optimal control of the cochineal insect impact on cacti plants

S. Khassal*, E.M. Moumine and O. Balatif

Abstract

We propose a mathematical model, SIM, that describes the dynamics of cochineal insect spread among cacti and examines the effects of various control strategies. The model is analyzed for the existence and uniqueness of solutions, and we investigate the equilibrium points and stability of the system using both local and global stability analyses. By performing numerical simulations in MATLAB, we validate our theoretical findings. Furthermore, we propose an optimal control strategy to minimize

*Corresponding author

Received 5 December 2024; revised 1 February 2025; accepted 14 February 2025

Sofiane Khassal
Department of Mathematics, Faculty of Sciences El Jadida, Chouaib Doukkali University, El Jadida, Morocco. e-mail: sofiane.k@ucd.ac.ma

El Mehdi Moumine
Department of Mathematics and Computer Science, Faculty of science Ben M'sik, University Hassan II, Casablanca, Morocco. e-mail: moumine.maths@gmail.com

Omar Balatif
Department of Mathematics, Faculty of Sciences El Jadida, Chouaib Doukkali University, El Jadida, Morocco. e-mail: balatif.maths@gmail.com

How to cite this article
Khassal, S., Moumine, E.M. and Balatif, O., Mathematical modeling, analysis, and optimal control of the cochineal insect impact on cacti plants. *Iran. J. Numer. Anal. Optim.*, 2025; 15(2): 823-851. https://doi.org/10.22067/ijnao.2025.91108.1558

the cochineal population in cacti fields. The optimal control problem is formulated using Pontryagin's maximum principle, and the corresponding optimality system is solved iteratively. Our study compares three control strategies: cutting and burning infected cacti, insecticide spraying, and a combined approach. The results demonstrate that the combined strategy is the most effective in reducing the cochineal population. This research provides valuable insights into managing cochineal infestations and offers practical recommendations for farmers to control the spread of these pests.

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

Keywords: Mathematical modeling; Stability analysis; Optimal control; Cochineal insect; Cacti plants.

1 Introduction

The cochineal insect, **Dactylopius coccus**, is a well-known pest that significantly affects cacti and other succulents. It is notorious for feeding on plant sap, weakening their health, and, in severe cases, leading to the death of the host plant. These infestations are of considerable economic concern, especially in regions where cacti are a major agricultural crop. The damage caused by cochineal infestations includes dehydration, discoloration, and, ultimately, the destruction of entire crops. Cochineal is particularly harmful to cacti, which are otherwise resilient [13, 1].

Historically, cochineal insects were highly valued for their use in producing carmine dye, a pigment that has been utilized for centuries in the textile industry and art. The ability of these insects to produce carmine made them a valuable commodity in pre-Columbian Mesoamerica, where they were a significant economic resource. During the colonial era, the Spanish established a monopoly on cochineal trade, making it one of the most important exports from the Americas to Europe. The cultivation of cochineal and its host plants later spread globally, particularly to Mediterranean regions and parts of Asia, following its introduction by European colonial powers [8].

Mathematical modeling has become an essential tool for understanding the dynamics of cochineal infestations and assessing the effectiveness of dif-

ferent control measures. Early mathematical models of pest populations date back to the works of Verhulst and Lotka, who developed foundational models describing population growth and interactions [17, 12]. More recently, mathematical models have been applied to agricultural pest spread, including cochineal. These models typically take the form of differential equations that describe the dynamics of both the pest and its host plants [14, 10]. For example, the SIR (Susceptible-Infected-Removed) model, widely used in epidemiology, has been adapted to model the spread of various agricultural pests, including cochineal [5, 4].

However, most previous models have focused on the population dynamics of the pest or the biological aspects of infestation, often neglecting the effects of control measures and external interventions. In this paper, we propose a new mathematical model, denoted as the SIM model, that describes the spread of cochineal insects on cacti while incorporating different types of control strategies. Our model builds upon the classical SIR framework by including terms that represent the impact of interventions such as chemical treatments, removal of infected plants, and natural predation. We also conduct a stability analysis to examine the system's behavior under various conditions and determine the effects of control measures on pest spread [16, 9].

The novelty of our work lies in integrating optimal control theory with pest management strategies. Using Pontryagin's maximum principle, we derive optimal control strategies for minimizing the spread of cochineal while considering resource constraints. This approach enables us to evaluate the most efficient allocation of resources for pest control and optimize management efforts [15, 6].

This study contributes to the scientific literature by offering a more comprehensive approach to understanding cochineal infestations. In addition to providing theoretical insights into the dynamics of cochineal spread, our model offers practical guidance for pest management in agricultural settings. The findings have significant implications for policymakers and farmers, suggesting effective strategies for reducing the impact of cochineal infestations and preventing their spread to new regions.

The remainder of the paper is structured as follows: Section 2 introduces the proposed mathematical model, Section 3 analyzes its fundamental properties and equilibrium points, and Section 4 presents a stability analysis. Section 5 provides numerical simulations to validate the theoretical results, while Section 6 discusses optimal control strategies for managing cochineal infestations. In Section 7, optimal control numerical results are considered. Finally, Section 8 concludes with a summary of our findings and potential applications.

2 Formulation of the mathematical model

Before introducing our model, we first examine the biology and reproductive behavior of the cochineal insect, as understanding these aspects is essential for modeling its spread among cacti.

Cochineal insects are soft-bodied, oval-shaped organisms. Males possess wings, while females remain wingless. After mating, females lay eggs that rapidly hatch into tiny larvae. These larvae secrete a white waxy substance on their bodies, which serves as protection against water loss and excessive sunlight. This waxy coating appears on cacti as white, cotton-like masses. The larvae migrate to the edges of cactus pads, where wind currents carry the wax threads, dispersing them onto adjacent plants.

Cochineal infestations can cause significant agricultural losses, as these insects feed on cacti by extracting their fluids. In severe cases, this leads to plant desiccation and eventual death. In most instances, if a cactus is infested, it either perishes or must be removed—farmers typically resort to cutting, burying, or burning infected plants to prevent further spread. The insect's rapid dispersion is facilitated by wind, as well as by transportation via farming equipment, trucks, livestock wool, and the movement of animals and people from infested areas.

Based on these observations, we propose a mathematical model, denoted as SIM, to describe the spread of cochineal insects among cacti. We categorize the total cacti population, N, into two compartments:

- **S**: Represents the healthy (susceptible) cacti plants in the region.

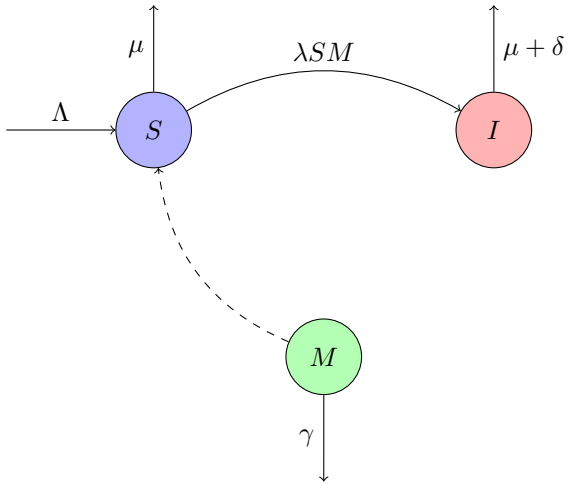


Figure 1: Compartments of cacti population

- **I**: Represents the infected cacti plants affected by the cochineal insect.

Additionally, we introduce a third compartment to represent the cochineal insect population, denoted as **M**.

In Figure 1, λ is the rate of cochineal insect encounters with cacti (measured in [cochineal/day]), γ is the cochineal insect mortality rate (measured in [cochineal/day]), δ is the death rate of cacti due to infection (measured in [cacti/day]), β is the growth rate of the cochineal insect population (measured in [cochineal/day] and $\beta > 0$), K is the carrying capacity of the cochineal insect population (measured in [cochineal/day] and $K > 0$), Λ is the recruitment rate of cacti (measured in [cacti/day]), and μ is the natural mortality rate of cacti (measured in [cacti/day]). Moreover,

$$\begin{cases} \frac{dS}{dt} = \Lambda - \lambda SM - \mu S, \\ \frac{dI}{dt} = \lambda SM - \mu I - \delta I, \\ \frac{dM}{dt} = \beta M \left(1 - \frac{M}{K}\right) - \gamma M, \end{cases} \tag{1}$$

with the following nonnegative initial conditions: $S(0) \geqslant 0, I(0) \geqslant 0, M(0) \geqslant 0$.

3 Model basic properties

The provided model describes the populations of both cacti and Cochineal insects. Therefore, it becomes necessary to investigate some fundamental properties of the system (1) such as the existence, positivity, and boundedness of the solutions.

3.1 Positivity of solutions

Theorem 1. If $S(0) \geq 0$, $I(0) \geq 0$ and $M(0) \geq 0$, then the solutions $S(t)$, $I(t)$, and $M(t)$ of system (1) are positive for all $t \geq 0$.

Proof. It follows from the first equation of system (1) that

$$\frac{\mathrm{d}S}{\mathrm{d}t} = \Lambda - \lambda SM - \mu S \geq -(\lambda M + \mu) S. \tag{2}$$

Then, we have

$$\frac{\mathrm{d}S}{S} \geq -(\lambda M + \mu)\,\mathrm{d}t. \tag{3}$$

Integrating this inequality from 0 to t gives

$$S(t) \geq S_b(0) e^{-\int_0^t (\lambda M + \mu) ds}.$$

That implies

$$S(t) \geq 0 \qquad \text{for all } t \geq 0.$$

Similarly, we prove that $I(t) \geq 0$ and $M(t) \geq 0$ for all $t \geq 0$.

□

3.2 Boundedness of the solutions

Theorem 2. Let $C = \max\{M(0), K\}$. Then the set

$$\Gamma = \left\{ (S, I) \in \mathbb{R}^2_+ : N(t) \leq \frac{\Lambda}{\mu} \right\} \times \{ M \in \mathbb{R}_+ : M(t) \leq C \}$$

is positively invariant under system (1) with nonnegative initial conditions $S(0)$, $I(0)$, and $M(0)$.

Proof. From the three equations of the system (1), we have

$$\frac{dN(t)}{dt} = \Lambda - \mu N(t) - \delta I \leq \Lambda - \mu N(t). \tag{4}$$

Then,

$$N(t) \leq N(0)e^{-\mu t} + \frac{\Lambda}{\mu}\left[1 - e^{-\mu t}\right]. \tag{5}$$

If we take limit $t \to \infty$, we have $0 \leq N(t) \leq \frac{\Lambda}{\mu}$.

From the last equation of (1), we have

$$\frac{\mathrm{d}M}{\mathrm{d}t} \leq \beta M\left(1 - \frac{M}{K}\right). \tag{6}$$

Hence, employing a typical comparison approach yields $\limsup_{t\to\infty} M(t) \leq K$.

Finally, the set Γ is positively invariant for the system (1). □

3.3 Existence of solutions

Theorem 3. The system (1) that satisfies a given initial condition $(S(0), I(0), M(0))$ has a unique solution.

Proof. Model (1) can be reformulated in matrix form as follows:
Let $X(t) = (S, I, M)^T$ and $F(X(t)) = \left(\frac{dS}{dt}, \frac{dI}{dt}, \frac{dM}{dt}\right)^T$.

The system (1) can be rewritten as follows:

$$F(X(t)) = AX + B(X(t)),$$

where

$$A = \begin{pmatrix} -\mu & 0 & 0 \\ 0 & -(\mu+\delta) & 0 \\ 0 & 0 & \beta - \gamma \end{pmatrix}$$

and

$$B(X(t)) = \begin{pmatrix} \Lambda - \lambda SM \\ \lambda SM \\ -\frac{\beta M^2}{K} \end{pmatrix}.$$

Let X_1 and X_2 be solutions of (1). Then

$$\begin{aligned}
|B(X_1) - B(X_2)| \leq & \mid \lambda(S_2M_2 - S_1M_1)| + \mid \lambda(S_1M_1 - S_2M_2)| + \mid \frac{\beta}{K}(M_2^2 - M_1^2)| \\
\leq & \mid \lambda(S_2M_2 - S_2M_1)| + \mid \lambda(S_2M_1 - S_1M_1)| \\
& + \mid \lambda(S_1M_1 - S_1M_2)| + \mid \lambda(S_1M_2 - S_2M_2)| + \mid \frac{\beta}{K}(M_2^2 - M_1^2)| \\
\leq & \lambda S_2 \mid M_2 - M_1| + \lambda M_1 \mid S_2 - S_1| + \lambda S_1 \mid M_1 - M_2| \\
& + \lambda M_2 \mid S_1 - S_2| + \frac{\beta}{K} \mid M_2 - M_1| \mid M_2 + M_1| \\
\leq & \frac{\lambda\Lambda}{\mu} \mid M_2 - M_1| + \lambda C \mid S_2 - S_1| \\
& + \frac{\lambda\Lambda}{\mu} \mid M_1 - M_2| + \lambda C \mid S_1 - S_2| + \frac{2\beta C}{K} \mid M_2 - M_1| \\
\leq & \left(\frac{2\lambda\Lambda}{\mu} + \frac{2\beta C}{K}\right) \mid M_1 - M_2| + 2\lambda C \mid S_1 - S_2| \\
\leq & N \|X_1 - X_2\|,
\end{aligned}$$

where

$$N = \max\left(\frac{2\lambda\Lambda}{\mu} + \frac{2\beta C}{K}, 2\lambda C, \|A\|\right).$$

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 $S(t) \geq 0$, $I(t) \geq 0$, and $M(t) \geq 0$ in $\mathbb{R}^3_+$. Therefore, a solution of the model (1) exists [3]. □

4 Stability analysis of the model

In this section, we explore the stability characteristics of model (1) for both the cochineal insect-free equilibrium point denoted as E_0 and the cochineal insect endemic equilibrium point denoted as E^*.

The equilibrium point of model (1) satisfies the following equations:

$$\Lambda - \lambda SM - \mu S = 0, \tag{7}$$

$$\lambda SM - \mu I - \delta I = 0, \tag{8}$$

$$\beta M\left(1 - \frac{M}{K}\right) - \gamma M = 0. \tag{9}$$

From (9), we have

$$M\left(\beta\left(1 - \frac{M}{K}\right) - \gamma\right) = 0 \Rightarrow M = 0 \quad \text{or} \quad M = K - \frac{\gamma K}{\beta}.$$

For $M = 0$: (7) $\Rightarrow \Lambda - \mu S = 0 \Rightarrow S = \frac{\Lambda}{\mu}$, (8) $\Rightarrow I = 0$, then $E_0 = \left(\frac{\Lambda}{\mu}, 0, 0\right)$.

For $M = M^* = K - \frac{\gamma K}{\beta} = K(1 - \frac{1}{\Re_0})$: From (7), we have

$$S^* = \frac{\Lambda\beta}{\lambda K(\beta - \gamma) + \mu\beta} = \frac{\Lambda}{\lambda K(1 - \frac{1}{\Re_0}) + \mu}.$$

From (8), we have

$$I^* = \frac{\lambda\Lambda K(\beta - \gamma)}{(\mu + \delta)\left(\lambda K(\beta - \gamma) + \mu\beta\right)} = \frac{\lambda\Lambda K(1 - \frac{1}{\Re_0})}{(\mu + \delta)\left(\lambda K(1 - \frac{1}{\Re_0}) + \mu\right)}.$$

Therefore,

$$E^* = \left(S^*, I^*, M^*\right),$$

where $\Re_0$ represents the basic reproduction number, defined as

$$\Re_0 = \frac{\beta}{\gamma}.$$

In epidemiology, the basic reproduction number $\Re_0$ signifies the average number of secondary infections caused by one infected individual in a population entirely susceptible. We calculated this value utilizing the next-generation matrix method as described in [2, 18]

Indeed, if we let $x = (S, I, M)$, then the model (1) can be expressed as

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \mathscr{F}(x) - \vartheta(x),$$

where

$$\mathscr{F}(x) = \begin{pmatrix} \Lambda \\ \lambda SM \\ \beta M\left(1 - \frac{M}{K}\right) \end{pmatrix},$$

$$\vartheta(x) = \begin{pmatrix} \lambda SM + \mu S \\ \mu I + \delta I \\ \gamma M \end{pmatrix}.$$

The Jacobian matrices of $\mathscr{F}(x)$ and $\vartheta(x)$ at the free equilibrium E_0 are, respectively,

$$D\mathscr{F}(E_0) = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & \frac{\lambda\Lambda}{\mu} \\ 0 & 0 & \beta \end{pmatrix},$$

$$D\vartheta(E_0) = \begin{pmatrix} \mu & 0 & \frac{\lambda\Lambda}{\mu} \\ 0 & \mu+\delta & 0 \\ 0 & 0 & \gamma \end{pmatrix}.$$

where

$$F = \begin{pmatrix} 0 & \frac{\lambda\Lambda}{\mu} \\ 0 & \beta \end{pmatrix},$$

$$V = \begin{pmatrix} \mu+\delta & 0 \\ 0 & \gamma \end{pmatrix}.$$

Finally, we have

$$\Re_0 = \rho\left(FV^{-1}\right) = \frac{\beta}{\gamma}.$$

4.1 Local stability

In this section, we analyze the local stability of both the free equilibrium point E_0 and the endemic equilibrium point E^*.

Theorem 4. If $\Re_0 < 1$, then E_0 is locally asymptotically stable.

Proof. The Jacobian matrix calculated for system (1) at E_0 is given by

$$J_{E_0} = \begin{pmatrix} -\mu & 0 & \frac{-\lambda\Lambda}{\mu} \\ 0 & -\mu-\delta & \frac{\lambda\Lambda}{\mu} \\ 0 & 0 & \beta-\gamma \end{pmatrix}.$$

We put $d_1 = -\mu < 0$, $d_2 = -\mu - \delta$ and $d_3 = \beta - \gamma = \gamma(\Re_0 - 1)$. Then, let us consider the matrix:

$$T_1 = \begin{pmatrix} d_1 - r & 0 & \frac{-\lambda\Lambda}{\mu} \\ 0 & d_2 - r & \frac{\lambda\Lambda}{\mu} \\ 0 & 0 & d_3 - r \end{pmatrix}.$$

Hence, the eigenvalues of J_{E_0} are $r_1 = d_1$, $r_2 = d_2$ and $r_3 = d_3$. Clearly, all the eigenvalues of the characteristic equation are negative if $\Re_0 < 1$. Then, E_0 is locally asymptotically stable.

□

Theorem 5. If $\Re_0 > 1$, then E^* is locally asymptotically stable.

Proof. Let J_E^* be the Jacobian matrix of the system (1) evaluated at E_1,

$$J_{E^*} = \begin{pmatrix} -\lambda M^* - \mu & 0 & -\lambda S^* \\ \lambda M^* & -\mu - \delta & \lambda S^* \\ 0 & 0 & \beta - \frac{2\beta}{K} M^* - \gamma \end{pmatrix}.$$

We put $d_1 = -\lambda M^* - \mu < 0$, $d_2 = -\lambda M^* - \mu$ and $d_3 = \beta - \gamma - \frac{2\beta}{K} M^*$. Then, we consider the following matrix:

$$T_2 = \begin{pmatrix} d_1 - r & 0 & -\lambda S^* \\ \lambda M^* & d_2 - r & \lambda S^* \\ 0 & 0 & d_3 - r \end{pmatrix}.$$

The characteristic equation of T_2 can be written as

$$(d_1 - r)(d_2 - r)(d_3 - r) = 0.$$

Hence, the eigenvalues of J_{E^*} are $r_1 = -d_1$, $r_2 = -d_2$, and $r_3 = d_3$, where

$$d_3 = \beta - \gamma - \frac{2\beta}{K} M^* = \gamma\,(1 - \Re_0)\,.$$

Clearly, all the eigenvalues of the characteristic equation are negative if $\Re_0 > 1$. Then, E^* is locally asymptotically stable if $\Re_0 > 1$, proving the desired result.

□

4.2 Global stability analysis

Now, our aim is to evaluate the global asymptotic stability of both the free equilibrium E_0 and the endemic equilibrium E^* of model (1).

Theorem 6. If $\Re_0 \leq 1$, then E_0 is globally asymptotically stable on Γ.

Proof. We consider the Lyapunov function $V(S, I, M) = M$. We have

$$\begin{aligned} \dot{V} = & \quad \dot{M} \\ = & \quad \beta M \left(1 - \frac{M}{K}\right) - \gamma M \\ = & \quad \gamma \left(\Re_0 - 1\right) M - \frac{\beta}{K} M^2. \end{aligned}$$

Therefore, it is clearly that if $\Re_0 \leq 1$, then $\dot{V} \leq 0$.

Indeed, if $\Re_0 \leq 1$, then $\dot{V}(S, I, M) = 0 \Leftrightarrow M = 0$.

Hence, according to LaSalle's invariance principle [11], the free equilibrium point E_0 is globally asymptotically stable on Γ. □

Theorem 7. If $\Re_0 > 1$, then E^* is globally asymptotically stable on Γ.

Proof. To investigate the global stability of the endemic equilibrium E^*, we examine the Lyapunov function represented as

$$V(S, I, M) = M - M^* \left(1 + \ln(\frac{M}{M^*})\right). \tag{10}$$

Then, the time derivative of the Lyapunov function is given by

$$\begin{aligned} \dot{V} = & \quad \frac{1}{M} \left(M - M^*\right) \dot{M} \\ = & \quad \frac{1}{M} \left(M - M^*\right) \left(\beta M \left(1 - \frac{M}{K}\right) - \gamma M\right) \\ = & \quad \left(M - M^*\right) \left(\beta - \frac{\beta}{K} M - \gamma\right) \\ = & \quad \left(M - M^*\right) \left(\beta - \gamma(\Re_0 - 1) \frac{M}{M^*} - \gamma\right) \\ = & \quad \gamma \left(M - M^*\right) \left(\Re_0 - 1\right) \left(1 - \frac{M}{M^*}\right) \\ = & \quad \gamma \left(1 - \Re_0\right) \frac{\left(M - M^*\right)^2}{M^*}. \end{aligned}$$

Therefore, it is obvious that if $\Re_0 > 1$, then $\dot{V} < 0$.

In addition, if $\Re_0 > 1$, then $\dot{V}(S, I, M) = 0 \Leftrightarrow M = M^*$.

Hence, by LaSalle's invariance principle [11], the endemic equilibrium E^* is globally asymptotically stable on Γ. □

5 Numerical simulation

In this section, we present several numerical solutions of system (1) using approximate parameter values, given the lack of actual data resulting from the absence of statistical studies on this phenomenon. The system (1) was solved using Euler's method. We employed different initial values for each state variable and utilized the parameter set: $\Lambda = 700$ [cacti/day], $\mu = 0.05$ [cacti/day], $\lambda = 0.02$ [cochineal/day], $\delta = 0.03$ [cacti/day], $\beta = 0.02$ [cochineal/day], $\gamma = 0.05$ [cochineal/day], and $K = 5000$ [cochineal/day]. It is noteworthy that we identified the disease-free equilibrium point $E_0 = (14000, 0, 0)$ with $\Re_0 = 0.4 < 1$.

The provided figures illustrate the evolution of the state variables S, I, and M over time (in days), representing the susceptible cacti, infected cacti, and cochineal insects, respectively. Three scenarios with different initial values are depicted in each figure:

- **Figure 2 (S): Evolution of susceptible cacti**
  - $S(0) = 150{,}000$ (blue line): The number of susceptible cacti decreases rapidly and stabilizes around 14,000 over time.
  - $S(0) = 100{,}000$ (orange line): Similarly, the population decreases and reaches the same steady-state value of 14,000.
  - $S(0) = 50{,}000$ (yellow line): Despite a lower initial population, the system stabilizes at the same steady-state level.

  This figure demonstrates that the system consistently converges to $S = 14{,}000$ regardless of initial conditions.

- **Figure 2 (I): Evolution of infected cacti**

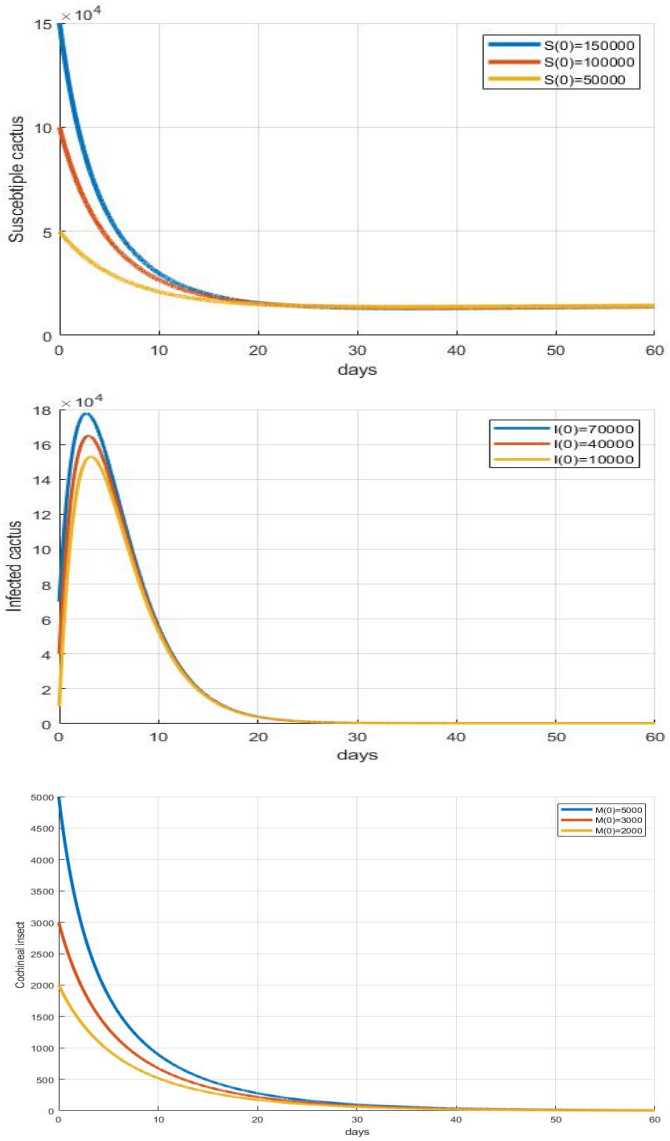


Figure 2: SIM model with a disease-free equilibrium

– $I(0) = 70{,}000$ (blue line): The number of infected cacti increases initially, reaching a peak near 18×10^4 before gradually declining to zero.

- $I(0) = 40{,}000$ (orange line): Following a similar pattern, the population rises to a peak before declining to zero.
- $I(0) = 10{,}000$ (yellow line): Even with smaller initial infections, the population exhibits the same behavior, with infection eradication over time.

This figure highlights that infection is eliminated in all scenarios as the system stabilizes.

- **Figure 2 (M): Evolution of cochineal insects**
  - $M(0) = 5{,}000$ (blue line): The cochineal insect population decreases steadily, approaching zero within approximately 60 days.
  - $M(0) = 3{,}000$ (orange line): Similarly, the population declines to zero at a slightly faster rate.
  - $M(0) = 2{,}000$ (yellow line): A smaller initial population also diminishes to zero over time.

  This figure demonstrates that the cochineal insect population is eradicated in all scenarios.

Across all figures, the application of the model confirms that the system is globally stable at $E_0 = (14{,}000, 0, 0)$. This stability is achieved regardless of initial conditions, as long as $\Re_0 < 1$. The results highlight the robustness of the system's dynamics in eliminating infections and cochineal insects while stabilizing the susceptible cactus population. According to Theorem 6, E_0 is stable on Γ.

Additionally, considering diverse initial values for each state variable and the set of parameters: $\Lambda = 700$ [cacti/day], $\mu = 0.05$ [cacti/day], $\lambda = 0.02$ [cochineal/day], $\delta = 0.03$ [cacti/day], $\beta = 0.07$ [cochineal/day], $\gamma = 0.01$ [cochineal/day], and $K = 5000$ [cochineal/day], we observe that the endemic equilibrium point $E^* = (20000, 72000, 3000)$ and $\Re_0 = 7 > 1$.

The provided figures illustrate the evolution of the state variables S, I, and M over time (in days), representing the susceptible cacti, infected cacti, and cochineal insects, respectively. Three scenarios with different initial values are depicted in each figure:

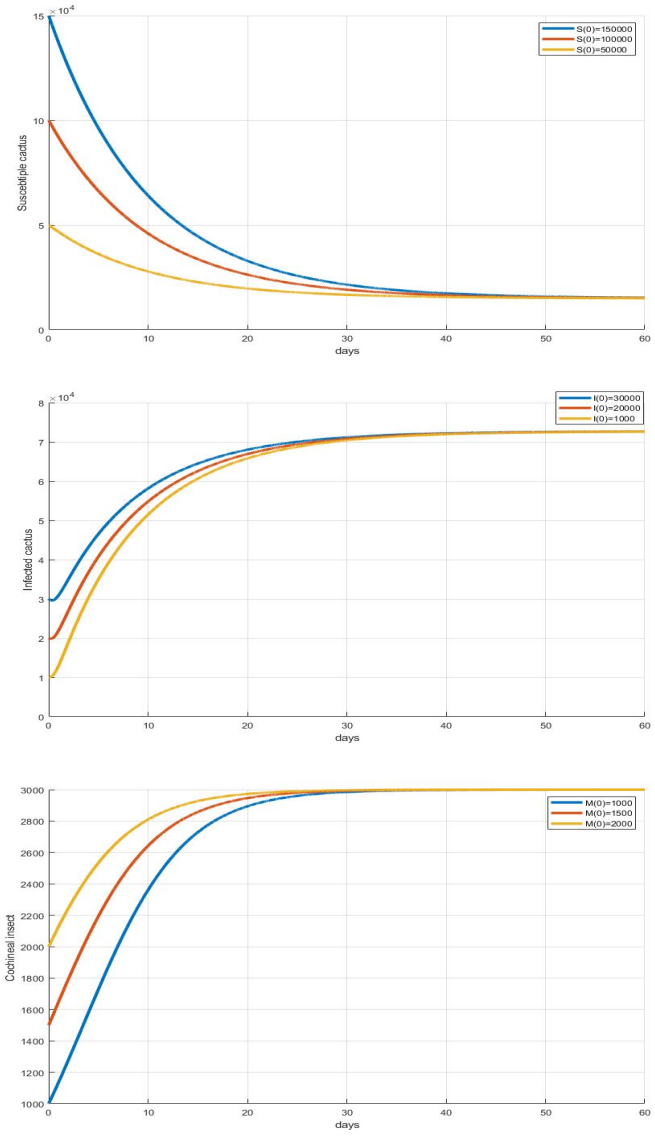


Figure 3: SIM model with an epidemic equilibrium

- **Figure 3 (S): Evolution of Susceptible Cacti**

  - $S(0) = 150{,}000$ (blue line): The susceptible cactus population decreases over time, stabilizing at a value near 20,000.

  - $S(0) = 100{,}000$ (orange line): The population follows a similar trend, reaching the same steady-state value of $S \approx 20{,}000$.
  - $S(0) = 50{,}000$ (yellow line): Despite a lower initial population, the system stabilizes at the same steady-state value.

- **Figure 3 (I): Evolution of Infected Cacti**
  - $I(0) = 30{,}000$ (blue line): The number of infected cacti rises and stabilizes at a value near 72,000.
  - $I(0) = 20{,}000$ (orange line): Similarly, the population stabilizes at $I \approx 72{,}000$ after an initial increase.
  - $I(0) = 10{,}000$ (yellow line): Even with a smaller starting population, the system converges to the same equilibrium.

- **Figure 3 (M): Evolution of Cochineal Insects**
  - $M(0) = 1{,}000$ (blue line): The population of cochineal insects grows steadily, approaching a value near 3,000 over time.
  - $M(0) = 1{,}500$ (orange line): Similarly, the population stabilizes near 3,000 after initial growth.
  - $M(0) = 2{,}000$ (yellow line): Regardless of the higher initial population, the system stabilizes at $M \approx 3{,}000$.

Over time, the system stabilizes with the susceptible cactus population S close to 20,000, the infected cactus population I near 72,000, and the cochineal insect population M at approximately 3,000. This behavior demonstrates that, for $\Re_0 > 1$, the solution curves converge towards the equilibrium point E^*. According to Theorem 7, the equilibrium E^* is stable on Γ.

6 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, $I(t)$, while maximizing the number of recovered plants, $R(t)$, over the period $[t_0, t_f]$. Additionally, the model aims to minimize the costs associated with control measures, ensuring a balance between effective pest management and economic feasibility.

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 $t \in [t_0, t_f]$, as follows:

$$\begin{cases} \frac{dS}{dt} &= \Lambda - \lambda S(t)M(t) - \mu S(t), \\ \frac{dI}{dt} &= \lambda S(t)M(t) - \delta I(t) - \mu I(t) - u_1(t)I(t), \\ \frac{dM}{dt} &= \beta M(t)\left(1 - \frac{M(t)}{K}\right) - \gamma M(t) - u_2(t)M(t), \end{cases} \tag{11}$$

with the initial conditions $S(0) \geq 0$, $I(0) \geq 0$, and $C(0) \geq 0$.

The control $u_1(t)$ represents cutting and burning cactus infected with cochineal, while the control $u_2(t)$ denotes the application of insecticide to combat cochineal.

The problem is to minimize the objective functional:

$$J(u_1, u_2) = I(t_f) + M(t_f) + \int_{t_0}^{t_f} \left[I(t) + M(t) + \frac{C_1}{2}\left(u_1(t)\right)^2 + \frac{C_2}{2}\left(u_2(t)\right)^2 \right] dt, \tag{12}$$

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

$$U = \left\{ (u_1(t), u_2(t)) : 0 \leq u_1(t) \leq 1 \ , 0 \leq u_2(t) \leq 1, \ / \ t \in [t_0, t_f] \right\}.$$

6.1 Existence of optimal controls

In this part, we present the theorem that proves the existence of an optimal control $(u_1^*,\, u_2^*)$ minimizing the cost function J.

Theorem 8. 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 [7], we must check the following properties:

(A_1): The set of controls and the corresponding state variables is nonempty.

(A_2): The control set U is convex and closed.

(A_3): The right-hand side of the state system is bounded by a linear function in the state and control variables.

(A_4): The integral $L(I, M, u_1, u_2)$ of the objective functional is convex on U, and there exist constants $\varkappa_1 > 0$, $\varkappa_2 > 0$, and $\varepsilon > 1$ such that

$$L(I, M, u_1, u_2) \geq -\varkappa_1 + \varkappa_2 \left(|u_1|^2 + |u_2|^2\right)^{\frac{\varepsilon}{2}}.$$

The first condition (A_1) is verified using the result in Lukes. The set U is convex and closed by definition, thus the condition (A_2) is satisfied. 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) is satisfied. Also, we have the last needed condition (A_4),

$$L(I, M, u_1, u_2) \geq -\varkappa_1 + \varkappa_2 \left(|u_1|^2 + |u_2|^2\right)^{\frac{\varepsilon}{2}}.$$

where

$$\varkappa_1 = 2 \sup_{t\in[t_0,t_f]} (I(t), M(t)), \qquad \varkappa_2 = \inf(\frac{C_1}{2}, \frac{C_2}{2}) \text{ and } \varepsilon = 2,$$

since $C_1 > 0$ and $C_2 > 0$. We conclude that there exists an optimal control $(u_1^*, u_2^*) \in U$ such that

$$J(u_1^*, u_2^*) = \min_{(u_1,u_2)\in U} J(u_1, u_2).$$

□

6.2 Characterization of the optimal controls

In this section, we utilize Pontryagin's principle [15]. 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

$$H(t) = I(t) + M(t) + \frac{C_1}{2}\left(u_1(t)\right)^2 + \frac{C_2}{2}\left(u_2(t)\right)^2 + \sum_{i=1}^{3} \lambda_i h_i, \tag{13}$$

where h_i is the right side of the system of differential equations (11) of ith state variable.

Theorem 9. Given the optimal controls (u_1^*, u_2^*) and solutions S^*, I^*, and M^* of the corresponding state system (11), there exist adjoint functions λ_1, λ_2, and λ_3 satisfying

$$\begin{cases} h_1 = \Lambda - \lambda S(t)M(t) - \mu S(t), \\ h_2 = \lambda S(t)M(t) - \delta I(t) - \mu I(t) - u_1(t)I(t), \\ h_3 = \beta M(t)\left(1 - \frac{M(t)}{K}\right) - \gamma M(t) - u_2(t)M(t), \end{cases} \tag{14}$$

such that the transversality conditions at time t_f are

$$\begin{cases} \lambda_1(t_f) = 0, \\ \lambda_2(t_f) = 1, \\ \lambda_3(t_f) = 1. \end{cases} \tag{15}$$

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

$$\begin{cases} u_1^*(t) = \min\left(1, \max\left(0, \ \frac{\lambda_2}{C_1} I(t)\right)\right), \\ u_2^*(t) = \min\left(1, \max\left(0, \frac{\lambda_3}{C_2} M(t)\right)\right). \end{cases} \tag{16}$$

Proof. The Hamiltonian H is defined as follows:

$$H(t) = I(t) + M(t) + \frac{C_1}{2}\left(u_1(t)\right)^2 + \frac{C_2}{2}\left(u_2(t)\right)^2 + \sum_{i=1}^{3} \lambda_i h_i,$$

where

$$\begin{cases} h_1 = \Lambda - \lambda S(t)M(t) - \mu S(t), \\ h_2 = \lambda S(t)M(t) - \delta I(t) - \mu I(t) - u_1(t)I(t), \\ h_3 = \beta M(t)\left(1 - \frac{M(t)}{K}\right) - \gamma M(t) - u_2(t)M(t). \end{cases}$$

For $t \in [t_0, t_f]$, the adjoint equations and transversality conditions can be obtained by using Pontryagin's maximum principle [15] such that

$$\begin{cases} \lambda_1' = -\frac{dH}{dS} = \lambda_1(\lambda M(t) + \mu) - \lambda_2 \lambda M(t), \\ \lambda_2' = -\frac{dH}{dI} = -1 + \lambda_2\left(\delta + \mu + u_1(t)\right), \\ \lambda_3' = -\frac{dH}{dM} = -1 + \lambda_1 \lambda S(t) - \lambda_2 \lambda S(t) - \lambda_3\left(\beta - \frac{2\beta M(t)}{K} - \gamma - u_2(t)\right), \end{cases}$$

for $t \in [t_0, t_f]$, the optimal controls u_1^* and u_2^* can be solved from the optimality condition. We have

$$\frac{dH}{du_1} = C_1 u_1(t) - \lambda_2 I(t) = 0.$$

So

$$u_1(t) = \frac{\lambda_2}{C_1} I(t).$$

We have

$$\frac{dH}{du_2} = C_2 u_2(t) - \lambda_3 M(t) = 0.$$

So

$$u_2(t) = \frac{\lambda_3}{C_2} M(t).$$

By the bounds in U of the controls, it is convenient to obtain u_1^* and u_2^* in the form of (16).

□

7 Optimal control numerical results

In this section, 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 initially 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 cochineal individuals both with and without control measures are generated based on these parameter values: $\Lambda = 1000$ [cacti/day], $\mu = 0.05$ [cacti/day], $\lambda = 0.02$ [cochineal/day], $\delta = 0.03$ [cacti/day], $\beta = 0.07$ [cochineal/day], $\gamma = 0.01$ [cochineal/day], and $K = 70000$ [cochineal/day]. When analyzing the graphs, please be aware that solid lines represent cochineal individuals without control measures, whereas dashed lines indicate those with control measures.

7.1 Control Strategy 1: Cutting and burning cactus infected with cochineal

The goal of this approach is to minimize the function (12), with a primary focus on reducing the cochineal population through cutting and burning cactus infected with cochineal. Figure 4 illustrates the effects of this spraying on the cacti plants.

This figure illustrates the evolution of the state variable I, representing the number of infected cacti, over time (in days). Two scenarios are depicted:

- **Without control** u_1 (solid red line): In this scenario, the number of infected cacti increases rapidly and levels off, approaching a steady-state at approximately 7×10^5 after around 50 days.
- **With control** u_1 (dashed red line): When control u_1 is applied, the number of infected cacti grows initially but peaks at a lower value.

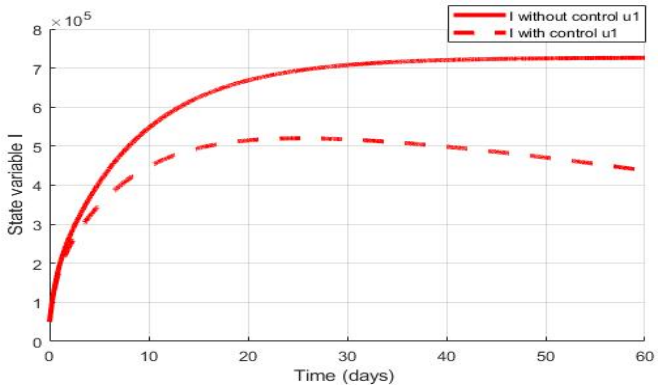


Figure 4: The evolution of the number of infected cacti with and without the control u_1

After reaching the peak, the number steadily declines, indicating that the control measure is effective in reducing the infection.

This comparison demonstrates the significant impact of control u_1 on managing the infection levels among the cacti.

7.2 Control Strategy 2: Effects of insecticide spraying

The main objective of treating cacti infested with cochineal, within the context of a strategy, is to minimize the function (12) while maintaining other control measures at zero. Figure 5 illustrates the cochineal dynamics, taking into account the presence or absence of this control measure.

This figure illustrates the evolution of the state variable M, representing the number of cochineals, over time (in days). Two scenarios are depicted:

- **Without control** u_2 (solid green line): In this scenario, the number of cochineals increases rapidly and levels off, approaching a steady-state at approximately 6×10^5 after around 30 days.

- **With control** u_2 (dashed green line): When control u_2 is applied, the number of cochineals grows initially but stabilizes at a much lower

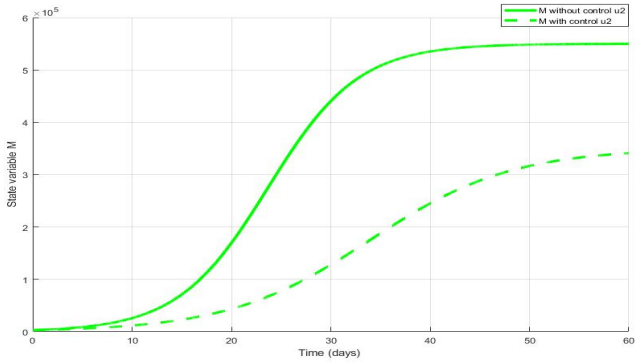


Figure 5: The evolution of the number of cochineals with and without the control u_2

value. This indicates that the control measure effectively limits the growth of the population.

This comparison demonstrates the significant impact of control u_2 in managing the cochineal population.

7.3 Control Strategy 3: Cutting and burning cactus infected with cochineal and insecticide spraying

This strategy aims to minimize the objective function (12) by implementing both control measures. Figure 6 illustrates the disease progression with both controls in effect, compared to the scenario where no control measures are utilized to manage the disease.

This first figure illustrates the evolution of the state variable I, representing the number of infected cacti, over time (in days). Two scenarios are depicted as follows:

- **Without control u_1 and u_2** (solid red line): In this scenario, the number of infected cacti increases rapidly at first and then levels off, approaching a steady-state at approximately 7×10^5 after around 50 days.

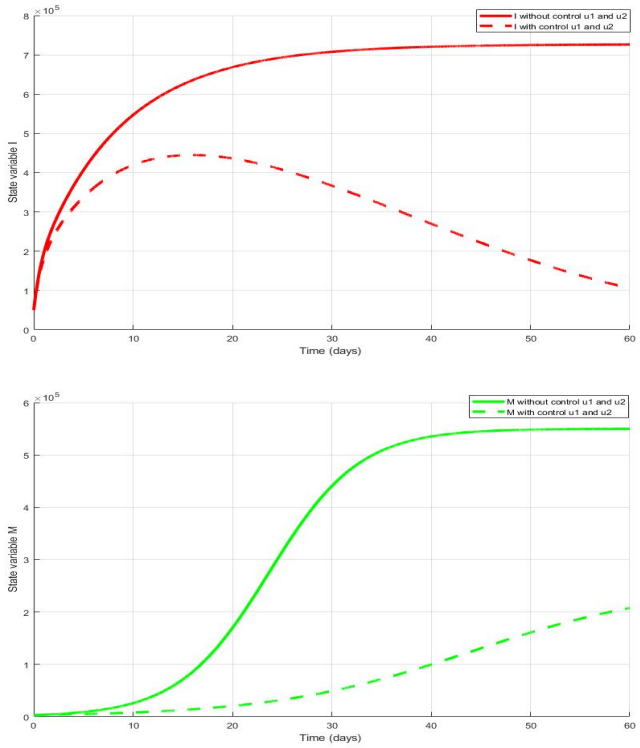


Figure 6: The evolution of the number of infected cacti and cochineals, both with and without u_1 and u_2

- **With control u_1 and u_2** (dashed red line): When control measures are applied, the number of infected cacti initially grows but peaks at a lower value compared to the uncontrolled scenario. After reaching the peak, the number of infected cacti steadily declines, indicating that the controls effectively reduce the infection levels over time.

This comparison highlights the effectiveness of control measures (u_1 and u_2) in managing the spread of the infection among cacti.

The second figure illustrates the evolution of the state variable M, representing the number of cochineals, over time (in days). Two scenarios are depicted:

- **Without control u_1 and u_2** (solid green line): In this scenario, the number of cochineals increases rapidly at first and then levels off, approaching a steady-state at approximately 6×10^5 after around 30 days.

- **With control u_1 and u_2** (dashed green line): When control measures are applied, the number of cochineals grows at a much slower rate and stabilizes at a significantly lower value, suggesting the effectiveness of the controls in limiting the population of cochineals.

This comparison highlights the impact of control measures (u_1 and u_2) in managing the population growth of cochineals.

8 Conclusion

In this paper, we presented a mathematical model, SIM, that describes the spread of the cochineal insect among cacti. We showed that the overall behavior of the model is entirely determined by a critical parameter, the basic reproduction number $\Re_0$. Specifically, we demonstrated that the infection-free equilibrium, E_0, is both locally and globally asymptotically stable when $\Re_0 < 1$, leading to a gradual decline in the cochineal population over time and its eventual eradication. Conversely, when $\Re_0 > 1$, the cochineal population increases. Furthermore, we established that the endemic equilibrium, E^*, is locally and globally asymptotically stable when $\Re_0 > 1$, ensuring the persistence of the infestation.

Our numerical simulations validated the analytical findings regarding the stability of both the infection-free equilibrium E_0 and the endemic equilibrium E^*. These results provide a solid foundation for the development of an optimal control strategy to help farmers minimize losses caused by cochineal infestations in cacti fields.

The model incorporates the dynamics of cochineal infestations and the effects of control measures. We analyzed the system's behavior under three control strategies aimed at managing the cochineal population:

- **Control Strategy 1 (cutting and burning infected cacti, u_1)**: This strategy effectively reduces the number of infected cacti, but the

decline is slow, and the population stabilizes at a higher level compared to other strategies. While effective, it requires continuous intervention over an extended period.

- **Control Strategy 2 (insecticide spraying, u_2)**: Insecticide application significantly accelerates the reduction of the cochineal population. The population stabilizes at a lower level more quickly, reflecting the effectiveness of chemical control in mitigating infestation rates.

- **Control Strategy 3 (combined approach, u_1 and u_2)**: The combined application of cutting and burning infected cacti along with insecticide spraying yields the most substantial reduction in the cochineal population. This strategy results in a faster decline and a lower steady-state population, showcasing the synergistic effect of integrating both physical and chemical control measures.

In conclusion, the combined strategy (Control Strategy 3) proves to be the most effective in reducing the cochineal population, offering a more comprehensive and sustainable solution. However, the choice of the most appropriate strategy depends on factors such as the severity of the infestation, available resources, and environmental considerations. The SIM model provides a valuable tool for evaluating these control measures and supporting decision-making in the management of cochineal infestations on cacti.

Acknowledgements

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

References

[1] Agri-Mag. *Cactus, Cochineal, and Biological Control*, https://www.agri-mag.com/2017/06/22/cactus-cochenille-et-lutte-biologique/#:~:text=Le%20cactus%

20est%20pr%20sent%20dans,%C3%A9cosyst%C3%A8mes%20%C3%A0%20travers%20le%20monde.

[2] 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.

[3] Birkhoff, G., and Rota, G.C. *Ordinary differential equations*, 4th ed., John Wiley and Sons, 1989.

[4] Chowell, G., Simonsen, L., Viboud, C., and Kuang, Y. *Is west Africa approaching a catastrophic phase or is the 2014 Ebola epidemic slowing down? Different models yield different answers for Liberia*, PLoS Currents, 6 (2014) ecurrents-outbreaks.

[5] El Baz, O., Ait Ichou, M., Laarabi, H., and Rachik, M. *Stability analysis of a fractional model for the transmission of the cochineal*, Math. Model. Comput. 10 (2) (2023) 379–386.

[6] Farkas, M., and Weinberger, H.F. *Optimal control theory in mathematical biology*, Springer, 1998.

[7] Fleming, W.H., and Rishel, R.W. *Deterministic and stochastic optimal control*, Springer, 1975.

[8] Gibernau, M. *Introduction to Cochineal and Its Role in Global Trade*, J. Econ. Bot. 28 (3) (2019) 101–105.

[9] Aït Ichou, M., and Laarabi, H. *Application of SIR model to study the spread of cochineal on cacti*, Math. Biosci., 2021.

[10] Khaleque, A., and Sen, P. *An empirical analysis of the Ebola outbreak in west Africa*, Sci. Rep., 7 (2017) 42594.

[11] LaSalle, J.P. *The stability of dynamical systems*, Regional Conference Series in Applied Mathematics, 25, SIAM, 1976.

[12] Lotka, A.J. *Elements of physical biology*, Williams and Wilkins, 1925.

[13] De Jesus Mendez-Gallegos, S., Tiberi, R. and Panzavolta, T. *Carmine Cochineal *Dactylopius Coccus* Costa (Rhynchota: Dactylopiidae): Significance, Production and Use*, Adv. Hortic. Sci. 17 (3) (2003) 165–171.

[14] Pell, B., Kuang, Y., Viboud, C., and Chowell, G. *Using phenomenological models for forecasting the 2015 Ebola challenge*, Epidemics, 22 (2018) 62–70.

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

[16] Rachik, M., and Laarabi, H. *Mathematical modeling of pest dynamics in agricultural systems: A review*, Int. J. Appl. Math. Comput. Sci. 2020.

[17] Verhulst, P.F. *Notice on the law of population growth*, Corresp Math. Phys. 10 (1838) 113–126.

[18] van den Driessche, P., and Watmough, J. *Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission*, Math. Biosci. 180 (2002) 29–48.

Aims and scope

Iranian Journal of Numerical Analysis and Optimization (IJNAO) is publishedby 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. 2, pp 424-851

web site: https://ijnao.um.ac.ir
Email: ijnao@um.ac.ir
ISSN-Print: 2423-6977
ISSN-Online: 2423-6969