Ferdowsi University of Mashhad

Iranian Journal of Numerical Analysis and Optimization

Volume 14, Number 1

March 2024

Serial Number: 28

Ferdowsi University of Mashhad, Iran

In the Name of God

Iranian Journal of Numerical Analysis and Optimization (IJNAO)

This journal is authorized under the registration No. 174/853 dated 1386/2/26 (2007/05/16), by the Ministry of Culture and Islamic Guidance.

Volume 14, Number 1, March 2024

ISSN-Print: 2423-6977, **ISSN-Online:** 2423-6969

Publisher: Faculty of Mathematical Sciences, Ferdowsi University of Mashhad

Published by: Ferdowsi University of Mashhad Press

Printing Method: Electronic

Address: Iranian Journal of Numerical Analysis and Optimization

Faculty of Mathematical Sciences, Ferdowsi University of Mashhad

P.O. Box 1159, Mashhad 91775, Iran.

Tel. : +98-51-38806222 , **Fax:** +98-51-38807358

E-mail: ijnao@um.ac.ir

Website: http://ijnao.um.ac.ir

This journal is indexed by:

- SCOPUS
- ZbMATH Open
- ISC
- DOAJ
- SID
- Civilica
- Magiran
- Mendeley
- Academia.edu
- Linkedin

- The Journal granted the International degree by the Iranian Ministry of Science, Research, and Technology.

Iranian Journal of Numerical Analysis and Optimization

Volume 14, Number 1, March 2024

Ferdowsi University of Mashhad - Iran

©2024 All rights reserved. Iranian Journal of Numerical Analysis and Optimization

Iranian Journal of Numerical Analysis and Optimization

Director

M. H. Farahi

Editor-in-Chief

Ali R. Soheili

Managing Editor

M. Gachpazan

EDITORIAL BOARD

Abbasbandi, Saeid*
(Numerical Analysis)
Imam Khomeini International University, Iran.
e-mail: abbasbandy@ikiu.ac.ir

Abdi, Ali*
(Numerical Analysis)
University of Tabriz, Iran.
e-mail: a_abdi@tabrizu.ac.ir

Area, Iván*
(Numerical Analysis)
Universidade de Vigo, Spain.
e-mail: area@uvigo.es

Babaie Kafaki, Saman*
(Optimization)
Semnan University, Iran.
e-mail: sbk@semnan.ac.ir

Babolian, Esmail*
(Numerical Analysis)
Kharazmi University, Iran.
e-mail: babolian@khu.ac.ir

Cardone, Angelamaria*
(Numerical Analysis)
Università degli Studi di Salerno, Italy.
e-mail: ancardone@unisa.it

Dehghan, Mehdi*
(Numerical Analysis)
Amirkabir University of Technology, Iran.
e-mail: mdehghan@aut.ac.ir

Effati, Sohrab*
(Optimal Control & Optimization)
Ferdowsi University of Mashhad, Iran.
e-mail: s-effati@um.ac.ir

Emrouznejad, Ali*
(Operations Research)
Aston University, UK.
e-mail: a.emrouznejad@aston.ac.uk

Farahi, Mohammad Hadi*
(Optimal Control & Optimization)
Ferdowsi University of Mashhad, Iran.
e-mail: farahi@um.ac.ir

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

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

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

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

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

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

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

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

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

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

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

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

Salehi Fathabadi, Hasan*
(Operations Research)
University of Tehran, Iran.
e-mail: hsalehi@ut.ac.ir

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

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

Toutounian, Faezeh*
(Numerical Analysis)
Ferdowsi University of Mashhad, Iran.
e-mail: toutouni@um.ac.ir

Türkyılmazoğlu, Mustafa*
(Applied Mathematics)
Hacettepe University, Turkey.
e-mail: turkyilm@hacettepe.edu.tr

Vahidian Kamyad, Ali*
(Optimal Control & Optimization)
Ferdowsi University of Mashhad, Iran.
e-mail: vahidian@um.ac.ir

Xu, Zeshui*
(Decision Making)
Sichuan University, China.
e-mail: xuzeshui@263.net

Vasagh, Zohreh
(English Text Editor)
Ferdowsi University of Mashhad, Iran.

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

* Full Professor

** Associate Professor

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

Letter from the Editor–in–Chief

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

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

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

Ali R. Soheili

Editor-in-Chief

Contents

Chebyshev wavelet-based method for solving various stochastic optimal control problems and its application in finance 1
M. Yarahmadi and S. Yaghobipour

A new numerical approach to the solution of the nonlinear Kawahara equation by using combined Taylor–Dickson approximation 20
D. Priyadarsini, M. Routaray and P.K. Sahu

A new approach for ranking decision-making units in data envelopment analysis by using communication game theory . 44
M. Amiri and A. Ashrafi

Efficient numerical methods on modified graded mesh for singularly perturbed parabolic problem with time delay . . . 77
A. Kumar and S. Gowrisankar

Approximate proper solutions in vector optimization with variable ordering structure 107
S. Shahbeyk

A study of global dynamics and sensitivity analysis of a discrete-time model of the COVID-19 epidemic 136
M. E.-B. Keddar and O. Belhamiti

Modal spectral Tchebyshev Petrov–Galerkin stratagem for the time-fractional nonlinear Burgers' equation 172
Y.H. Youssri and A.G. Atta

Minimization of sub-topical functions over a simplex 200
M.H. Daryaei and M.A. Yaghoobi

Richardson extrapolation technique on a modified graded mesh for singularly perturbed parabolic convection-diffusion problems 219
K. K. Sah and S. Gowrisankar

Noisy label relabeling by nonparallel support vector machine 265
A. Sahleh and M. Salahi

On generalized one-step derivative-free iterative family for evaluating multiple roots 291
H. Arora, A. Cordero and J. R. Torregrosa

Wavelet approximation with Chebyshev wavelets 315

S. M. Jesmani, H. Mazaheri and S. Shojaeian

Iranian Journal of Numerical Analysis and Optimization
Vol. 14, No. 1, 2024, pp 1–19
https://doi.org/10.22067/ijnao.2023.82445.1265
https://ijnao.um.ac.ir/

CC BY NC

Research Article

Chebyshev wavelet-based method for solving various stochastic optimal control problems and its application in finance†

M. Yarahmadi*, and S. Yaghobipour

Abstract

In this paper, a computational method based on parameterizing state and control variables is presented for solving Stochastic Optimal Control (SOC) problems. By using Chebyshev wavelets with unknown coefficients, state and control variables are parameterized, and then a stochastic optimal control problem is converted to a stochastic optimization problem. The expected cost functional of the resulting stochastic optimization problem is approximated by sample average approximation thereby the problem can

*Corresponding author

Received 17 May 2023; revised 16 July 2023; accepted 30 August 2023

Majid Yarahmadi
Department of Mathematics and Computer Science, Lorestan University. e-mail: yarahmadi.m@lu.ac.ir

Saba Yaghobipour
Department of Mathematics and Computer Sciences, Lorestan University, Lorestan, Iran. e-mail: sabaphd88@gmail.com.

† This article was suggested by the scientific committee of the 5th national seminar on control and optimization for publication in IJNAO, which was accepted after independent review.

How to cite this article
Yarahmadi, M. and Yaghobipour, S., Chebyshev wavelet-based method for solving various stochastic optimal control problems and its application in finance. *Iran. J. Numer. Anal. Optim.*, 2024; 14(1): 1-19. https://doi.org/10.22067/ijnao.2023.82445.1265

be solved by optimization methods more easily. For facilitating and guaranteeing convergence of the presented method, a new theorem is proved. Finally, the proposed method is implemented based on a newly designed algorithm for solving one of the well-known problems in mathematical finance, the Merton portfolio allocation problem in finite horizon. The simulation results illustrate the improvement of the constructed portfolio return.

AMS subject classifications (2020): Primary 93E20; Secondary 91G80, 91G10.

Keywords: Stochastic optimal control; Chebyshev wavelets; Expansion; Optimal asset allocation.

1 Introduction

Most practical problems in various fields, such as economics and finance, lead us to solve continuous finite-horizon Stochastic Optimal Control (SOC) problems. These problems are driven by Stochastic Differential Equations (SDEs) in mathematical finance and especially in portfolio management, which is started with classical equations such as Merton [10].

The solution of an SDE requires the evaluation of stochastic Itô integral $\int_0^T S(u)dB(u)$, where $\{B(t)\}_t \geq 0$ is a Brownian motion and S is a stochastic process. The common methods for evaluating of the integral, are direct numerical ideas [13]. For example, the approaches of [20, 3] use an optimal wavelet approximation of the Browning motion $\{B(t)\}_t \geq 0$. Also, the approaches of [30, 23, 22] apply the Legendre wavelets to compute numerical solutions of integral and differential equations. For instance, in [22], coefficient functions of the SDE are approximated in terms of Legendre wavelets. Then according to the relation between the block pulse functions and the Itô integral of this function, the integrals of the coefficient function are approximated.

The most common approach for solve SOC problems is the indirect dynamic programming approach in which the corresponding Hamiltonian–Jacobi–Bellman (HJB) equation has to be considered [14, 27, 33]. However, the second-order partial differential HJB equation is usually solved under

some prior assumptions, and it is not solvable analytically in most cases, which is why applying appropriate numerical ideas can be applicable to design optimal decision (control) rules. Among the numerical ideas, the direct ones are more common methods. In these methods, the dynamical system under consideration is discretized in space and time domain, based on the application of the Markov chain. For example, the presented methods of [17, 31] obtain the transition probabilities of the Markov chain from the HJB equation by using finite difference and finite-element ideas. Also, there are some approaches that apply approximation schemes to SOC problems. For instance, the approaches of [4, 16] are Markov decision chain methods from the original SOC problem by approximating the solution of the corresponding SDE and then solving the resulting Bellman equation by using value iteration. Additionally, Huschto and Sager [9] utilized the Wiener chaos expansion developed by Ledoux [19] to reformulate the original SOC problem as a deterministic optimal control problem. Indeed, the considered stochastic process is expressed in terms of deterministic coefficient and orthonormal basis polynomials spanning the underlying Wiener chaos space. Therefore, the approach of [9] is a direct numerical method based on using orthonormal polynomials to solve SOC problems. For this purpose, wavelet polynomials can be used, because they have the property of local approximating discontinuous functions [29]. For example, Haar wavelet orthogonal functions [26], Legendre wavelets [30], and Chebyshev wavelets [35] can be applied to SOC problems, and then the resulting parameterized problems can be solved. Indeed, the application of wavelets as a basis function in a numerical solution of integral equations and optimal control have been attended recently [29, 1, 32]. In comparison with other algorithms, the wavelets techniques provide very fast algorithms. This is because of specific properties when it is used as a basis function. In fact, the wavelets techniques convert SOC problems to stochastic optimization (SO) ones [21, 2, 8] that can be solved by using common methods of SO problems. These methods are commonly divided into two categories, solution methods for single-stage problems and multistage problems [8]. These methods minimize or maximize an expected objective function due to random variables in the formulation of stochastic problems. There are four well-known solution methods, sample average ap-

proximation (SAA) [8, 5], stochastic approximation [18], response surfaces [34], and metamodels [6].

In this paper, a new direct numerical method based on using Chebyshev wavelets is applied to solve SOC problems. For this purpose, stochastic state and control processes are approximated by the finite series of Chebyshev wavelet polynomials. The resulting Chebyshev approximation is the best approximation for the continuous state and control variables. Chebyshev wavelets are well-behaved basic functions that are orthonormal on the interval $[0, 1]$. The advantages of Chebyshev wavelets are that the values of the degree of polynomial M and the number of subintervals 2^k are adjustable. Therefore, it can yield a more accurate approximation than piecewise constant orthogonal functions. Thereby, stochastic state and control processes and the expected cost functional of the SOC problem are parameterized. Therefore, the SOC problem is converted to an SO problem. The resulting SO problem is a single-stage problem. Thus, the parameterized expected cost functional is approximated by sample average approximation thereby the problem is formulated as a deterministic optimal control problem, which can be solved by optimization methods more easily. Also, the proposed method does not require operational matrix that is an obstacle for works which use Chebychev wavelets to convert a differential equation into an algebraic one. Additionally, to solve an SOC problem by the proposed method does not require to solve a partial differential HJB equation, which is a limitation for solving SOC problems, whereas the optimal trajectory and optimal control are approximated properly. The convergence of the proposed method is proved through a new theorem. Finally, the proposed method is implemented for solving one of the well-known problems in mathematical finance, the Merton portfolio allocation problem in finite horizon. The simulation results illustrate the improvement of the constructed portfolio return.

This paper is organized as follows: The mathematical preliminaries of the proposed method are presented in section 2. The Chebyshev wavelet-based method is presented in section 3. The Merton portfolio allocation problem as an application of the proposed method is simulated in section 4. Finally, the conclusion is presented in section 5.

2 Mathematical preliminaries

2.1 Formulation of a stochastic optimal control problem

Assume that $\{X_t\}_{t\in[0,T]}$ is an n-dimensional stochastic process within the probability space $(\Omega, \mathcal{F}, \mathbb{R})$ and that $\{B_t\}_{t\in[0,T]}$ is a d-dimensional Brownian motion. The finite-horizon SOC problem is defined as follows [28]:

$$\min_u E\Big[\int_0^T L(X_t, u_t)dt + G(T, X_T)\Big], \tag{1a}$$

$$s.t. \qquad dX_t = b(X_t, u_t)dt + \sigma(X_t, u_t)dB_t, \tag{1b}$$

$$X(0) = X_0. \tag{1c}$$

The process $\{X_t\}_{t\in[0,T]}$ in $L^2(\Omega \times [0,T])$ is driven by controlled Ito stochastic differential equation (1b) with initial condition (1c). Also, b and σ are describing the drift and diffusion parts of the random state process $X_t \in \mathbb{R}^n$ for all $t \in [0,T]$, respectively. The stochastic control process $\{u_t\} = \{u(\omega, t)\}_{u_t\in\mathbb{R}^n}$ is chosen over a set $\mathcal{A}$ of admissible controls such that the cost function (1a) be minimized.

Additionally, the taken decision at time $t \in [0,T]$ depends on what already has happened up to t. Therefore, the stochastic control process must be $\mathcal{F}_t$-adapted [12]. The admissible control functions can be chosen as "deterministic" controls $u(t,\omega) = u(t)$, "open-loop" $u(t,\omega)$ controls, which are nonanticipative with respect to the Brownian motion $\{B_t\}$ and "Markov" controls $u(t,\omega) = u_M(t, X(\omega))$, where $u_M(\cdot)$ is a nonrandom and Lebesgue-measurable function. The process$\{X_t\}$ becomes an Ito diffusion under Markov control [25]. Also, the function u_M must be smooth.

2.2 Chebyshev wavelets

In this subsection, Chebyshev wavelet polynomials are described, briefly. A family of wavelets can be constituted by dilation and translation of a function Ψ called mother wavelet.

Consider the following family of continuous wavelet:

$$\Psi_{a,b}(t) = |a|^{-\frac{1}{2}}\Psi(\frac{t-a}{b}), \quad a, b \in \mathbb{R}, a \neq 0, \tag{2}$$

where the dilation parameter a and translation parameter b are varying continuously. The family of Chebyshev wavelets $\Psi_{nm}(t) = \Psi(k, n, m, t)$ is one of the applicable wavelets that is defined on interval $[0, 1]$ as follows [29]:

$$\Psi_{nm}(t) = \begin{cases} \frac{\alpha_m 2^{\frac{k}{2}}}{\sqrt{\pi}} T_m(2^{k+1}t - 2n + 1), & \frac{n-1}{2^k} \leqslant t \leqslant \frac{n}{2^k}, \\ 0, & otherwise, \end{cases} \tag{3}$$

where $k \in \mathbb{N}$ and $n = 1, 2, 3, \ldots, 2^k$. Also, $m = 0, 1, 2, \ldots, M-1$ for M-order Chebyshev polynomials and

$$\alpha_m = \begin{cases} \sqrt{2}, & m = 0, \\ 2, & m = 1, 2, \ldots, \end{cases} \tag{4}$$

Additionally, $T_m(t)$'s in (3) are Chebyshev polynomials that can be constructed by the following recursive relations:

$$\begin{cases} T_0(t) = 1, \\ T_1(t) = t, \\ \vdots \\ T_{m+1}(t) = 2tT_m(t) - T_{m-1}(t). \end{cases} \tag{5}$$

3 Solving stochastic optimal control problem via Chebyshev wavelet-based method

In this section, a proposed Chebyshev wavelet-based method for solving general SOC problems is presented. This procedure is presented in three subsections. In subsection 3.1, stochastic state and control process are expanded in Chebyshev Wavelet polynomials. In subsection 3.2, the expected cost function approximation is described. Finally, the convergence analysis of the method is proved in subsection 3.3

3.1 Chebyshev wavelet expansion of stochastic state and control process

Let $Q \subseteq C^1[0,T]$ be the set of all piecewise-continuous functions with initial condition (1c) and let $Q_{2^k,M-1} \subseteq Q$ be the class of combinations of Chebyshev wavelet polynomials of degree up to $(M-1)$.
Consider the Chebyshev wavelet approximations of state process $\{X_t\}_{t\in[0,T]}$ and control process $\{u_t\}$ as follows:

$$\hat{X}(t) = \sum_{n=1}^{2^k} \sum_{m=0}^{M-1} a_{nm} \Psi_{nm}(t), \tag{6}$$

and

$$\hat{U}(t) = \sum_{n=1}^{2^k} \sum_{m=0}^{M-1} c_{nm} \Psi_{nm}(t). \tag{7}$$

Therefore, the SOC problem (1) can be interpreted as a stochastic minimization problem on Q. For this purpose, the interval $[0,T]$ is divided to 2^k subintervals; that is,

$$[0,T] = [0, \frac{1}{2^k}T] \bigcup [\frac{1}{2^k}T, \frac{2}{2^k}T] \bigcup \cdots \bigcup [\frac{2^k-1}{2^k}T, T]. \tag{8}$$

Thus, the state variable (6) and the control variable (7) is rewritten as follows:

$$\hat{X}(t) = \begin{cases} \hat{x}_1(t) = \sum_{m=0}^{M-1} a_{1m}\Psi_{1m}(t), & 0 \leqslant t \leqslant \frac{1}{2^k}T, \\ \hat{x}_2(t) = \sum_{m=0}^{M-1} a_{2m}\Psi_{2m}(t), & \frac{1}{2^k}T \leqslant t \leqslant \frac{2}{2^k}T, \\ \vdots & \\ \hat{x}_{2^k}(t) = \sum_{m=0}^{M-1} a_{2^k m}\Psi_{2^k m}(t), & \frac{2^k-1}{2^k}T \leqslant t \leqslant T. \end{cases} \tag{9}$$

Also,

$$\hat{U}(t) = \begin{cases} \hat{u}_1(t) = \sum_{m=0}^{M-1} c_{1m}\Psi_{1m}(t), & 0 \leqslant t \leqslant \frac{1}{2^k}T, \\ \hat{u}_2(t) = \sum_{m=0}^{M-1} c_{2m}\Psi_{2m}(t), & \frac{1}{2^k}T \leqslant t \leqslant \frac{2}{2^k}T, \\ \vdots & \\ \hat{u}_{2^k}(t) = \sum_{m=0}^{M-1} c_{2^k m}\Psi_{2^k m}(t), & \frac{2^k-1}{2^k}T \leqslant t \leqslant T. \end{cases} \tag{10}$$

Now, from (9) and (10), the cost functional (1a) becomes

$$E[\hat{J}(a_{10},\ldots,a_{2^kM-1},c_{10},\ldots,c_{2^kM-1})] = E\Big[\int_0^T L(\hat{X},\hat{U})dt + G(T,\hat{X}_T)\Big].$$

In other words, we have

$$E[\hat{J}] = E\Big[\int_0^{\frac{1}{2^k}T} L(\hat{x}_1(t),\hat{u}_1(t))dt + \int_{\frac{1}{2^k}T}^{\frac{2}{2^k}T} L(\hat{x}_2(t),\hat{u}_2(t))dt \\ + \cdots + \int_{\frac{2^k-1}{2^k}T}^{T} L(\hat{x}_{2^k}(t),\hat{u}_{2^k}(t))dt + G(T,\hat{x}_{2^k}(T))\Big]. \tag{11}$$

Additionally, from (9) and (10), the controlled Ito SDE (1b) becomes

$$\dot{\hat{X}}_t = b(\hat{X}_t,\hat{U}_t) + \sigma(\hat{X}_t,\hat{U}_t)\dot{B}_t. \tag{12}$$

More precisely, the following constraints are yielded

$$\begin{cases} \dot{\hat{x}}_1(t) = b(\hat{x}_1,\hat{u}_1) + \sigma(\hat{x}_1,\hat{u}_1)\dot{B}_1, \\ \dot{\hat{x}}_2(t) = b(\hat{x}_2,\hat{u}_2) + \sigma(\hat{x}_2,\hat{u}_2)\dot{B}_2, \\ \vdots \\ \dot{\hat{x}}_{2^k}(t) = b(\hat{x}_{2^k},\hat{u}_{2^k}) + \sigma(\hat{x}_{2^k},\hat{u}_{2^k})\dot{B}_{2^k}. \end{cases} \tag{13}$$

Also, according to the initial condition (1c), we have the following constraint:

$$\sum_{m=0}^{M-1} a_{1m}\Psi_{1m}(0) - X_0 = 0. \tag{14}$$

According to the continuity property of the state variable and (8) for

$$t_i = \frac{i}{2^k}, \qquad i = 1,2,\ldots,2^k-1, \tag{15}$$

one can yield

$$\begin{array}{l} \sum_{m=0}^{M-1} a_{1m}\Psi_{1m}(t_1) = \sum_{m=0}^{M-1} a_{2m}\Psi_{2m}(t_1), \\ \sum_{m=0}^{M-1} a_{2m}\Psi_{1m}(t_2) = \sum_{m=0}^{M-1} a_{3m}\Psi_{3m}(t_2), \\ \vdots \\ \sum_{m=0}^{M-1} a_{2^k-1m}\Psi_{2^k-1m}(t_{2^k-1}) = \sum_{m=0}^{M-1} a_{t_{2^k}m}\Psi_{2^k m}(t_{2^k-1}). \end{array} \tag{16}$$

According to the SOC problem (1) and the parameterization process, which was mentioned above, the following SO problem with objective function (11) subject to constraints (13)–(14) and (16) is designed, which can be solved by SO methods [8, 18], more easily.

$$\min E[\hat{J}(\alpha_t,\gamma_t,B_t)], \tag{17a}$$

$$subject\ to: \qquad P[\alpha_t, \gamma_t, B_t]^T = C, \tag{17b}$$

where $\alpha = [a_{10}, \ldots, a_{1M-2}, a_{20}, \ldots, a_{2M-1}, \ldots, a_{2^k 0}, \ldots, a_{2^k M-1}]$, $\gamma = [c_{10}, \ldots, c_{1M-2}, c_{20}, \ldots, c_{2M-1}, \ldots, c_{2^k 0}, \ldots, c_{2^k M-1}]$ and $\{B_t\}$ is the Brownian motion.

3.2 Expected cost function approximation

Consider the SO problem (17). Let $\mathbb{R}^{2k+1M}$ be the domain of all the feasible decisions and let $(\alpha_t, \gamma_t) = \mathcal{X}$ be a specific decision. The goal is to find a decision that minimizes the cost function, $\hat{J}$. Let the Brownian motion $\{\xi_t\}$ denote random information that is available only after a taken decision. Since we cannot directly optimize $\hat{J}(\mathcal{X}, \xi)$, we compute minimize the expected value, $E[\hat{J}(\mathcal{X}, \xi)]$. SAA is the most common approach of a choice for the approximation of $E[\hat{J}(\mathcal{X}, \xi)]$. The first step in SAA is sampling. let $\{\xi_1, \ldots, \xi_n\}$ be a set of independent identically distributed realizations of random process $\{\xi_t\}$ and let $\hat{J}(\alpha_t, \gamma_t, \xi_i)$ be the cost function realization for ξ_i. According to the SAA approach, the expected cost function is approximated by the average of the realizations:

$$E[\hat{J}(\mathcal{X}, \xi)] \approx \frac{1}{n} \sum_{i=1}^{n} \hat{J}(\mathcal{X}, \xi_i). \tag{18}$$

The second step in SAA is the solving an optimal problem. The right-hand side of (18) is deterministic, so deterministic optimization methods can be used for solving the following approximated problem for all $n \in \mathbb{N}$:

$$\beta_n^* = \min_{(\alpha_t, \gamma_t) \in \mathbb{R}^{2k+1M}} E[\hat{J}(\alpha_t, \gamma_t, \xi)] = \min_{\mathcal{X} \in \mathbb{R}^{2k+1M}} \frac{1}{n} \sum_{i=1}^{n} \hat{J}(\mathcal{X}, \xi_i). \tag{19}$$

Deterministic search is the main benefit of SAA [8].

3.3 Convergence analysis

In this section, the convergence of the proposed method is investigated.

Theorem 1. Suppose that $f(x) \in L^2[0,1]$ with bounded second derivative $|\ddot{f}| \leq L$, and let $\sum_{n=0}^{\infty}\sum_{m=0}^{\infty} c_{nm}\Psi_{nm}(x)$ be its infinite Chebyshev wavelet expansion. Then

$$|c_{nm}| \leq \frac{\sqrt{(2\pi)}L}{(2n)^{\frac{5}{2}}(m^2-1)}. \tag{20}$$

This means that the Chebyshev wavelets series converges uniformly to $f(x)$; that is,

$$f(x) = \sum_{n=1}^{\infty}\sum_{m=0}^{\infty} c_{nm}\Psi_{nm}(x). \tag{21}$$

Proof. See [7]. □

Theorem 2. Let $\beta = \inf_{(X,u)\in Q} E[J]$. If $\beta_{2^kM-1} = \inf_{(X,u)\in Q_{2^kM-1}} E[J(X,u)]$ for $k, M = 1, 2, \ldots$, where Q_{2^kM-1} is a subset of Q, then $\lim_{k,M\to\infty} \beta_{2^kM-1} = \beta$.

Proof. Since $(X_t, u_t) \in Q_{2^kM-1}$, by considering the Chebyshev expansions (6) and (7) for stochastic processes $\{X_t\}$ and $\{u_t\}$, respectively, we have

$$\beta_{2^kM-1} = \min_{(\alpha_t,\gamma_t)\in\mathbb{R}^{2k+1M}} E[\hat{J}(\alpha_t, \gamma_t, \xi)]. \tag{22}$$

Now, let $\{\xi_1, \ldots, \xi_n\}$ be a set of independent identically distributed realizations of random process $\{\xi_t\}$ and let $\hat{J}(\alpha_t, \gamma_t, \xi_i)$ be the cost function realization for ξ_i. According to the SAA approach, the expected cost function is approximated by the average of the realizations:

$$E[\hat{J}(\alpha, \gamma, \xi)] \approx \frac{1}{n}\sum_{i=1}^{n} \hat{J}(\alpha, \gamma, \xi_i). \tag{23}$$

If we define

$$\beta^*_{n,2^kM-1} = \min_{(\alpha,\gamma)\in\mathbb{R}^{2k+1M}} \frac{1}{n}\sum_{i=1}^{n} \hat{J}(\alpha, \gamma, \xi_i), \tag{24}$$

then there exists an optimal value $(\alpha^*, \gamma^*, \xi_i)$ for

$$\arg\min\{\frac{1}{n}\sum_{i=1}^{n} \hat{J}(\alpha, \gamma, \xi_i) : (\alpha, \gamma, \xi_i) \in (\mathbb{R}^{2k+1M} \times \mathbb{R})\}, \tag{25}$$

such that

$$\beta^*_{n,2^kM-1} = \frac{1}{n}\sum_{i=1}^{n} \hat{J}(\alpha^*, \gamma^*, \xi_i). \tag{26}$$

Also, one can conclude that there exists a corresponding optimal value (X^*, u^*) from $\arg\min\{\frac{1}{n}\sum_{i=1}^{n} \hat{J}(X,u) : (X,u) \in Q_{2^kM-1}\}$ such that

$$\beta^*_{n,2^kM-1} = \frac{1}{n}\sum_{i=1}^{n} \hat{J}(X^*, u^*). \tag{27}$$

Furthermore, since $Q_{2^kM-1} \subseteq Q_{2^kM}$, we have

$$\beta^*_{n,2^kM} = \frac{1}{n}\sum_{i=1}^{n} \hat{J}(X^*, u^*) = \min_{(X,u)\in Q_{2^kM}} \frac{1}{n}\sum_{i=1}^{n} \hat{J}(X,u) \tag{28}$$

$$\leq \min_{(X,u)\in Q_{2^kM-1}} \frac{1}{n}\sum_{i=1}^{n} \hat{J}(X,u) \tag{29}$$

$$= \beta^*_{n,2^kM-1}. \tag{30}$$

Therefore, for all k, M and n, we have $\beta^*_{n,2^kM} \leq \beta^*_{n,2^kM-1}$. Since, the sequence $\{\beta^*_{n,2^kM}\}$ is a nonincreasing sequence. Additionally, this sequence is upper bounded and therefore is convergent and this completes the proof. □

For implementing of the proposed method, Chebyshev Wavelet-based Method algorithm for solving SOC problems, is presented as follows:

Chebyshev wavelet-based method algorithm for stochastic optimal control problems

Input : SOC problem (1).

Output : The approximated Stochastic optimal trajectory and approximated optimal control.

Step 1 : Choose N, k, and M, and set $i = 1$

Step 2 : Approximate the state and control variable by Chebyshev wavelet series from equations (9) and (10).

Step 3 : Find an expression of $E[\hat{J}]$ from equation (11).

Step 4 : Determine the set of equality constraints, by (13), (14), and (16).

Step 5 : Consider a set of independent identically distributed realizations of Brownian motion $\{B_t\}$ and approximate the expected cost function $E[\hat{J}(\alpha, \gamma, B)]$ from (18).

Step 6 : Determine the optimal parameters $[\alpha^*, \gamma^*]$ by solving the optimization problem with the resulting cost function of **Step 5** subject to (17b) and substitute these parameters into (9) and (10) to find the approximated optimal trajectory, approximated optimal control.

Step 7 : If $i = N$ go to **Step 8** else set $i = i + 1$ and go to **Step 2**.

Step 8 : End.

4 Simulation results

In this section, one of the well-known problems in continuous-time finance, the Merton portfolio allocation problem, is considered for illustrating the advantages and efficiency of the proposed method.

Example 1 (Merton portfolio allocation problem in finite horizon[15])**.** Consider an investing process on time interval $[0, T]$, and denote wealth at time $t \in [0, T]$ by W_t. Also, suppose that this investing is started with a known initial wealth W_0. At time t, we must choose a fraction of wealth to invest in a (set of risky assets) stock portfolio $U_1(t)$ whereas the remaining fraction,

$1-U_1(t)$ is invested in a (the risk-free asset) bond . Additionally, the variable $U_2(t)$ is considered as wealth consumption value at time t. Therefore, the control vector has two component, $0 \leq U_1(t) \leq 1$ and $U_2(t) \geq 0$.

Let the interest rate for risky and risk-free investment be R and r, respectively, such that $0 < r < R$. Thus, the bond price b_t and the stock price S_t evolve according to the following Black-Scholes models:

$$\frac{db(t)}{b} = rdt, \qquad \frac{dS_t}{S} = Rdt + \sigma dB_t, \tag{31}$$

where σ is a real positive constant that denotes the volatility of S, and B_t is a Brownian motion described. The objective is to find an optimal two-dimensional control vector $(U_1(t), U_2(t))$, which maximizes the following expected cost functional:

$$E[J] = E\Big[\int_0^T e^{-\beta t} F(U_2(t))dt\Big] \tag{32}$$

subject to

$$\begin{aligned} dW_t &= \tfrac{db(t)}{b} W_t + \tfrac{dS(t)}{S} W_t - U_2(t)dt \\ &= [(1-U_1(t))r + U_1 R]W_t dt + U_1(t)\sigma W_t dB_t - U_2(t)dt, \end{aligned} \tag{33}$$

$$W(0) = W_0, \tag{34}$$

where β and F are discount rate and utility function, respectively.

Consider the utility function $F(U_2(t)) = U_2^{\alpha}(t), 0 < \alpha < 1$. This problem is solved by the proposed method for $k = 1, M = 3, r = 0.05, R = 0.11, \alpha = 0.5, \beta = 0.11, W_0 = 10^5$ and $\sigma = 0.4$ on the interval $[0, 1]$. Therefore, the state and control variables are approximated as follows:

$$\hat{W}(t) = \begin{cases} \hat{w}_1(t) = \sum_{m=0}^{2} a_{1m}\Psi_{1m}(t), & 0 \leqslant t \leqslant \frac{1}{2}, \\ \hat{w}_2(t) = \sum_{m=0}^{2} a_{2m}\Psi_{2m}(t), & \frac{1}{2} \leqslant t \leqslant 1, \end{cases} \tag{35}$$

$$\hat{U_1}(t) = \begin{cases} \hat{u}_1(t) = \sum_{m=0}^{2} c_{1m}\Psi_{1m}(t), & 0 \leqslant t \leqslant \frac{1}{2}, \\ \hat{u}_2(t) = \sum_{m=0}^{2} c_{2m}\Psi_{2m}(t), & \frac{1}{2} \leqslant t \leqslant 1, \end{cases} \tag{36}$$

$$\hat{U_2}(t) = \begin{cases} \hat{v}_1(t) = \sum_{m=0}^{2} d_{1m}\Psi_{1m}(t), & 0 \leqslant t \leqslant \frac{1}{2}, \\ \hat{v}_2(t) = \sum_{m=0}^{2} d_{2m}\Psi_{2m}(t), & \frac{1}{2} \leqslant t \leqslant 1, \end{cases} \tag{37}$$

where

$$\Psi(t) = [\Psi_{10}(t), \Psi_{11}(t), \Psi_{12}(t), \Psi_{20}(t), \Psi_{21}(t), \Psi_{22}(t)],$$
$$a(t) = [a_{10}, a_{11}, a_{12}, a_{20}, a_{21}, a_{22}],$$
$$\gamma(t) = [c_{10}, c_{11}, c_{12}, c_{20}, c_{21}, c_{22}, d_{10}, d_{11}, d_{12}, d_{20}, d_{21}, d_{22}].$$

From (35) and (36), the expected cost functional (32) is parameterized as follows:

$$E[\hat{J}] = E\Big[\int_0^{\frac{1}{2}} e^{-\beta t}\hat{v}_1^{\alpha}(t)dt + \int_{\frac{1}{2}}^{1} e^{-\beta t}\hat{v}_2^{\alpha}(t)dt\Big]. \tag{38}$$

By substituting (35) and (36) into the SED equation (33) and applying the differential transformation $\frac{dB}{dt} = t_k B(t_k)$ [24], the following equality constraints can be obtained:

$$\begin{cases} [\dot{\hat{w}}_1 - [(1-\hat{u}_1)r + \hat{u}_1 R]\hat{w}_1]_{t=t_k} + [\hat{u}_1\sigma\hat{w}_1]_{t=t_k} t_k B_k - \hat{v}_1(t_k) = 0, t_k \in \{0, \frac{1}{2}\}, \\ (\dot{\hat{w}}_2 - [(1-\hat{u}_2)r + \hat{u}_2 R]\hat{w}_2)_{t=t_k} + [\hat{u}_2\sigma\hat{w}_2]_{t=t_k} t_k B_k - \hat{v}_2(t_k) = 0, t_k \in \{\frac{1}{2}, 1\}. \end{cases} \tag{39}$$

The continuity of state variable must be satisfied at point $t_1 = \frac{1}{2}$. Therefore, the following equality constraint is considered as well:

$$\hat{w}_1(t_1) - \hat{w}_2(t_1) = 0. \tag{40}$$

Also, from initial condition (34) another constraint is produced as follows:

$$\hat{w}_1(0) - 10^5 = 0. \tag{41}$$

Now, consider a Brownian motion, $B_t \sim N(0,1)$. Let $\{B_t(t_1), B_t(t_2), B_t(t_3)\}$ be a sample of random process B_t. According to the SAA method, the expected cost functional (38) is approximated as

$$E[\hat{J}(a,\gamma,B_t)] \approx \frac{1}{3}\sum_{i=1}^{3} \hat{J}(a,\gamma,B_i). \tag{42}$$

By maximizing (42) subject to constraints (39)–(41), the optimal vector $[a^*, \gamma^*]$ is obtained. Therefore, the optimal approximated trajectory of wealth $\{W_t\}$ and the optimal approximated control $\{U_2(t)\}$ can be obtained. The simulation results are shown in Figures 1 and 2.

In the following, the simulation results based on presented method in this paper are comprised with the simulation result of method proposed in [11].

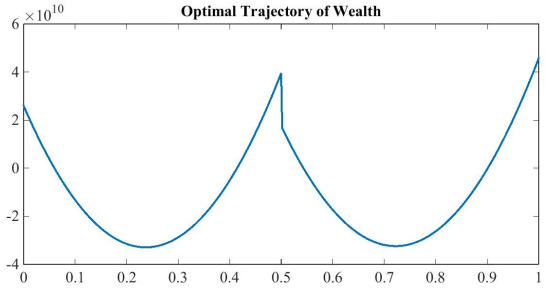


Figure 1: The wealth trajectory by Chebyshev wavelet-based method.

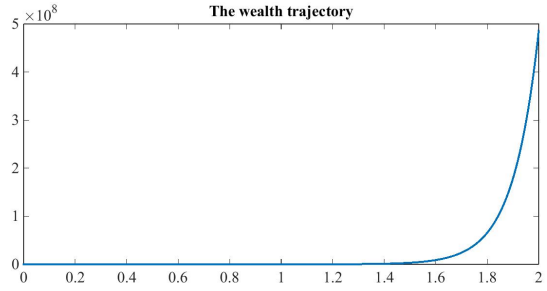


Figure 2: The wealth trajectory by method proposed in [11]

The optimal trajectory of the wealth $\{W_t\}$ by the Chebyshev wavelet-based method and method proposed in [11] is given in Figures 1 and 2, respectively. Figure 1 shows that the Chebyshev wavelet-based portfolio increases the initial wealth value from $10^5\$$ to $5*10^{10}\$$ while according to Figure 2, the proposed method in [11] has increased it to $5 * 10^8\$$. Therefore, from comparison point of view, the resulting portfolio of Chebyshev wavelet-based method has a higher return.

5 Conclusion

In this paper, a new Chebyshev wavelet-based algorithm, which is the state-control parameterization method for solving problems, has been presented. The state and control process were parameterized via orthogonal Chebyshev polynomial basis functions. Therefore, the SOC problem was converted to

an SO, and an optimal approximation of the solution was produced without requiring compute the operational matrix of the derivative. An advantage of the proposed method is that an SOC problem is solved without encountering a partial differential HJB equation. The convergence of the proposed method was proved via a new theorem. One of the well-known problems in mathematical finance, the Merton portfolio allocation problem in finite horizon, was simulated by the proposed method. The simulation results showed the capability and efficiency of the proposed method in comparison with other similar works. For instance, the return of the constructed portfolio was improved by the proposed method.

Acknowledgements

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

References

[1] Aidoo, A.Y. and Wilson, M. *A review of wavelets solution to stochastic heat equation with random inputs*, Applied Mathematics, 6(14) (2015) 2226–2239.

[2] Aoki, M. *Optimization of stochastic systems. Topics in discrete-time systems*, Mathematics in Science and Engineering, Vol. 32. Academic Press, New York-London, 1967.

[3] Ayache, A. and Taqqu, M.S. *Rate optimality of wavelet series approximations of fractional Brownian motion*, J. Fourier Anal. Appl. 9(5), (2003) 451–471.

[4] Azzato, J.D. and Krawczyk, J. *An improved MATLAB package for approximating the solution to a continuous-time stochastic optimal control problem*, Working paper of the Victoria University of Wellington, 2006.

[5] Cherukuri, A. *Sample average approximation of conditional value-at-risk based variational inequalities*, Optim. Lett. (2023) 1–26.

[6] Cinquegrana, D., Zollo, A.L., Montesarchio, M. and Bucchignani, E. *A Metamodel-Based Optimization of Physical Parameters of High Resolution NWP ICON-LAM over Southern Italy.* Atmosphere 14(5), (2023) 788.

[7] Guariglia, E. and Guido, R.C. *Chebyshev wavelet analysis,* J. Funct. Spaces 2022, Art. ID 5542054, 17 pp.

[8] Hannah, L.A. *Stochastic optimization*, International Encyclopedia of the Social & Behavioral Sciences, 2 (2015) 473–481.

[9] Huschto, T. and Sager, S. *Solving stochastic optimal control problems by a Wiener chaos approach*, Vietnam J. Math. 42(1), (2014) 83–113.

[10] Jadamba, B., Khan, A.A., Migórski, S. and Sama, M. eds. *Deterministic and stochastic optimal control and inverse problems*, CRC Press, Boca Raton, FL, 2022.

[11] Kafash, B. and Nadizadeh, A. *Solution of stochastic optimal control problems and financial applications*, J. Math. Ext. 11 (2017) 27–44.

[12] Kappen, H.J. *Stochastic optimal control theory*, ICML, Helsinki, Radbound University, Nijmegen, Netherlands, 2008.

[13] Kloeden, P.E. and Platen, E. *The numerical solution of stochastic differential equations*, 3rd edn. Springer-Verlag Berlin Heidelberg, 2011.

[14] Korn, R. and Korn, E. *Option pricing and portfolio optimization: Modern methods of financial mathematics*, Vol. 31. American Mathematical Soc., 2001.

[15] Kraft, H., Meyer-Wehmann, A. and Seifried, F.T.. Holger, K., *Dynamic asset allocation with relative wealth concerns in incomplete markets*, J. Econom. Dynam. Control 113 (2020), 103857, 20 pp.

[16] Krawczyk, J.B. *A Markovian approximated solution to a portfolio management problem*,ITEM. Inf. Technol. Econ. Manag. No. 1 (2001), Paper 2, 33 pp.

[17] Kushner, H.J. and Dupuis, P. *Numerical methods for stochastic control problems in continuous time,* Second edition. Applications of Mathematics (New York), 24. Stochastic Modelling and Applied Probability. Springer-Verlag, New York, 2001.

[18] Lan, G. and Zhou, Z. *Dynamic stochastic approximation for multi-stage stochastic optimization*, Math. Program. 187(1-2), (2021), Ser. A, 487–532.

[19] Ledoux, M. *A note on large deviations for wiener chaos*, Séminaire de Probabilités, XXIV, 1988/89, 1–14, Lecture Notes in Math., 1426, Springer, Berlin, 1990.

[20] Lisei, H. and Soos, A. *Wavelet approximation of the solutions of some stochastic differential equations*, Fourth Joint Conference on Applied Mathematics. Pure Math. Appl. 15 (2004), no. 2-3, 213–223 (2005).

[21] Marti, K. *Stochastic optimization methods*, Springer-Verlag, Berlin, 2005.

[22] Mohammadi, F. *A efficient computational method for solving stochastic Itô-Volterra integral equations*, TWMS J. Appl. Eng. Math. 5(2), (2015) 286–297.

[23] Mohammadi, F. and Hosseini, M. *A new Legendre wavelet operational matrix of derivative and its applications in solving the singular ordinary differential equations*, J. Franklin Inst. 348(8), (2011) 1787–1796.

[24] Odibat, Z., Erturk, V.S., Kumar, P., Ben Makhlouf, A. and Govindaraj, V., *An implementation of the generalized differential transform scheme for simulating impulsive fractional differential equations*, Hindawi Mathematical Problems in Engineering, 2022, Article ID 8280203, (2022) 11 pages.

[25] Øksendal, B. *Stochastic differential equations. An introduction with applications*,Sixth edition. Universitext. Springer-Verlag, Berlin, 2003.

[26] Pargaei, M. and Kumar, B.V. *A 3D Haar wavelet method for a coupled degenerate system of parabolic equations with nonlinear source coupled with non-linear ODEs*, Appl. Numer. Math. 185 (2023), 141–164.

[27] Petkovi, N. and and Božinović, M. *The application of the dynamic programming method in investment optimization*,Megatrend revija, 13(3) (2016) 171–182.

[28] Pham, H. *Continuous-time stochastic control and optimization with financial applications*, Stochastic Modelling and Applied Probability, 61. Springer-Verlag, Berlin, 2009.

[29] Rafiei, Z., Kafash, B. and Karbassi S.M. *A new approach based on using Chebyshev wavelets for solving various optimal control problems*, Comput. Appl. Math. 37 (2018), S144–S157.

[30] Razzaghi, M. and Yousefi, S. *The Legendre wavelets operational matrix of integration*, Internat. J. Systems Sci. 32(4), (2001) 495–502.

[31] Simpkins, A. and Todorov, E. *Practical numerical methods for stochastic optimal control of biological systems in continuous time and space*, In 2009 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning, pp. 212–218. IEEE, (2009).

[32] Sohrabi, S. *Comparison Chebyshev wavelets method with BPFs method for solving Abel's integral equation*, Ain Shams Eng. J. 2(3-4), (2011) 249–254.

[33] Tourin, A. and Yan, R. *Dynamic pairs trading using the stochastic control approach*, J. Econom. Dynam. Control 37(10) (2013) 1972–1981.

[34] Yang, R., Li, W. and Liu, Y. *A novel response surface method for structural reliability*, AIP Advances 12, (2022) 015205.

[35] Yang, Y., Heydari, M., Avazzadeh, Z. and Atangana, A., *Chebyshev wavelets operational matrices for solving nonlinear variable-order fractional integral equations*, Adv. Difference Equ. 2020, Paper No. 611, 24 pp.

Iranian Journal of Numerical Analysis and Optimization
Vol. 14, No. 1, 2024, pp 20–43
https://doi.org/10.22067/ijnao.2023.83184.1287
https://ijnao.um.ac.ir/

CC BY NC

Research Article

A new numerical approach to the solution of the nonlinear Kawahara equation by using combined Taylor–Dickson approximation

D. Priyadarsini, M. Routaray and P.K. Sahu*

Abstract

This article presents a novel numerical approach to the solution of the nonlinear Kawahara equation. The desired approximations are obtained from the combination of Dickson polynomials and Taylor's expansion. The combined approach is based on Taylor's expansion for discretizing the time

*Corresponding author

Received 28 June 2023; revised 28 August 2023; accepted 10 September 2023

Dibyajyoti Priyadarsini
Department of Mathematics, School of Applied Sciences, KIIT University, Odisha-751024, India. e-mail: pdibyajyoti7@gmail.com

Mitali Routaray
Department of Mathematics, School of Applied Sciences, KIIT University, Odisha-751024, India. e-mail: mitaray8@gmail.com

Prakash Kumar Sahu
Department of Mathematics, School of Applied Sciences, KIIT University, Odisha-751024, India. e-mail: prakash.2901@gmail.com

How to cite this article
Priyadarsini, D., Routaray, M., and Sahu, P.K., A new numerical approach to the solution of the nonlinear Kawahara equation by using combined Taylor–Dickson approximation. *Iran. J. Numer. Anal. Optim.*, 2024; 14(1): 20-43. https://doi.org/10.22067/ijnao.2023.83184.1287

derivative and Dickson polynomials for space derivatives. The problem will be converted into a system of linear algebraic equations for each time step via some suitable collocation points. Error estimation is presented after obtaining the approximate solution. The newly proposed technique is compared with some existing numerical methods to show the method's applicability, accuracy, and efficacy. Two problems are solved to demonstrate the method's power and effect, and the results are presented as a table and graphics.

AMS subject classifications (2020): 35C11; 35G20; 65N22.

Keywords: Taylor's expansion; Dickson polynomials; Kawahara equation; Discretization.

1 Introduction

Partial differential equations (PDEs), including linear and nonlinear PDEs, are used to describe many real-world phenomena [9, 25, 16]. We must solve differential equations to see the nature of the background of these phenomena. In general, we are better at dealing with linear problems than nonlinear problems. However, in practice, we face an increasing number of nonlinear problems. Nonlinear partial differential equations (NPDEs) have a significant role in science and engineering [27, 26]. In the last few decades, many researchers have solved numerous numerical approximations for solving NPDEs. The radial basis function partition of the unity method was used by Nikan and Avazzadeh [22] to solve the sine-Gordon system. Başhan et al. [5] proposed a numerical solution of the complex modified Korteweg–de Vries equation by the differential quadrature method (DQM). Also, Başhan et al. [6] proposed a finite difference method combined with DQM for numerical computation of the modified equal-width wave equation. Many numerical methods, such as finite difference, decomposition, homotopy perturbation, finite element, radial basis function, finite volume, and B-spline collocation methods, have been developed in recent years to solve these NPDEs [23, 29, 33, 30, 6]. Nonlinear evolution equations are employed in many engineering and mathematics disciplines, including solid state physics, fluid mechanics, chemical

physics, plasma physics, geochemistry, and chemical kinematics[14, 4]. One of the well-known nonlinear evolution equations is the fifth-order Kawahara equation, which appears in the theories of shallow water waves having surface tension, magneto-acoustic waves in a plasma, and capillary gravity equations [8, 31].

Consider the nonlinear Kawahara equation [16] is of the form

$$u_t + uu_s + u_{3s} + u_{5s} = 0, \qquad (s,t) \in [0,1] \times [0,T], \tag{1}$$

based on the boundary conditions

$$u(0,t) = -\frac{72}{169} + \frac{105}{169}\operatorname{sech}^4 kct, \tag{2}$$

$$u(1,t) = -\frac{72}{169} + \frac{105}{169}\operatorname{sech}^4 k(1+ct), \tag{3}$$

as well as the initial condition

$$u(s,0) = -\frac{72}{169} + \frac{105}{169}\operatorname{sech}^4 ks, \tag{4}$$

where $k = \frac{1}{2\sqrt{13}}$ and $c = \frac{36}{169}$.

Researchers have developed a wide range of approximate algorithms and analytical techniques over the last few decades to solve the Kawahara equation. In scholarly literature, there are numerous effective methods for approximating the solutions of nonlinear PDE. The homotopy analysis method was used by Abbasbandy [1] to solve the Kawahara equation. Turgut and Karakoc [2] proposed a numerical approach based on the collocation method to solve the modified Kawahara problem. To address the Kawahara problem numerically, Rasoulizadeh and Rashidinia [28] developed a local radial basis function-finite dimension (RBF-FD) meshless approach. Haq and Uddin [13] looked at the RBFs approximation method for the Kawahara problem. Soltanalizadeh [32] used the differential transformation approach to solve the Kawahara equation numerically. Lahiji and Aziz [21] solved the Kawahara equation numerically using spectral methods. Karakoç, Zeybek, and Ak [15] to compute numerical solutions of the Kawahara equation using the septic B-spline collocation method. Zara et al. [34] applied the kernel smoothing method to approximate the modified Kawahara equation numerically.

Several scholars have also developed different approaches for solving this equation.

In this article, we describe a newly developed, extremely simple numerical technique based on Dickson approximation. There are some studies, such as [20, 19, 18], that employed Dickson polynomials as a basis for numerically solving integral equations, but the present method is different from them. This study seeks to develop a newly combined approximation method that combines Dickson polynomial approximation with Taylor-series approximation to solve the nonlinear Kawahara equation. We discretize the time variables after first estimating the time-dependent term with Taylor's series. The underlying model problem is then assumed to have a solution for the novel Dickson series expansion of the unknown function at each time step. Then, using an appropriate set of collocation points, all associated unknown functions are represented in the Dickson matrix form. Error analysis of the combined technique is also presented here. There are few works on solving polynomial approximation since it is difficult to solve the NPDEs. So we combine Taylor's scheme for time derivative and Dickson polynomial for space derivative to create a new technique to solve the Kawahara equation. The obtained results have been compared with the same obtained by some existing methods. It manifests that the present method is more accurate than other methods and the computational time taken by the present technique is much less (less than a second).

The rest of the paper is organized as follows: Section "Introduction" highlights the importance of the Kawahara equation and its application to a certain field of study. It also mentions the main objectives or goals of the research, as well as any background information required for understanding the next parts. The "Preliminaries" Section emphasizes Taylor's time discretization scheme and the Dickson polynomials. The next part discusses the "Error analysis " to quantify how closely the numerical solution approximates the exact solution. The section "Numerical methods for the Kawahara equation" details the suggested numerical technique for approximating solutions to the Kawahara equation. It also explains the step-by-step procedure of the numerical method, possibly making use of the earlier discussed Dickson polynomials and Taylor's time discretization scheme. The proposed method

is used to solve nonlinear Kawahara equations problems about their accuracy and applicability in Section "Illustrative Examples." The contributions of the work are summarized in the "Conclusion" section, which highlights the achievements of the proposed method and identifies possible directions for further study.

2 Preliminaries

Taylor's time discretization scheme:

First, the NPDE is attempted to be discretized with respect to the time variable; that is, the interval $[0, T]$ can be divided into $(M + 1)$ grid points as

$$0 =: t_0 < t_1 = \Delta t < \cdots < t_m = M\Delta t = T, \tag{5}$$

with the uniform step size $\Delta t = t_n - t_{n-1}$. To obtain a discretization scheme that is time accurate, we get the following from the Taylor series representation for $u^n = u(s, t_n)$:

$$u_t^n = \frac{u^{n+1} - u^n}{\Delta t} - \frac{\Delta t u_{tt}^n}{2} + O(\Delta t^2). \tag{6}$$

Dickson polynomial:

The Dickson polynomial was first introduced by Dickson [11] in 1896; afterward, it was rediscovered by Brewer[10] in 1961. For integer $n > 0$ and α is a commutative ring $\mathbb{R}$ with identity. The first kind Dickson polynomial $D_n(s, \alpha)$ of degree n with the parameter $\alpha \in \mathbb{R}$ is defined by

$$D_n(s, \alpha) = \sum_{p=0}^{\lfloor \frac{n}{2} \rfloor} \frac{n}{n-p} (-\alpha)^p s^{n-2p}, \qquad n \geq 1, -\infty < t < \infty.$$

The first few Dickson polynomials are

1. $D_0(s,\alpha) = 2$,
2. $D_1(s,\alpha) = s$,
3. $D_2(s,\alpha) = s^2 - 2\alpha$,
4. $D_3(s,\alpha) = s^3 - 3s\alpha$,
5. $D_4(s,\alpha) = s^4 - 4s^2\alpha$,
6. $D_5(s,\alpha) = s^5 - 5s^3\alpha + 5s\alpha^2$.

They may also be generated by the recurrence relation for $n \geq 2$ as

$D_n(s,\alpha) = sD_{n-1}(s,\alpha) - \alpha D_{n-2}(s,\alpha)$, $n \geq 2$. The polynomials $u = D_n(s,\alpha)$ read the second-order differential equation:

$$(s^2 - 4\alpha)u'' + su' - n^2u = 0, \qquad n = 0, 1, 2, 3, \ldots.$$

Function approximation:

Since $f(s) \in L^2(\mathbb{R})$ is a continuous function defined over $[0,1]$, then the Dickson polynomials can be represented as

$$f(s) = \sum_{i=0}^{N} c_i D_i(s), \qquad s \in [0,1], \tag{7}$$

where $D_n(s,\alpha) = D_n(s)$ with any arbitrary constant α.

In this case, it is necessary to figure out the unknown coefficients, c_i, $i = 0, 1, \ldots, N$. The Dickson vector $D_N(s)$ and the unknown vector C_N will be introduced in the following ways:

$$D_N(s) = \Big[D_0(s)\ D_1(s)\ \ldots\ D_n(s)\Big]^T,$$

and

$$C_N = \Big[c_0\ c_1\ \ldots\ c_n\Big]^T.$$

We can express (7) in a compact representation using these vectors as follows:

$$f(s) = C_N^T D_N(s). \tag{8}$$

Additionally, by introducing the matrix Ψ,

$$\Psi = \begin{bmatrix} 2 & 0 & 0 & \dots 0 \\ 0 & \frac{1}{1}\binom{1}{0}(-\alpha)^0 & 0 & \dots 0 \\ \frac{2}{1}\binom{2}{1}(-\alpha)^1 & 0 & \frac{2}{2}\binom{2}{0}(-\alpha)^0 & \dots 0 \\ 0 & \frac{3}{2}\binom{2}{1}(-\alpha)^1 & 0 & \dots 0 \\ \vdots & \vdots & \vdots & \vdots \ 0 \\ \frac{N}{\frac{N}{2}}\binom{\frac{N}{2}}{\frac{N}{2}}(-\alpha)^{\frac{N}{2}} & 0 & \frac{N}{\frac{N}{2}+1}\binom{\frac{N}{2}+1}{\frac{N}{2}-1}(-\alpha)^{\frac{N}{2}-1} & \dots 0 \end{bmatrix}_{(N+1)\times(N+1)},$$

and the monomial vector

$$S_n(s) = \begin{bmatrix} 1 & s & s^2 & \dots & s^n \end{bmatrix}^T,$$

we shall express $D_N(s)$ as follows:

$$D_N(s) = \Psi S_n(s). \tag{9}$$

For finding the operational matrix of the derivative, a straightforward calculation shows that

$$\begin{aligned} \frac{d}{ds}D_N(s) &= \psi \frac{d}{ds}S_n(s) \\ &= \psi \lambda S_n(s), \end{aligned} \tag{10}$$

where

$$\lambda = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & 0 \\ 0 & 0 & \dots & 0 & N & 0 \\ 0 & 0 & \vdots & 0 & 0 & 0 \end{bmatrix}_{(N+1)\times(N+1)}. \tag{11}$$

3 Error analysis:

Kürkçü, Aslan, and Sezer [19] employed the residual function in the Banach space method to do convergence research on the Dickson polynomial solution for integro-differential equations. We can find the Dickson polynomial-based

error analysis for approximation for the nonlinear Kawahara equation. Finally, using the Dickson and Taylor's approximation, we derive the error analysis of approximation.

Theorem 1. [17] Suppose that H is a Hilbert space and that Y is a closed subspace of H such that $dim\ Y < \infty$ and $y_1, y_2, \ldots, y_n$ is any basis for Y. Let x be an arbitrary element in H and let y_0 be the most accurate approximation of $x \in Y$. Then $\|x - y_0\|_2^2 = \frac{G(x,y_1,y_2,\ldots,y_n)}{G(y_1,y_2,\ldots,y_n)}$, where

$$G(x, y_1, y_2, \ldots, y_n) = \begin{vmatrix} \langle x, x\rangle & \langle x, y_1\rangle & \langle x, y_2\rangle & \cdots & \langle x, y_n\rangle \\ \langle y_1, x\rangle & \langle y_1, y_1\rangle & \langle y_1, y_2\rangle & \cdots & \langle y_1, y_n\rangle \\ & & & & \\ \langle y_n, x\rangle & \langle y_n, y_1\rangle & \langle y_n, y_2\rangle & \cdots & \langle y_n, y_n\rangle \end{vmatrix}.$$

Theorem 2. Let $u \in C([0,1]\times[0,1])$ and let $Y = Span\{D_0, D_1, D_2, \ldots, D_n\}$, where D_i's are Dickson polynomials. If $u_N = c^T\phi(x)$ is the most accurate approximation to u for particular time step Δt, then the error bound is as follows:

$$\|u - u_N\|_2 \leq O(\Delta t)^2 + \left(\frac{G(u, D_0, D_1, \ldots, D_n)}{G(D_0, D_1, \ldots, D_n)}\right)^{\frac{1}{2}},$$

Proof. Let $u(x,t)$ be the exact solution and let $u(x,t_n)$ be an approximate solution after time discretization. For a particular time step Δt, from (6), we have

$$\|u(x,t) - u(x,t_n)\|_2 \leq O(\Delta t)^2.$$

Let $u_N(x,t_n) = c^T\phi(x)$ be the most accurate approximate solution of $u(x,t_n)$ using Dickson polynomials for a particular time step $t = t_n$. The error bound is as follows:

$$\begin{aligned} \|u(x,t) - u_N(x,t_n)\|_2 &= \|u(x,t) - u(x,t_n) + u(x,t_n) - u_N(x,t_n)\|_2 \\ &\leq \|u(x,t) - u(x,t_n)\|_2 + \|u(x,t_n) - u_N(x,t_n)\|_2 \\ &\leq O(\Delta t)^2 + \left(\frac{G(u, D_0, D_1, \ldots, D_n)}{G(D_0, D_1, \ldots, D_n)}\right)^{\frac{1}{2}}. \end{aligned}$$

□

4 Numerical method for Kawahara equation

Consider the nonlinear Kawahara equation defined in (1)–(4) and the differentiate equation (1) with respect to t, we get

$$u_{tt}^n = (u_{5s})_t - (u_{3s})_t - (uu_s)_t. \tag{12}$$

By replacing first-order derivatives $u_t^n \approx \frac{(u^{n+1}-u^n)}{\Delta t}$ in all occurrences, we may write the equation as

$$u_{tt}^n = \Big(\frac{u^{n+1}-u^n}{\Delta t}\Big)_{5s} - \Big(\frac{u^{n+1}-u^n}{\Delta t}\Big)_{3s} - \Big(\frac{u^{n+1}-u^n}{\Delta t}\Big)u_s - u\Big(\frac{u_s^{n+1}-u_s^n}{\Delta t}\Big),$$

$$\Delta t u_{tt}^n = (u_{5s}^{n+1}-u_{5s}^n) - (u_{3s}^{n+1}-u_{3s}^n) - (u^{n+1}-u^n)u_s - u(u_s^{n+1}-u_s^n). \tag{13}$$

Next, using the time discretized form, (13) is inserted into the right-hand side of (6) to get

$$\begin{aligned} u_t^n = \Big(\frac{u^{n+1}-u^n}{\Delta t}\Big) - \frac{1}{2}\Big[&(u_{5s}^{n+1}-u_{5s}^n) - (u_{3s}^{n+1}-u_{3s}^n) \\ &-(u^{n+1}-u^n)u_s - u(u_s^{n+1}-u_s^n)\Big]. \end{aligned} \tag{14}$$

Now, discretizing the time-dependent times of (1), we get

$$u_t^n = u_{5s}^n - u_{3s}^n - u^n u_s^n. \tag{15}$$

Again, replace (15)in the left side of (14); that is,

$$\begin{aligned} u_{5s}^n - u_{3s}^n - u^n u_s^n = &\Big(\frac{u^{n+1}-u^n}{\Delta t}\Big) - \frac{1}{2}\Big[\Delta t u_{tt}^n = (u_{5s}^{n+1}-u_{5s}^n) \\ &- (u_{3s}^{n+1}-u_{3s}^n) - (u^{n+1}-u^n)u_s - u(u_s^{n+1}-u_s^n)\Big]. \end{aligned} \tag{16}$$

Equation (16) can be simplified as

$$\begin{aligned} \Big[\frac{1}{\Delta t} + \frac{1}{2}u_s^n\Big]u^{n+1} + \Big[\frac{1}{2}u^n\Big]u_s^{n+1} + \Big[\frac{1}{2}\Big]u_{3s}^{n+1} + \Big[\frac{-1}{2}\Big]u_{5s}^{n+1} \\ = \frac{1}{2}u_{5s}^n - \frac{1}{2}u_{3s}^n + \frac{u^n}{\Delta t}. \end{aligned} \tag{17}$$

The unknown functions $u^{n+1}, u_s^{n+1}, u_{3s}^{n+1}$, and u_{5s}^{n+1} can be approximated by Dickson polynomials defined in (17) as

$$u^{n+1}(s) = C_N^T D(s) = C_N^T \Psi S_n(s). \tag{18}$$

Similarly,

$$\begin{aligned} u_s^{n+1} &= C_N^T \Psi \lambda S_n(s), \\ u_{2s}^{n+1} &= C_N^T \Psi \lambda^2 S_n(s), \\ u_{3s}^{n+1} &= C_N^T \Psi \lambda^3 S_n(s), \\ u_{5s}^{n+1} &= C_N^T \Psi \lambda^5 S_n(s). \end{aligned}$$

Now, (17) can be described as follows:

$$\begin{aligned} &A_3(s) C_N^T \Psi \lambda^5 S_n(s) + A_2(s) C_N^T \Psi \lambda^3 S_n(s) \\ &+ A_1(s) C_N^T \Psi \lambda S_n(s) + A_0(s) C_N^T \Psi S_n(s) = F_n(s), \end{aligned} \tag{19}$$

where

$$\begin{aligned} A_0(s) &= \frac{1}{\Delta t} + \frac{1}{2} u_s^n, \qquad A_1(s) = \frac{1}{2} u^n, \qquad A_2(s) = \frac{1}{2}, \\ A_3(s) &= \frac{-1}{2}, \qquad F_n(s) = \frac{1}{2} u_{5s}^n - \frac{1}{2} u_{3s}^n + \frac{u^n}{\Delta t}, \end{aligned}$$

and (19) can be simplified as

$$C_N^T \Psi [A_3(s)\lambda^5 + A_2(s)\lambda^3 + A_1(s)\lambda + A_0(s)] S_n(s) = F_n(s). \tag{20}$$

Putting the collocation points $s_i = (\frac{i}{n})$ in (20), we get a system of the linear equation as follows:

$$C_N^T \Psi [A_3(s_i)\lambda^5 + A_2(s_i)\lambda^3 + A_1(s_i)\lambda + A_0(s_i)] S_n(s_i) = F_n(s_i) \tag{21}$$

for $i = 0, 1, 2, \ldots, N$. Equation (21) is clearly a set of $(N+1)$ linear equations in terms of $(N+1)$ unknown coefficients $c_0, c_1, c_2, \ldots, c_N$ that need to be found.

Again, (21) can be expressed as a matrix equation as

$$D_N C_N = F_n, \tag{22}$$

where

$$D_n := (\Phi A_3(\lambda^T)^5 + \Phi A_2(\lambda^T)^3 + \Phi A_1(\lambda^T) + \Phi A_0)\Psi^T$$

and

$$\Phi = \left[S_n(s_0)\ S_n(s_1)\ S_n(s_2)\ \ldots\ S_n(s_n)\right],$$

$$A_m = \begin{bmatrix} A_m(s_0) & 0 & \ldots & 0 \\ 0 & A_m(s_1) & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & A_m(s_N) \end{bmatrix}_{(N+1)\times(N+1)},$$

$$F_n = \begin{bmatrix} f_n(s_0) \\ f_n(s_1) \\ \vdots \\ f_n(s_N) \end{bmatrix}_{(N+1)\times 1}.$$

From the given boundary condition, we have $u_{n+1}(0) = h_0^{n+1} = -\frac{72}{169} + \frac{105}{169}\operatorname{sech}^4 kct$ and $u_{n+1}(1) = h_1^{n+1} = -\frac{72}{169} + \frac{105}{169}\operatorname{sech}^4 k(1+ct)$, using the matrix notation, it may be represented as

$D_n(0)C_N = h_0^{n+1}$, $D_n(0) = S_n(s_0)\Psi^T$,

$D_n(1)C_N = h_1^{n+1}$, $D_n(1) = S_n(s_1)\Psi^T$.

Next, we substitute the first two rows of (22) by the boundary conditions $\left[D_n(0); h_0^{n+1}\right]$ and $\left[D_n(1); h_1^{n+1}\right]$ for the convenience. From the above system, we can obtain the approximate value of the unknown coefficients $c_0, c_1, c_2, \ldots, c_N$ and hence obtain the solution of (18).

5 Illustrative examples

To demonstrate the performance of the Combined Taylor–Dickson Approximation (CTDA) technique, numerical simulations based on the nonlinear Kawahara equation for specific initial and boundary values are provided. We utilize an ASUS laptop with Processor (CPU) Ryzen 5 (3500U) and RAM(8GB) processing capabilities for conducting the computation. The MATHEMATICA software is used for the problem implementation and 3D

plotting of exact solution and approximate solution graphs with various values of s, t, and Δt.

Example 1. Consider the nonlinear Kawahara equation, (1)–(4). The exact solution of this problem is given byKaya and Al-Khaled [16] as $u(s,t) = -\frac{72}{169} + \frac{105}{169}\operatorname{sech}^4 k(s+ct)$, where $k = \frac{1}{2\sqrt{13}}$ and $c = \frac{36}{169}$. Using time steps $\Delta t = 0.1$, $\Delta t = 0.01$, $\Delta t = 0.001$, and $N = 6$, the absolute errors obtained by the present method for $0 \le s \le L = 1$ is given in Table 1. Despite the fact that there are few approximation methods for solving the Kawahara problem, the authors have compared their findings to the exact answer that has been verified.

The maximum absolute errors solved by the RBF method [13] with the nonlinear Kawahara equation are compared in Table 2. Table 3 displays the absolute error for Example 1 determined using the CTDA technique for various values of the Dickson parameter $\alpha = 1, 2$, and the solution is not affected for different values of α. Figures 1 and 2 show the behavior of the exact and approximation solutions for $N = 6$ and $\Delta t = 0.1$, respectively. Figure 3 shows a two-dimensional graphical representation of the exact and approximate solutions for $t = 1$ and $\Delta t = 0.1$. The graph represents how these possibilities compare visually. Figure 4 displays an "Absolute error graph" that shows the differences between the exact and approximate solutions for $N = 6$ and $\Delta t = 0.1$. In the literature, the computation time of the RBF method has not been provided. However, the proposed method provides more accurate results than the RBF method, and the computational time of the present method is much less than 0.267 seconds for $N = 6$.

From Theorem 2, the maximum absolute error for $N = 6$ and $\Delta t = 0.1$ is computed as

$$O(\Delta t)^2 + \left(\frac{G(u, D_0, D_1, \ldots, D_n)}{G(D_0, D_1, \ldots, D_n)}\right)^{\frac{1}{2}} = (0.1)^2 + 2.01111 \times 10^{-7}$$
$$= 1.00002 \times 10^{-2}.$$

Similarly, for $\Delta t = 0.01$ and $\Delta t = 0.001$, the maximum error can be calculated as 1.00201×10^{-4} and 1.20111×10^{-6}, respectively. So, this present method satisfies the error-bound condition for different time steps.

Table 1: Absolute error for Example 1 for $N = 6$.

t	s	$\Delta t = 0.1$	$\Delta t = 0.01$	$\Delta t = 0.001$
0.2	0.2	2.02289×10^{-6}	2.35173×10^{-8}	2.66443×10^{-11}
	0.4	2.2706×10^{-6}	1.01181×10^{-7}	2.66443×10^{-10}
	0.6	1.33716×10^{-6}	2.24034×10^{-8}	2.38497×10^{-10}
	0.8	9.9872×10^{-6}	4.80536×10^{-8}	2.66443×10^{-10}
0.4	0.2	3.82575×10^{-6}	4.60788×10^{-7}	3.20035×10^{-10}
	0.4	5.57816×10^{-6}	4.54266×10^{-7}	2.24848×10^{-9}
	0.6	1.9494×10^{-6}	1.72914×10^{-8}	1.91345×10^{-9}
	0.8	8.80955×10^{-6}	4.29541×10^{-8}	1.91345×10^{-9}
0.6	0.2	4.51829×10^{-6}	5.54659×10^{-7}	1.83046×10^{-9}
	0.4	7.4988×10^{-6}	7.01482×10^{-7}	1.6928×10^{-9}
	0.6	5.48562×10^{-6}	7.81906×10^{-7}	1.29511×10^{-9}
	0.8	3.23781×10^{-6}	2.91407×10^{-8}	1.91345×10^{-9}
0.8	0.2	3.41209×10^{-6}	4.23962×10^{-8}	8.60722×10^{-9}
	0.4	6.12709×10^{-6}	6.07458×10^{-7}	1.10209×10^{-10}
	0.6	5.48562×10^{-6}	1.03009×10^{-7}	8.43569×10^{-10}
	0.8	1.20464×10^{-6}	1.47048×10^{-8}	1.91345×10^{-9}

Table 2: Comparison of L_∞ errors of Example 1 for $\Delta t = 0.1$ and $N = 6, 8, 10$.

Present method($N = 6$)	9.9872×10^{-4}
Present method($N = 8$)	2.64883×10^{-6}
Present method($N = 10$)	2.223943×10^{-7}
RBF[13] method($N = 51$)	9.38×10^{-6}

Table 3: Absolute error of Example 1 for Dickson parameter $\alpha = 1, 2$.

t	**x**	$\alpha = 1$		$\alpha = 2$	
		$\Delta t = 0.1$	$\Delta t = 0.01$	$\Delta t = 0.1$	$\Delta t = 0.01$
0.1	0.1	3.967×10^{-6}	7.253×10^{-8}	3.967×10^{-6}	7.253×10^{-8}
	0.2	9.746×10^{-6}	1.098×10^{-8}	9.746×10^{-6}	1.098×10^{-8}
	0.3	1.114×10^{-6}	9.507×10^{-8}	1.114×10^{-6}	9.507×10^{-8}
0.2	0.1	7.927×10^{-6}	1.463×10^{-8}	7.927×10^{-6}	1.463×10^{-8}
	0.2	2.022×10^{-6}	2.351×10^{-8}	2.022×10^{-6}	2.351×10^{-8}
	0.3	2.559×10^{-6}	2.382×10^{-8}	2.559×10^{-6}	2.382×10^{-8}
0.3	0.1	1.114×10^{-6}	2.134×10^{-8}	1.114×10^{-6}	2.134×10^{-8}
	0.2	3.012×10^{-6}	3.575×10^{-8}	3.012×10^{-6}	3.575×10^{-8}
	0.3	4.006×10^{-6}	3.966×10^{-8}	4.006×10^{-6}	3.966×10^{-8}

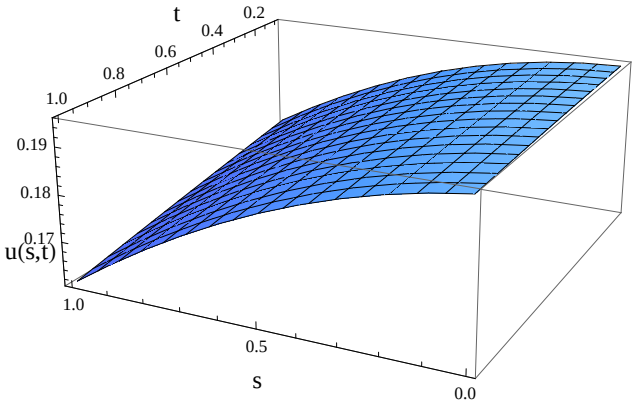


Figure 1: Exact solution of Example 1 for $\Delta t = 0.1$ and $N = 6$.

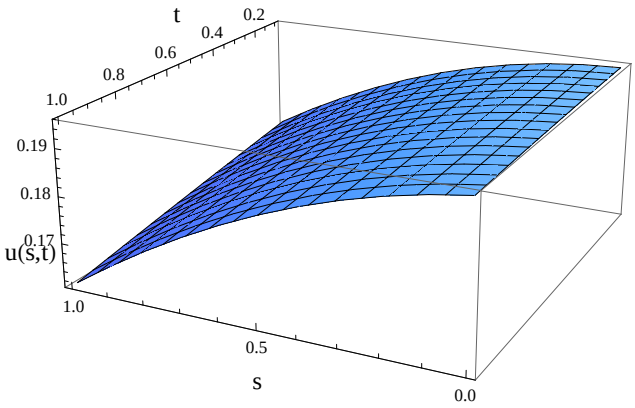


Figure 2: Approximate solution of Example 1 for $\Delta t = 0.1$ and $N = 6$.

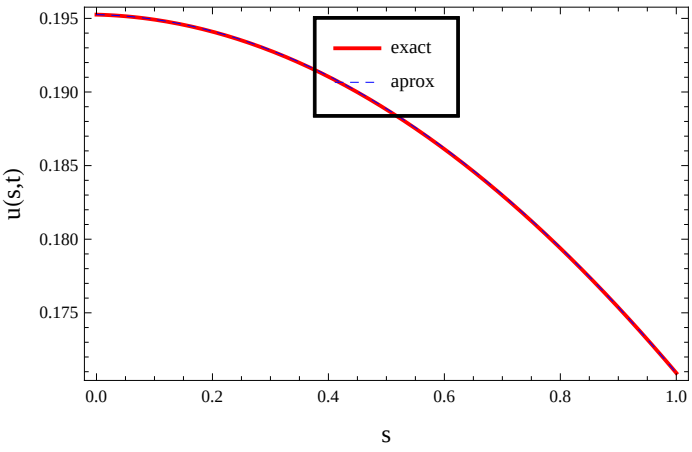


Figure 3: Two-dimensional graph of Example 1 for $\Delta t = 0.1$ and $N = 6$ and $t = 1$.

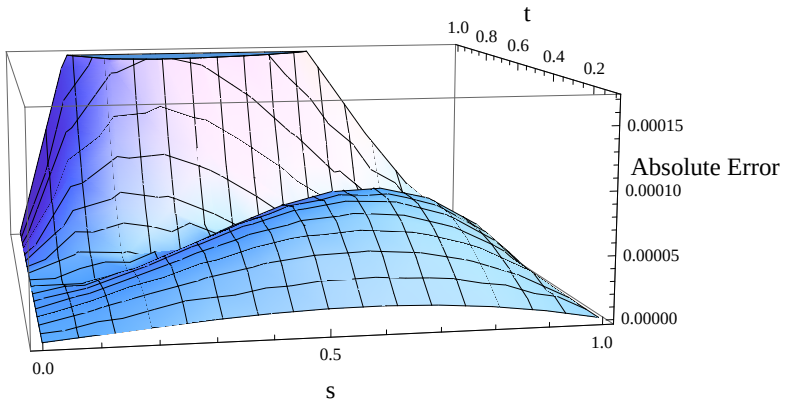


Figure 4: Absolute error of Example 1 for $\Delta t = 0.1$ and $N = 6$.

Example 2. Here, we consider a Kawahara-type equation given as

$u_t + uu_s + u_{3s} + u_s - u_{5s} = 0.$

Whose exact solution is given as

$u(s,t) = -\frac{105}{169}\operatorname{sech}^4 k(s - 2 - ct)$, where $k = \frac{1}{2\sqrt{13}}$ and $c = \frac{205}{169}$. Using time steps $\Delta t = 0.1$, $\Delta t = 0.01$, $\Delta t = 0.001$ and the Kawahara equation with $N = 8$, the approximation for $0 \leq s \leq L = 1$ is given in Table 4. Despite the fact that there are few approximation methods for solving the Kawahara problem, the authors compare their findings to the exact answer that has been verified.

The maximum absolute errors solved by the RBF method [13] with the nonlinear Kawahara equation are compared in Table 5. Table 6 displays the absolute error for Example 2 determined using the CTDA technique for various values of the Dickson parameter $\alpha = 1, 2$, and the solution is not affected for different values of α. Figures 5 and 6 show the behavior of the exact and approximation solutions for $N = 8$ and $\Delta t = 0.1$, respectively. Figure 7 shows a two-dimensional graphical representation of the exact and approximate solutions for $t = 1$ and $\Delta t = 0.1$. The graph represents how these possibilities compare visually. Figure 8 displays an "Absolute error graph" that shows the differences between the exact and approximate solutions for $N = 6$, $\Delta t = 0.1$. In the literature, the computation time of the RBF method has not been provided. However, the proposed method provides more accu-

Table 4: Absolute error for Example 2 for $N = 8$.

t	s	$\Delta t = 0.1$	$\Delta t = 0.01$	$\Delta t = 0.001$
0.2	0.2	1.894×10^{-6}	1.73465×10^{-8}	3.13242×10^{-11}
	0.4	9.10089×10^{-6}	1.21102×10^{-7}	1.23812×10^{-10}
	0.6	1.71853×10^{-6}	2.04784×10^{-8}	2.06468×10^{-10}
	0.8	2.99047×10^{-6}	2.01218×10^{-8}	2.01002×10^{-10}
0.4	0.2	2.64883×10^{-6}	4.13838×10^{-7}	4.36423×10^{-10}
	0.4	1.32887×10^{-6}	1.77252×10^{-7}	1.81323×10^{-9}
	0.6	2.59029×10^{-6}	3.0947×10^{-8}	3.12158×10^{-9}
	0.8	2.89771×10^{-6}	3.16247×10^{-8}	3.161×10^{-9}
0.6	0.2	2.46179×10^{-6}	3.81904×10^{-7}	4.03194×10^{-9}
	0.4	1.28578×10^{-6}	1.71872×10^{-7}	1.75922×10^{-9}
	0.6	2.5832×10^{-6}	3.09405×10^{-6}	3.12232×10^{-9}
	0.8	2.99047×10^{-6}	3.27557×10^{-8}	3.27585×10^{-9}
0.8	0.2	1.5832×10^{-6}	2.32987×10^{-6}	2.46238×10^{-9}
	0.4	8.23146×10^{-6}	1.110252×10^{-7}	1.112919×10^{-10}
	0.6	1.70397×10^{-6}	2.04625×10^{-7}	2.0659×10^{-10}
	0.8	2.03731×10^{-6}	2.223943×10^{-7}	2.224081×10^{-10}

rate results than the RBF method and the computational time of the present method is much less that is 0.326 seconds for $N = 8$.

6 Conclusion

The Kawahara equation was solved using a combination of Taylor's and Dickson polynomial approximation. The authors employed derivative operational matrices as a tool to implement their chosen approximation methods. These matrices help in the numerical computation of function derivatives, which is necessary when dealing with differential equations. For small values of N, the current method yields very precise results with exact solutions. We get a more accurate solution if we increase the value of N. Also, taking small values of Δt yields a more accurate result. The validity and convenience of the existing techniques were demonstrated using numerical examples. These ex-

Table 5: Comparison of L_∞ errors of Example 2 for $\Delta t = 0.1$ for $N = 6, 8, 10$.

Present method($N = 6$)	1.3288×10^{-4}
Present method($N = 8$)	1.28578×10^{-6}
Present method($N = 10$)	4.36423×10^{-8}
RBF[13] method($N = 51$)	6.159×10^{-5}

Table 6: Absolute error of Example 2 for Dickson parameter $\alpha = 1, 2$.

t	**x**	$\alpha = 1$		$\alpha = 2$	
		$\Delta t = 0.1$	$\Delta t = 0.01$	$\Delta t = 0.1$	$\Delta t = 0.01$
0.1	0.1	1.118×10^{-6}	2.285×10^{-8}	1.118×10^{-6}	2.285×10^{-8}
	0.2	3.220×10^{-6}	4.155×10^{-8}	3.220×10^{-6}	4.155×10^{-8}
	0.3	5.039×10^{-6}	4.523×10^{-8}	5.039×10^{-6}	4.523×10^{-8}
0.2	0.1	2.138×10^{-6}	4.117×10^{-8}	2.138×10^{-6}	4.117×10^{-8}
	0.2	5.860×10^{-6}	7.590×10^{-8}	5.860×10^{-6}	7.590×10^{-8}
	0.3	8.355×10^{-6}	9.451×10^{-8}	8.355×10^{-6}	9.451×10^{-8}
0.3	0.1	2.835×10^{-6}	5.463×10^{-8}	2.835×10^{-6}	5.463×10^{-8}
	0.2	7.847×10^{-6}	1.019×10^{-8}	7.847×10^{-6}	1.019×10^{-8}
	0.3	1.133×10^{-6}	1.296×10^{-8}	1.133×10^{-6}	1.296×10^{-8}

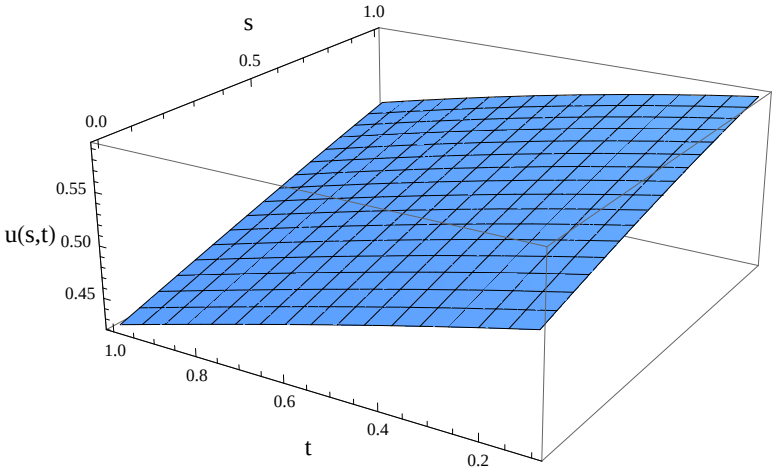


Figure 5: Exact solution of Example 2 for $\Delta t = 0.1$ and $N = 8$.

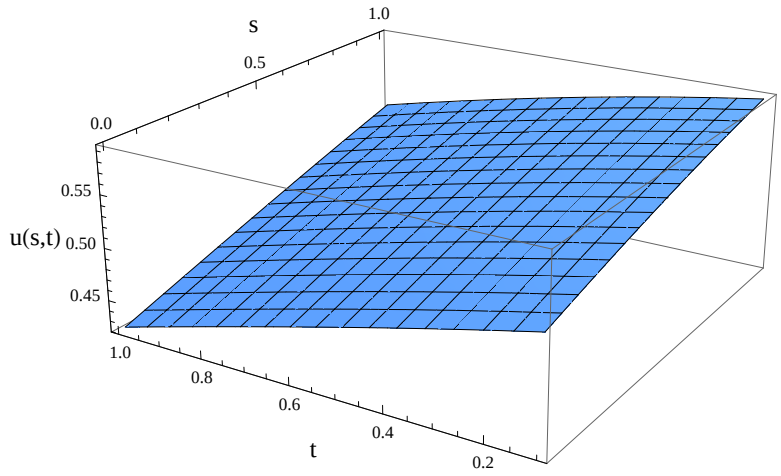


Figure 6: Approximate solution of Example 2 for $\Delta t = 0.1$ and $N = 8$.

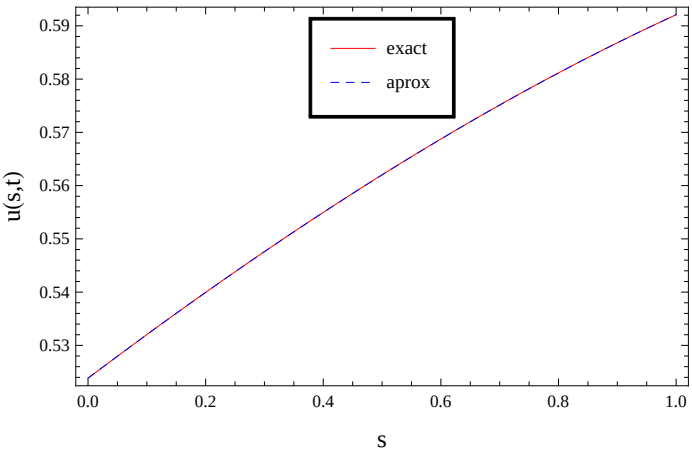


Figure 7: Two-dimensional graph of Example 2 for $\Delta t = 0.1$, $N = 8$, and $t = 1$.

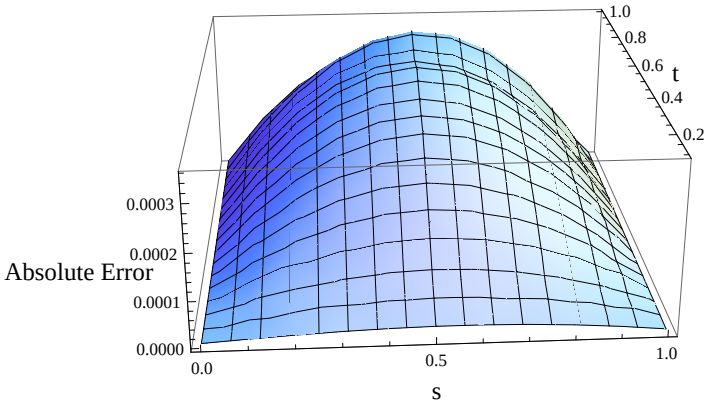


Figure 8: Absolute error of Example 2 for $\Delta t = 0.1$ and $N = 6$.

amples also proved the accuracy and efficiency of the proposed method. This method established a balance between accuracy and computational efficiency, which is important when dealing with challenging mathematical problems. This method can also be used to solve various NPDEs derived from physical models, which are subject to specified initial and boundary conditions. The method's future direction is that it can be used for NPDE with higher-order time derivatives.

Acknowledgements

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

References

[1] Abbasbandy, S. *Homotopy analysis method for the Kawahara equation,* Nonlinear Anal. Real World Appl. 11(1) (2010), 307–312.

[2] Ak, T. and Karakoc, S.B.G. *A numerical technique based on collocation*

method for solving modified Kawahara equation, J. Ocean Eng. Sci. 3(1) (2018), 67–75.

[3] Alipour, M. and Soradi Zeid, S. *Optimal control of Volterra integro-differential equations based on interpolation polynomials and collocation method,* Comput. Methods Differ. Equ. 11(1) (2023), 52–64.

[4] Aljahdaly, N.H. and Alweldi, A.M. *On the modified Laplace homotopy perturbation method for solving damped modified kawahara equation and its application in a fluid,* Symmetry, 15(2) (2023), 394.

[5] Başhan, A., Uçar, Y., Yağmurlu, N.M. and Esen, A., 2016, October. *Numerical solution of the complex modified Korteweg-de Vries equation by DQM,* In Journal of Physics: Conference Series (Vol. 766, No. 1, p. 012028). IOP Publishing.

[6] Başhan, A., Yağmurlu, N.M., Ucar, Y. and Esen, A., 2021. *Finite difference method combined with differential quadrature method for numerical computation of the modified equal width wave equation,* Numer. Methods Partial Differential Equations 37(1) (2021), 690–706.

[7] Bhargava, M. and Zieve, M.E. *Factoring Dickson polynomials over finite fields*, Finite Fields and Their Appl. 5(2) (1999), 103–111.

[8] Bhatter, S., Mathur, A., Kumar, D., Nisar, K.S. and Singh, J. *Fractional modified Kawahara equation with Mittag–Leffler law,* Chaos, Solitons Fractals, 131 (2020), 109508.

[9] Biswas, A. *Solitary wave solution for the generalized Kawahara equation,* Appl. Math. Lett. 22(2) (2009), 208–210.

[10] Brewer, B.W., *On certain character sums,* Trans. Amer. Math. Soc. 99(2) (1961), 241–245.

[11] Dickson, L.E., *The analytic representation of substitutions on a power of a prime number of letters with a discussion of the linear group,* Ann. of Math. 11(1-6) (1896/97), 65–120.

[12] Fernando, N., *A study of permutation polynomials over finite fields*, Thesis (Ph.D.)–University of South Florida. ProQuest LLC, Ann Arbor, MI, 2013. 109 pp.

[13] Haq, S. and Uddin, M. *RBFs approximation method for Kawahara equation,* Eng. Anal. Bound. Elem. 35(3) (2011), 575–580.

[14] He, D., *New solitary solutions and a conservative numerical method for the Rosenau–Kawahara equation with power law nonlinearity,* Nonlinear Dynam. 82 (2015), 1177–1190.

[15] Karakoç, S.B.G., Zeybek, H. and Ak, T. *Numerical solutions of the Kawahara equation by the septic B spline collocation method,* Stat. Optim. Inf. Comput. 2(3) (2014), 211–221.

[16] Kaya, D. and Al-Khaled, K. *A numerical comparison of a Kawahara equation,* Phys. Lett. A, 363 (5-6) (2007), 433–439.

[17] Kreyszig, E. *Introductory functional analysis with applications,* John Wiley & Sons, New York-London-Sydney, 1978.

[18] Kürkçü, Ö.K., Aslan, E. and Sezer, M. *A numerical approach with error estimation to solve general integro-differential-difference equations using Dickson polynomials,* Appl. Math. Comput 276 (2016), 324–339.

[19] Kürkçü, Ö.K., Aslan, E. and Sezer, M. *A novel collocation method based on residual error analysis for solving integro-differential equations using hybrid Dickson and Taylor polynomials,* Sains Malays, 46 (2017), 335–347.

[20] Kürkçü, Ö.K., Aslan, E., Sezer, M. and İlhan, Ö. *A numerical approach technique for solving generalized delay integro-differential equations with functional bounds by means of Dickson polynomials,* Int. J. Comput. Methods, 15(05) (2018), 1850039.

[21] Lahiji, M.A. and Aziz, Z.A. *Numerical solution for Kawahara equation by using spectral methods,* IERI Procedia, 10 (2014), 259–265.

[22] Nikan, O. and Avazzadeh, Z. *A locally stabilized radial basis function partition of unity technique for the sine-Gordon system in nonlinear optics,* Math. Comput. Simul. 199 (2022), 394–413.

[23] Nikan, O., Avazzadeh, Z., Machado, J.T. and Rasoulizadeh, M.N., *An accurate localized meshfree collocation technique for the telegraph equation in propagation of electrical signals,* Eng. Comput. 39(3) (2023), 2327–2344.

[24] Nikan, O., Avazzadeh, Z. and Rasoulizadeh, M.N., *Soliton solutions of the nonlinear sine-Gordon model with Neumann boundary conditions arising in crystal dislocation theory,* Nonlinear Dynam. 106(1) (2021), 783–813.

[25] Nikan, O., Avazzadeh, Z. and Rasoulizadeh, M.N., *Soliton wave solutions of nonlinear mathematical models in elastic rods and bistable surfaces,* Eng. Anal. Bound. Elem. 143 (2022), 14–27.

[26] Özer, S. *Numerical solution by quintic B-spline collocation finite element method of generalized Rosenau–Kawahara equation,* Math. Sci. 16(3) (2022), 213–224.

[27] Polat, N., Kaya, D. and Tutalar, H.I. *A analytic and numerical solution to a modified Kawahara equation and a convergence analysis of the method,* Appl. Math. Comput. 179(2) (2006), 466–472.

[28] Rasoulizadeh, M.N. and Rashidinia, J. *Numerical solution for the Kawahara equation using local RBF-FD meshless method,* J. King Saud Univ. Sci. 32(4) (2020), 2277–2283.

[29] Sahu, P.K., Ray, S.S. *Numerical solutions for Volterra integro-differential forms of Lane-Emden equations of first and second kind using Legendre multiwavelets,* Electron. J. Differ. Equ. 2015 (28) (2015), 1–11.

[30] Sahu, P.K. and Ray, S.S., *Legendre wavelets operational method for the numerical solutions of nonlinear Volterra integro-differential equations system,* Appl. Math. Comput. 256 (2015), 715–723.

[31] Saldır, O., Sakar, M.G. and Erdogan, F., *Numerical solution of time-fractional Kawahara equation using reproducing kernel method with error estimate,* Comput. Appl. Math. 38 (2019), 1–23.

[32] Soltanalizadeh, B., *Application of differential transformation method for numerical analysis of Kawahara equation,* Aust. J. Basic Appl. Sci. 5(12) (2011), 490–495.

[33] Wang, X. and Cheng, H. *Solitary wave solution and a linear mass-conservative difference scheme for the generalized Korteweg–de Vries–Kawahara equation,* Comput. Appl. Math. 40 (2021), 1–26.

[34] Zara, A., Rehman, S.U., Ahmad, F., Kouser, S. and Pervaiz, A. *Numerical approximation of modified Kawahara equation using Kernel smoothing method,* Math. Comput. Simul. 194 (2022), 169–184.

Iranian Journal of Numerical Analysis and Optimization
Vol. 14, No. 1, 2024, pp 44–76
https://doi.org/ 10.22067/ijnao.2023.82236.1251
https://ijnao.um.ac.ir/

CC BY NC

Research Article

A new approach for ranking decision-making units in data envelopment analysis by using communication game theory

M. Amiri, A. Ashrafi*

Abstract

Ranking decision making units (DMUs) is an important topic in data envelopment analysis (DEA). When efficient DMUs or inefficient DMUs have the same efficiency score, the traditional DEA model usually fails to rank all DMUs. For the sake of comparing and improving the discrimination power of DMUs, some proposed approaches use cooperative game theory for ranking DMUs. In this paper, communication game theory, which includes a transferable utility cooperative game and an undirected graph describing limited cooperation between players, can be used to rank DMUs. The idea

*Corresponding author

Received 3 May 2023; revised 30 July 2023; accepted 28 September 2023

Ali Ashrafi
Department of Mathematics, Faculty of Mathematics, Statistics and Computer Sciences, Semnan University, Semnan, Iran. e-mail: a_ashrafi@semnan.com

Mahdie Amiri
Ph.D Student, Faculty of Mathematics, Statistics and Computer Sciences, Semnan University, Semnan, Iran. e-mail: mahdie.amiri@semnan.ac.ir

How to cite this article
Amiri, M. and Ashrafi, A., A new approach for ranking decision-making units in data envelopment analysis by using communication game theory. *Iran. J. Numer. Anal. Optim.*, 2024; 14(1): 44-76. https://doi.org/ 10.22067/ijnao.2023.82236.1251

is that the ranking of DMUs can be done by measuring the effect of removing a subset of DMUs on the total share of the remaining DMUs obtained by the reference frontier share model. In the proposed approach, the players are the DMUs, and the characteristic function measures the increase and decrease in the total share of each DMU. The current paper considers the total share for efficient and inefficient DMUs to rank all DMUs. The proposed approach has been tested on several datasets and compared with the results of the previous ranking methods, which sometimes coincide. In the empirical study, a complete ranking of DMUs is useful and reasonable.

AMS subject classifications (2020): Primary 45D05; Secondary 42C10, 65G99.

Keywords: Data envelopment analysis, Communication game, Myerson value, Reference frontier share model.

1 Introduction

Data envelopment analysis (DEA) is a popular nonparametric approach in operations research and economics to evaluate the relative efficiency of a group of homogeneous decision-making units (DMUs) that consume multiple inputs to produce multiple outputs. Charnes, Cooper, and Rhodes [11] proposed the first DEA model under the assumption of constant returns to scale (CRS) technology. DEA traditional models can define the efficiency score for each DMU as the sum of weighted outputs divided by the sum of weighted inputs. The efficiency score is constrained to lie between zero and one, and if it is equal to one, then it means that the evaluated DMU is efficient, and if it is less than one, then the evaluated DMU is inefficient.

The traditional DEA models evaluate many DMUs as efficient, but they are not able to attain ranking for these units. The lack of discrimination power among the efficient DMUs leads to the major drawbacks of the traditional DEA models. Some scholars have proposed some new approaches for improving the discrimination power of the traditional DEA model, among them, we focus on the reference frontier share model proposed by Rezaeiani and Foroughi [42]. They defined the concept of the reference frontier to rank the efficient DMUs. The reference frontier of an inefficient DMU is the set

of all efficient DMUs that can make better the efficiency of that inefficient DMU. Their model measures the maximum value of the efficient DMU coefficient in all convex combinations of the reference frontier points of inefficient DMUs. The importance of each efficient DMU is specified by these values, which can be defined as the share of each efficient DMU in the creation of the reference frontiers for inefficient DMUs.

In classical transferable utility (TU) cooperative game theory, it is generally assumed that all players can form a cooperation coalition and that there are no limitations on the communication and cooperation possibilities among the players [9]. The presence or absence of restrictions in communication and cooperation possibilities between players may also cause or prevent their communication and cooperation. In many practical situations, there may be limited cooperation and communication between players because of certain social, economic, hierarchical, or technical structures [23]. Several models have been proposed and studied for limited cooperation and communication situations. Among them, we focus on the communication game introduced by Myerson [35]. A communication game consists of the TU cooperative game and an undirected graph in which vertices and links are the players and direct bilateral communication channels between players, respectively; see [26].

This paper first reviews the literature to explain the various proposed approaches to solve the major drawbacks of the traditional DEA model. Next, we present the ranking problem of DMUs in DEA from a communication game perspective, which is cooperative games enriched by a communication structure. The idea of the proposed approach is that if some efficient DMUs were not present, then the total share of other efficient DMUs would increase or decrease, and even some inefficient DMUs could become efficient. If some inefficient DMUs were not present, then the total share of efficient DMUs would decrease. The total share of an efficient DMU is the sum of the reference frontier share of that efficient DMU for each inefficient DMU. The total share of an inefficient DMU is the sum of the reference frontier share of each efficient DMU for that inefficient DMU.

The contribution of an efficient DMU can be measured by increasing or decreasing the total share of other efficient DMUs when this efficient unit is removed from the sample. The contribution of an inefficient DMU can

be measured by decreasing the total share of the efficient DMUs when this inefficient unit is removed from the sample. This means that an important efficient DMU is one in that the total share of other efficient DMUs increases by a large amount, and its total share decreases by a large amount whenever it is dropped from the sample. The important inefficient DMU is one that the total share of efficient DMUs decreases by a large amount whenever it is dropped from the sample. However, in the end, the total share of efficient and inefficient DMUs is collectively obtained by defining the frontier set of inefficient and efficient DMUs, respectively. Thus, we can extend the removal of a DMU to the removal of the different subsets of DMUs in the sample. The production possibility set (PPS) is derived collectively from DMUs; thus, the ranking DMUs can be considered a cooperative game. Since the efficiency of each inefficient DMU can only be improved by its reference frontier DMUs and the efficient DMUs can impact the shape of the efficient frontier, there are two cooperation and communication possibilities between inefficient units and their reference frontier units, and among two efficient DMUs that directly linked in the efficient frontier. Therefore, we define the limited TU cooperative game, called the communication game, to rank all DMUs. For TU communication games with a graph structure, a well-known allocation rule is the Myerson value. In the proposed game, the players are the DMUs, and the characteristic function measures the increase and decrease in the total share of each DMU.

DMUs cooperate in sharing their information to maximize efficiency and reach a win-win situation. For this purpose, cooperative games allow for calculating the maximum efficiency of any possible coalition. In DEA, efficient DMUs are the reference frontier for inefficient DMUs to improve their efficiency. In other words, the efficiency of inefficient DMUs is improved by their reference frontier, and efficient DMUs only improve the efficiency of their improved frontier. For this purpose, communication games provide to maximize efficiency in a situation where there is limited communication between efficient and inefficient DMUs. In some cases, the results of the ranking DMUs have been consistent or different for certain data sets, which is not unusual. Each ranking approach has a logic that gives it validity. The logic of our approach is defined based on the communication games that the coop-

eration of DMUs in defining the reference frontier and the Improved frontier gives validity to the approach. Therefore, our proposed method can provide acceptable results as a ranking approach.

The structure of this paper is as follows. In Section 2, five major DEA ranking approaches are reviewed. Section 3 briefly presents some preliminaries. Section 4 presents the approach proposed for ranking the DMUs. Some numerical examples related to datasets taken from the literature are reported in Section 5. Finally, we summarize and conclude in Section 6.

2 Literature review

Several approaches have been proposed to rank DMUs in DEA. Among them, we can refer to five major groups of approaches. The first group of DEA ranking approaches is those based on cross-efficiency. The cross-efficiency method evaluates the efficiency scores of each DMU by its favorable weights and the favorable weights of other DMUs obtained by the CRS multiplier DEA formulation. The average of these efficiency scores is the cross-efficiency scores of this DMU [44]. Many alternative secondary goals have been proposed because of the existence of alternative optimal weights. The first secondary goal models proposed by Doyle and Green [13] are the benevolent and aggressive strategic models. The benevolent and aggressive strategic models, respectively, maximize and minimize the cross-efficiency of other DMUs simultaneously while keeping their score optima. Liang et al. [29, 30] developed benevolent and aggressive secondary target models using slack variables for solving nonlinear models proposed by Doyle and Green. Wang and Chin [48] maximized the efficiency of each output of the DMU while keeping the efficiency of all DMUs at the maximum efficiency level. Subsequently, this model was extended by Wang, Chin, and Jiang [49] to maximize the efficiency of each input of the DMU simultaneously. The peer-restricted cross-efficiency model was proposed by Ramon, Ruiz, and Sirvent [39] that limits the differences among weights and removes variables with zero weights. Wang, Chin, and Luo[50] first used a virtual DMU in the cross-efficiency model and then increased and reduced the gaps between the weighted inputs or outputs associated with the evaluated DMU and the virtual DMU. Scholars have also

presented new ideas for secondary goal models [12, 54, 25, 32]. Table 1 lists a number of representative methods of this group.

Table 1: Some ranking methods based on cross-efficiency approaches.

Reference	Remarks
sexton, Silkman, and Hogan [44]	Cross-efficiency
Doyle and Green [13]	Secondary goals (aggressive/benevolent formulation)
Liang et al. [29, 30]	Extend the model of Doyle and Green
Wang [48]	New alternative models
Wang, Chin, and Jiang [49]	Simultaneously input- and output-oriented weight
Ramon, Ruiz, and Sirvent [39]	Select of the profiles of weights
Wang, Chin, and Luo [50]	A virtual ideal DMU and a virtual anti-ideal DMU

The second group corresponds to methods based on super-efficiency. Andersen and Petersen [4] introduced (AP) the super-efficiency for ranking extreme efficient DMUs. AP model drops the DMU under evaluation from the set of DMUs and then obtains a super efficiency score for this DMU. A major concern related to the AP models is the crucial infeasibility issue. Scholars have also proposed several models for solving the infeasibility problem of envelopment models or unboundedness of multiplier models [52, 45, 34]. Table 2 lists a number of representative methods of this group.

Table 2: Some ranking methods based on super-efficiency approaches.

Reference	Remarks
Andersen and Petersen [4]	Radial, infeasibility issues for VRS
Ma et al. [34]	Calculate more reliably by super efficiency
Xie et al. [52]	The efficiency of thermal power plants
Sojoodi, Dastmalchi, and Neshat [45]	Two super-efficiency models

A third major group of DEA ranking approaches is those based on reference-based approaches. This method defines references for all inefficient to determine the importance of efficient DMUs and improve the ranking of efficient DMUs. Bergendahl [7] developed the DEA model to evaluate the efficiency of a single bank and formed a reference set of banks, which is a

convex combination of efficient banks. Boljuncic [8] assessed the changes of inputs and outputs of an extreme efficient DMU thus obtained the region of efficiency for this DMU. Roshdi, Mehdiloozad, and Margaritis [43] used the least distance projection to determine the maximal closest reference set, which is the set of all possible reference DMUs associated with a given closest projection. Rezaeiani and Foroughi [42] introduced a reference frontier model to discriminate among efficient DMUs. This model measures the reference frontier share of each efficient DMU in improving the efficiency of inefficient DMU. Scholars have also presented new ideas for reference-based models [6, 5, 2]. Table 3 lists a number of representative methods of this group.

Table 3: Some ranking methods based on reference-based approaches.

Reference	Remarks
Bergendahl [7]	Use form a reference bank
Boljuncic [8]	Assess changes in an extreme efficient DMU
Roshdi, Mehdiloozad, and Margaritis [43]	The maximal closest reference set
Rezaeiani and Foroughi [42]	The reference frontier share of efficient DMUs
Behmanesh, Rahimi, and Gandomi [6]	test evolutionary many-objective algorithms

The fourth major group of DEA ranking approaches is those based on cross-influence. The cross-influence method excludes certain subsets of DMUs from the sample that defines the technology and then measures the impact on all the DMUs. A ranking model for extreme efficient DMUs was proposed by Jahanshahloo et al. [22]. This model first excludes the efficient DMUs from a set of DMUs and then forms the new efficient frontier. This new frontier gets closer to the inefficient DMUs, and some inefficient DMUs even become efficient. Du et al. [14] introduced an approach for ranking all DMUs based on two approaches of the influence of each efficient DMU on all the other DMUs and the standard efficiency scores. Hosseinzadeh Lotfi et al. [20] proposed a ranking approach based on applying aggregate units, which are artificial and are defined over efficient DMUs. This method first deletes one efficient DMU with better performance and then selects a set of input and output weights that can maximize the group efficiency once again. Izadikhah

and Farzipoor [21] proposed a new model to fully rank all DMUs. Their model first defines the concept of virtual DMU as an average of all inefficient DMUs and then measures the impact of efficient DMUs on virtual DMU and the impact of efficient DMUs on the influences of other efficient DMUs. Scholars have also presented new ideas for cross-influence models [40, 15, 18]. Table 4 lists a number of representative methods of this group.

Table 4: Some ranking methods based on cross-influence approaches.

Reference	Remarks
Jahanshahloo et al. [22]	Omission of efficient DMUs from reference set
Du et al. [14]	The influence of efficient DMUs on other DMUs
Hosseinzadeh Lotfi et al. [20]	Deletion of an efficient DMU on another ones
Izadikhah and Farzipoor [21]	Single virtual inefficient DMU
Ramon, Ruiz, and Sirvent [40]	The benchmarking of DMUs

The last major group of DEA ranking approaches is those based on game theory. In recent years, mechanisms of cooperation and non-cooperation in the DEA have attracted increasing attention. Nakabayashi and Tone [36] introduced a model based on DEA and the game theory approach for the consensus-making problems between individuals or organizations with multiple criteria to allocate and impute the given benefit or cost to the players. Wu et al. [51] proposed a method based on the Nakabayashi and Tone approach and cross-efficiency DEA model and obtained the common set of weights for fully ranking DMUs. Liang et al. [31] generalized the original DEA cross-efficiency concept to noncooperative game cross-efficiency. In their game model, each DMU (player) maximizes its cross-efficiency score when the cross-efficiency scores of each of the other DMUs do not deteriorate. The obtained game cross-efficiency scores were unique and constituted a Nash equilibrium point. Abing et al. [1] proposed an approach based on the multi-objective DEA model and Shapley value to solve the problem of many inputs and outputs and the small number of DMUs. Omrani, Beiragh, and Kaleibari [37] proposed the principal component analysis DEA model with the cooperative game for evaluating the performance of Iranian electricity distribution companies to increase the discriminant power of the DEA

model with many inputs and outputs. Rezaee, Izadbakhsh, and Yousef [41] introduced a cooperative game theory approach based on the DEA and Nash bargaining game for evaluating the performance of transportation systems by a large scale of measures. The first ranking problem of the efficient DMUs based on cooperative games and DEA was proposed by Li et al. [27]. They first dropped a nonempty subset of efficient DMUs from the set of DMUs and then measured the impact of its removal on the super-efficiency scores of the efficient units. Hinojosa et al. [19] proposed a new cooperative game approach in which first a nonempty subset of efficient DMUs from the set of DMUs dropped and then measured the impact of its removal on the efficiency scores of the inefficient DMUs. In Li et al. and Lozano et al. models, the set of players is the set of efficient DMUs. Omrani, Fahimi, and Mahmoodi [38] applied the game theory approach to increase distinguish power in the DEA model and obtained the fair weights in the cross-efficiency DEA context. Sun, Li, and Wang [47] proposed a comprehensive model based on game theory and the DEA model to improve resource utilization efficiencies and reduce pollutant emissions in the circular economic system. Soofizadeh and Fallahnejad [46] used the theory of cooperative games and the Shapley value method as a fair method to solve real practical problems with imprecise dates and rank airline groups. Their ranking for groups is based on the average marginal shares of groups in different coalitions. Fallahnejad, Asadi Rahmati, and Moradipour [16] used the Shannon entropy to create a new Shapley Vvalue obtained from aggregating the marginal effects of efficient DMUs weighted for ranking efficient DMUs. An et al. [3] proposed a cooperative social network partition approach by using DEA and game theory. The relationship among DMUs is defined by measuring the increased payoff of two cooperative DMUs. They employed the Shapley value to achieve community partition. Ghaeminasab et al. [17] proposed a coalition game to solve the revenue allocation problem by considering cooperative relations among DMUs. They used the concept of DEA efficiency, a new characteristic function, and the Shapley value. Chang, Lin, and Ouennich [10] applied the first DEA-based Nash bargaining models to help acquirer companies and obtain their most desired target companies. Lozano [33] proposed a bargaining model where players are input or output variables for each DMU. Players try

to reduce an input variable or an undesirable output and increase a desirable output. Yu and Rakshit [53] computed the input and output targets using the bargaining approach and DEA for a sample of major global airlines. This model performs an unbiased, rational negotiation between the inputs and outputs to achieve reasonable and optimal solutions for inputs and outputs. Table 5 lists a number of representative methods of this group.

Table 5: Some ranking methods based on game theory.

Reference	Remarks
Nakabayashi and Tone [36]	Solve egoists dilemma using cooperative game
Wu et al. [51]	Bargaining game model to evaluate DMUs
Liang et al. [31]	The DEA game cross-efficiency model
Abing et al. [1]	Shapley value-based multi-objective DEA
Omrani, Beiragh, and Kaleibari [37]	DEA cooperative game
Jahangoshai, Izadbakhsh, and Yousef [41]	DEA-game for ranking of operational
Li et al. [27]	Super-efficiency evaluation by a cooperative game
Hinojosa et al. [19]	Ranking efficient DMUs using cooperative game
Omrani, Fahimi, and Mahmoodi [38]	Game theory to find weights in cross-efficiency
Sun, Li, and Wang [47]	A game meta-frontier DEA approach
Soofizadeh and Fallahnejad [46]	Evaluation by using cooperative game fuzzy DEA
Fallahnejad, Asadi Rahmati, and Moradipour [16]	An entropy based Shapley value to rank
An et al. [3]	Cooperative social network community partition
Ghaeminasab et al. [17]	The Coalitional Game and DEA
Chang, Lin, and Ouennich [10]	DEA-based Nash bargaining approach
Lozano [33]	Bargaining approach for efficiency assessment
Yu and Rakshit [53]	The DEA bargaining approach

In recent studies, it is generally considered that players can form a cooperation coalition and there are no limitations on the communication and cooperation possibilities among them. The purpose of this paper is to propose a model from a perspective of limited communication situations to rank all DMUs. The basic assumption in this paper is that there are limitations on the communication and cooperation possibilities among the players. In other words, we develop a ranking model of the DMUs through the DEA reference frontier share model and using TU game theory ideas with restrictions in communication and cooperation. Based on these, we define the communication game and use the Myerson value defined as the Shapley value of the modified game.

3 Preliminaries

The TU cooperative games describe situations where players can form coalitions and payoffs of coalitions can be freely distributed among its members. Mathematically, a TU cooperative game is a pair (N, v), where N is a finite set of players and $v : 2^N \to R$ is a characteristic function with $v(\emptyset) = 0$. A subset $S \subseteq N$ called a coalition. For a subset $S \subseteq N$, the real number $v(S)$ represents the worth of the coalition S.

A value is a function f that assigns to every cooperative game (N, v), a subset $f(v)$ of R^n is called an allocation rule. The most well-known allocation rule that distributes the worth of cooperation among players in TU cooperative game theory is the Shapley value. The Shapley value of the cooperative game for all $i \in N$ is defined by

$$\psi_i(N, v) = \sum_{S \subset N \setminus \{i\}} \frac{s!(n-s-1)!}{n!} \, v(S \cup \{i\}) - v(S), \tag{1}$$

where n and s are the cardinality of N and S, respectively.

The Shapley value is the unique allocation rule that satisfies the following four appealing properties:

Efficiency. This property requires that the worth of coalition N is fully distributed among the n players. Mathematically, for every game (N, v), we have

$$\sum_{i \in N} \psi_i(N, v) = v(N).$$

Symmetry. This property demands that two symmetric players get the same payoff. Mathematically, for every game (N, v) and every pair (i, j) of symmetric players, we have

$$\psi_i(N, v) = \psi_j(N, v).$$

Null Player Property. For every game (N, v) and every null player $i \in N$, we have

$$\psi_i(N, v) = 0.$$

Additivity. This property requires the allocation rule to be additive. For every pair of games (N, v) and (N, w), we have

$$\psi(N, v + w) = \psi(N, v) + \psi(N, w).$$

The existence of communication and cooperation restrictions between players in cooperative game theory led to the introduction of new theoretical models commonly referred to as communication games. Mathematically, a communication game is denoted by the triplet (N, v, g), where (N, v) is a TU cooperative game and (N, g) is a graph on N describing the cooperation and communication possibilities between players. In a graph on N, N is a set of vertices and links are direct bilateral communication channels between players. Let us consider a graph (N, g) and two players $i, j \in S$. A link ij exists in the graph g if and only if the players i and j can directly communicate. For any $i, j \in S$, i and j are connected in S if and only if $i = j$ or there are $i = i_0$, i_1, ..., $i_k = j$ such that $\{i_t, i_{t+1}\} \in g$ for $0 \leq t < k$, namely $i = j$ or i and j are directly linked or indirectly linked in S. If each pair of players in S is connected, then S is connected. A maximal (not necessarily maximum) connected subset of S is called a component of S in g, which is called a component simply if $S = N$. We denote by $N|_g$ and $S|_g$ the set of all components belonging to N in g and the set of all components belonging to S in g, respectively.

The most well-known allocation rule for TU communication games is the Myerson value. Myerson defined the restricted characteristic function $v^g : 2^N \to R$ associated with the communication game (v, g) as follows:

$$v^g(S) = \sum_{C \in S|_g} v(C), \qquad \text{for all } S \subseteq N. \tag{2}$$

The interpretation of the restricted characteristic function v^g can be that if $S \subseteq N$ is connected, then players in coalition S can communicate and obtain their initial payoff $v(S)$. If S is nonconnected, then players can only communicate with members of the same connected component and receive the sum of the worth obtained by that component. If g is complete, then every coalition $S \subseteq N$ is connected, and v^g coincides with the initial game v.

The Myerson value of a communication game in TU communication game is a vector $M(N,v) \in \Re^N$ for all $i \in N$ defined by

$$M_i(N,v,g) = \sum_{S \subset N \setminus \{i\}} \frac{s!(n-s-1)!}{n!} \, v^g(S \cup \{i\}) - v^g(S). \tag{3}$$

The Myerson value is the Shapley value of the TU communication game in which the worth of each coalition equals the sum of the worth of each connected component belonging to the coalition. The Myerson value can be uniquely determined by component efficiency and fairness properties.

Component efficiency. This property requires that the payoff of each coalition, corresponding to one of the connected components of the communication game is fully distributed among its members. Mathematically, for every communication game and component $C \in N|_g$, we have

$$\sum_{i \in C} M_i(N,v,g) = v(C).$$

The component efficiency property coincides exactly with the efficiency property of Shapley value if the graph g is connected.

Fairness. This property requires that the removal of a link in the TU communication game leads to changes in the payoffs of both players by the same amount. Mathematically, for every communication game (N,v,g) and every link $(i,j) \in g$, we have

$$M_i(N,v,g) - M_i(N,v,g \setminus ij) = M_j(N,v,g) - M_j(N,v,g \setminus ij).$$

We will define the TU communication game and use the Myerson value as the allocation rule to rank DMUs by using a DEA reference frontier share model. The DEA reference frontier share model is defined as follows. Suppose for a sample with m independent DMUs and $j \in M = \{1,2,\ldots,m\}$ that each DMU consumes k inputs x_{ij} $(i \in I = \{1,2,\ldots,k\})$ to generate h outputs $y_{rj} (r \in H = \{1,2,\ldots,h\})$. Let DMU_o be a member of the set of efficient DMUs and let $DMUp$ be a member of the set of inefficient DMUs. The reference frontier of DMU_p, denoted by N_p, is all possible efficient DMUs for DMU_p that dominate it. The reference frontier share of an efficient DMU_o for inefficient DMU_p is defined as the maximum value of the coefficient asso-

ciated with DMU_o in all convex combinations of reference frontier points of DMU_p. The reference frontier share of DMU_o for DMU_p, denoted by λ_{po}, is defined as follows:

$$\lambda_{po} = \max \{\lambda_o | (\sum_{j\in E} \lambda_j x_j, \sum_{j\in E} \lambda_j y_j) \in N_p \quad \sum_{j\in E} \lambda_j = 1, \lambda_j \geq 0, \quad (j \in E)\}. \tag{4}$$

Let E and IE be the set of efficient and inefficient DMUs, respectively. The total share of DMU_o is defined by $\overline{\lambda}_o = \sum_{p\in IE} \lambda_{po}$, requiring that IE is nonempty. The reference frontier share model for ranking efficient DMUs is as follows:

$$\begin{aligned}
&\lambda_{po}(M) = \max \ \ \lambda_o \\
&\text{s.t.} \qquad \sum_{j\in E} \lambda_j x_j \leq x_{ip}, && i = 1, \ldots, m, \\
&\qquad \sum_{j\in E} \lambda_j y_j \geq y_{rp}, && r = 1, \ldots, s, \\
&\qquad \sum_{j\in E} \lambda_j = 1, \\
&\qquad \sum_{i=1}^{m} v_i x_{ij} - \sum_{r=1}^{s} u_r y_{rj} + u_o = d_j, && j \in E, \\
&\qquad \lambda_j \leq t_j, && j \in E, \\
&\qquad d_j \leq M'(1 - t_j), && j \in E, \\
&\qquad v_i \geq 1, && i = 1, \ldots, m, \\
&\qquad u_r \geq 1, && r = 1, \ldots, s, \\
&\qquad u_o \ \ free, \\
&\qquad \lambda_j \geq 0 \ \ d_j \geq 0 \ \ t_j \in \{0, 1\}, && j \in E.
\end{aligned} \tag{5}$$

The constraint $\sum_{j\in E} \lambda_j = 1$ assures us that for some DMU_j, we have $\lambda_j > 0$. Hence, for these DMU_j, we have $t_j = 1$ and $d_j = 0$. For each $T \subseteq M$, let $\lambda_{po}(T)$ be the reference frontier share of the DMU_o for DMU_p when the sample of DMUs under evaluation is the DMUs in T. Rezaeiani and Foroughi proved that model (5) is always feasible and includes an efficient point of the PPS. In their model, M' is a sufficiently large positive number.

The first ranking approach of efficient DMUs based on cooperative game theory is proposed by Li et al. [27]. In their model, players are the efficient DMUs, and the characteristic function measures the changes in the super-efficiency of the efficient DMUs belonging to S. Formally, the proposed game

of Li et al. is a cooperative TU game (N, v), where the characteristic function, for each $S \subseteq N$, is $v(s) = \sum_{k\in N} E_k(M \setminus S) - 1$. Next, Hinojosa et al. [19] introduced a new TU cooperative game approach. In this game, the characteristic function measures the changes in the efficiency scores of the inefficient DMUs when a given coalition of efficient units is dropped from the set of DMUs. Formally, the game defined by Lozano et al. is a cooperative TU game (N, v), where the characteristic function for each $S \subset N$ is $v(S) = \sum_{j\in M\setminus N} E_j(M \setminus S) - E_j(M)$. Moreover, $E_j(T)$ is the efficiency score of the $DMU_j \in T$ obtained by the input-oriented CCR model when the sample includes the DMUs belonging to T.

4 The proposed approach

In the present paper, the total share is proposed for each of the inefficient and efficient DMUs to increase the discrimination power among DMUs. We define the concept of the reference frontier share for inefficient DMUs. We concentrate on the total share of DMUs that are removed from the sample and efficient DMUs outside coalition S; that is, the role of the deleted units in changing the total share of other DMUs and the role of the remaining efficient units in improving the efficiency of the inefficient DMUs in the modified sample are considered. A TU communication game approach is proposed in which the characteristic function measures the effect of removing a subset of DMUs on the total share of all DMUs. This communication game consists of a TU cooperative game and an undirected graph in which the vertices are the DMUs. In the proposed graph, a link represents a direct bilateral communication channel between an inefficient DMU and its reference DMU and among two efficient DMUs that are directly linked in the efficient frontier.

The reference frontier of an inefficient DMU is the set of all efficient DMUs of the efficient frontier that dominate it. The improved frontier of an efficient DMU is the set of all inefficient DMUs of observed DMUs that this efficient DMU dominates them. Each point of the reference frontier of an inefficient DMU can improve that inefficient DMU and each point of the improved frontier of an efficient DMU can be improved by that efficient DMU. The reference frontier share of an efficient DMU is defined as the maximum value

of the efficient DMU coefficient in all convex combinations of the reference frontier points of an inefficient DMU. The reference frontier share of efficient DMUs for inefficient DMUs is presented in a matrix. In this matrix, each column corresponds to the efficient DMUs, and each row corresponds to all DMUs. The total share of the efficient DMU is the sum of the reference frontier share of that efficient DMU for its improved frontier points. This share is equal to the sum of the values in the column associated with that DMU. The total share of the inefficient DMU is the sum of the reference frontier share of its reference frontier points for that inefficient DMU. This share is equal to the sum of the values in the row associated with that DMU. The closer an inefficient unit is to the efficient frontier, the less efficient units are involved in improving its efficiency, and the reference frontier of this inefficient unit is smaller. As a result, the reference frontier share of this inefficient unit is lower. Therefore, the lower the reference frontier share of an inefficient unit, the higher the efficiency of that inefficient unit than other inefficient units.

Since the reference frontier share is defined by efficient and inefficient DMUs, removing an efficient or inefficient DMU leads to changes in the total share for each DMU. When the efficient DMU is eliminated, the total share of other efficient DMUs would decrease or increase. If the weak efficient or inefficient DMUs become extreme efficient by removing one efficient DMU, then the total share of the efficient DMUs would decrease. The total share of efficient DMUs would increase when the weak efficient or inefficient DMUs do not become extreme efficient by removing one efficient DMU. Let DMU_o be dropped from the original sample M, $S = \{DMU_o\}$, let the inequality $\overline{\lambda}_j(M) \leq \overline{\lambda}_j(M \setminus S)$ or $\overline{\lambda}_j(M) \geq \overline{\lambda}_j\ (M \setminus S)$ hold for efficient $DMU_j \in M \setminus \{DMUo\}$, and let $\overline{\lambda}_o(M) \geq \overline{\lambda}_o(M \setminus S)$. When the inefficient DMU is eliminated, the total share of efficient units would decrease. If DMU_p is dropped from M, $S = \{DMU_p\}$, then the inequality $\overline{\lambda}_j(M) \geq \overline{\lambda}_j(M \setminus S)$ holds for each efficient DMU. Therefore, if only a certain subset of the efficient DMUs is dropped from the set of DMUs, then the total share of the deleted efficient DMUs for all inefficient DMUs decreases by an amount equal to $\overline{\lambda}_j(M) - \overline{\lambda}_j(M \setminus S)$ and the total share of the remaining efficient DMUs would increase or decrease by an amount equal to $\overline{\lambda}_j(M \setminus S) - \overline{\lambda}_j(M)$ or

$\overline{\lambda}_j(M)-\overline{\lambda}_j(M\setminus S)$. If only a certain subset of the inefficient DMUs is dropped from the set of DMUs, then the total share of efficient DMUs decreases by an amount equal to $\overline{\lambda}_j(M)-\overline{\lambda}_j(M\setminus S)$. The total share of DMUs in $M\setminus S$ decreases by an amount equal to $\overline{\lambda}_j(M)-\overline{\lambda}_j(M\setminus S)$ or increases by an amount equal to $\overline{\lambda}_j(M\setminus S)-\overline{\lambda}_j(M)$ when a certain subset of the efficient and inefficient DMUs is dropped from the set of DMUs. Moreover, $\overline{\lambda}_j(M\setminus S)-\overline{\lambda}_j(M)$ is the increment on the total share that jth DMU enjoys when the DMUs in coalition S drop from the sample. The difference $\overline{\lambda}_j(M)-\overline{\lambda}_j(M\setminus S)$ is the reduction on the total share that DMU_j suffers when the DMUs in coalition S drop from the sample.

Therefore, we can define a cooperative TU game (N,v) in which the players are the DMUs, and the characteristic function for each nonempty coalition $S\subseteq N$ is as follows:

$$v(S)=\sum_{k\in S}\overline{\lambda}_k(M)-\overline{\lambda}_k(M\setminus S)+\sum_{j\in M\setminus(S\cup P)}\overline{\lambda}_j(M)-\overline{\lambda}_j(M\setminus S), \qquad (6)$$

where P is the set of all inefficient DMUs belonging to $M\setminus S$.

Since the reference frontier share model aims to measure the share of each efficient DMU in improving the efficiency of inefficient DMUs and the reference frontier share of efficient units is defined by the reference frontier of inefficient units, the inefficient DMUs and their reference frontier DMUs are invited to the game. The game only runs between efficient and inefficient DMUs. In our proposed approach, there exists a commodity that is the corresponding benchmark credit for all DMUs. This commodity represents the total share by the reference frontier share model, which can be transferred among the players. Since the reference frontier share is defined by the efficient and inefficient DMUs, there is only cooperation between inefficient DMUs and their reference frontier. Therefore, there are communication restrictions in a cooperative game. The allocation rule is defined as distributing the total share among the efficient and inefficient DMUs. We use the well-known allocation rule called Myerson value to rank DMUs. Myerson value is the Shapley value in the modified game (N,v^g) and can be specified by the performance and fairness features of the component. The proposed game on a player set N is a triplet (N,v,g) such that (N,v) is a TU game and

(N, g) is a graph. A graph (N, g) consists of vertices in which the vertices are the DMUs and a link represents a direct bilateral communication channel between inefficient DMU and its reference DMU and among two efficient DMUs that are directly linked in the efficient frontier. The characteristic function is the changes in the total share of the DMUs in the modified sample in which the DMUs belonging to coalition S were removed.

The TU in cooperative game theory assumes that one player can freely transfer part of their utility to another player and the players have a common commodity that is valued equally by all and enjoys the same total utility. In the proposed approach, we consider a commodity that stores utility and which can be transferred among the players. The characteristic function of the grand coalition represents the total share uncovered by the DEA reference frontier share. Thus, the implicit commodity is the benchmark credit for contributing to uncovering the total share present in the observations. It is clear that, since the reference frontier share is defined by efficient and inefficient DMUs, removing an efficient or inefficient DMU leads to changes in the total share for each DMU. These are the DMUs that deserve the credit. The credit allocation obtained with the Myerson value is used to rank the efficient and inefficient DMUs.

The following algorithm expresses the importance of DMUs in our proposed approach:

Step 1. Let $DMU_k \in N$ be as a player and list all possible subsets N without player k.

Step 2. List the reference frontier or improved frontier of DMU_k, define the graph (N, g) and obtain $S|_g$.

Step 3. Solve model (5) four times and obtain $\overline{\lambda}_k(M)$ and $\overline{\lambda}_k(M \setminus S)$ for $k \in S$, and $\overline{\lambda}_j(M)$ and $\overline{\lambda}_j(M \setminus S)$ for $j \in M \setminus (S \cup P)$.

Step 4. Based upon (6), Calculate $v^g(S)$.

Step 5. Calculate the Myerson value of DMU_k on the use of formulations (2) and (3).

Example 1. [42] Table 6 shows the dataset related to eight DMUs with one input and one output. Consider the sample that includes all DMUs, $S = \emptyset$. In this case, DMUs A, B, C, F, and G appear to be efficient. The reference

frontier of inefficient DMU_D for $S = \emptyset$ is shown in Figure 1. The reference frontier share model obtains the reference frontier share of the efficient DMUs for all DMUs and is represented as a matrix below:

DMU	A	B	C	F	G
A	1	0	0	0	0
B	0.5	1	0.5	0	0
C	0	0	1	0	0
D	0.75	1	1	0.5	0.333
E	0.25	0.5	1	0	0
F	0	0	0.333	1	0.666
G	0	0	0	0	1
H	0.25	0.5	1	1	0.666

Table 6: Dataset and the total share of all DMUs

DMUs	A	B	C	D	E	F	G	H
x_1	2	3	4	5	4	6	7	6
y_1	3	5	7	4	6	8	8.5	6
$\overline{\lambda}_j(M)$	1.75	2	3.833	3.5833	1.75	1.5	1.667	3.4167
Ranking	3	2	1	8	6	5	4	7

The total share of each efficient DMU is equal to the sum of the values of each column, for example, the result $\overline{\lambda}_A = 1.75$, $\overline{\lambda}_B = 2$, $\overline{\lambda}_C = 3.83$, $\overline{\lambda}_F = 1.5$ and $\overline{\lambda}_G = 1.67$. The more the total share of the efficient DMU, the higher its ranking. The total share of each inefficient DMU is equal to the sum of the values for each row, for example, $\overline{\lambda}_D = 3.584$, $\overline{\lambda}_E = 1.75$ and $\overline{\lambda}_H = 3.417$. The higher the total share of the inefficient DMU, the lower its efficiency. In an undirected graph, vertices are efficient and inefficient DMUs, and links represent direct bilateral communication channels between inefficient DMUs and their reference frontier or efficient DMUs and their improved frontier. To illustrate this graph, we consider two efficient DMU A and inefficient DMU D. Two DMUs A and D are vertices A and D in the graph. The reference frontier of DMU D is $\{A, B, C, F, G\}$, and the improved frontier

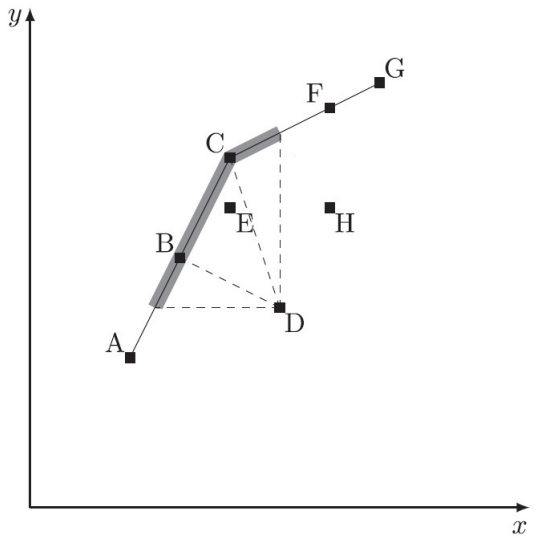


Figure 1: The reference frontier of DMU_D in $S = \emptyset$

of DMU A is $\{B, D, E, F\}$. Therefore, vertices A and D connect to their improved frontier and reference frontier, respectively. The cooperation and communication in the graph (N, g) are shown in Figure 2. According to the graph (N, g) and $S \subseteq N$, the set of all components that belong to S in g, $S|_g$, is defined and then $v^g(S)$ is calculated.

For the coalition $S = \{B, C\}$, the modified sample includes the inefficient DMUs D and H and efficient DMUs A, E, F, and G. Thus, the DMUs outside S coalition define a new efficient frontier. It means that DMUs B and C belonging to the coalition S are not present in the sample. The reference frontier of D inefficient DMU on the modified sample and new efficient frontier is shown in Figure 3.

Let us consider two coalitions $S = \{B, C\}$ and $\overline{S} = \{D, E, F, G, H\}$. The coalition S is connected, $S|g = S$, and therefore players belonging to S obtain their initial worth $v(S)$, that is, $v^g(S) = S$. The coalition $\overline{S}$ is not connected. As shown in Figure 4, $\overline{S}|g = \{\{D, F, G, H\}, \{E\}\}$, and we have $v^g(\overline{S}) = v(\{D, F, G, H\}) + v(\{E\})$.

For the coalition $S = \{B, C\}$, the modified reference frontier share of DMU_A, DMU_E, DMU_F, and DMU_G for inefficient DMU D are 0.8, 1,

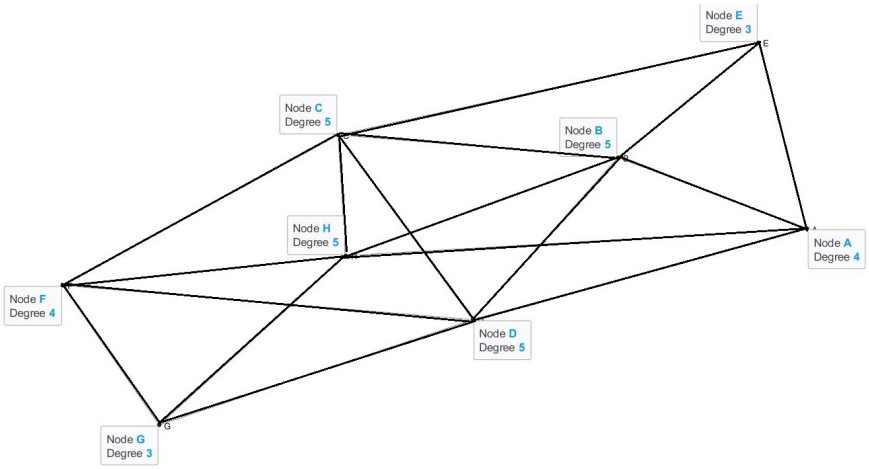


Figure 2: The graph (N, g).

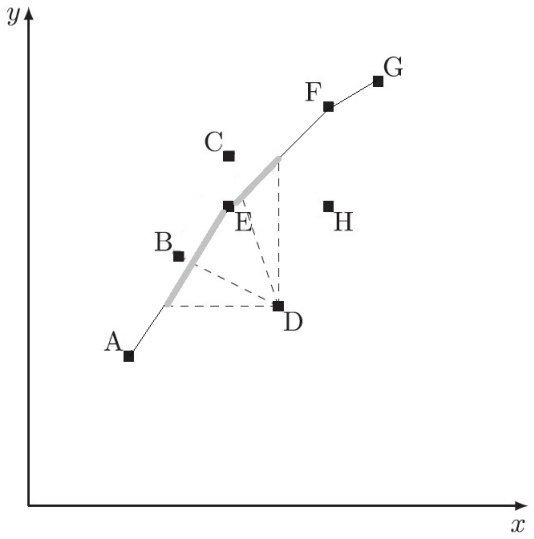


Figure 3: The reference frontier of DMU_D and new efficient frontier for $S = \{B, C\}$ coalition

0.75, and 0.6, respectively, the modified reference frontier share of DMU_A, DMU_E, DMU_F, and DMU_G for inefficient DMU H is 0.4545, 1, 1, and 0.67, respectively. Therefore, the modified reference frontier share of efficient

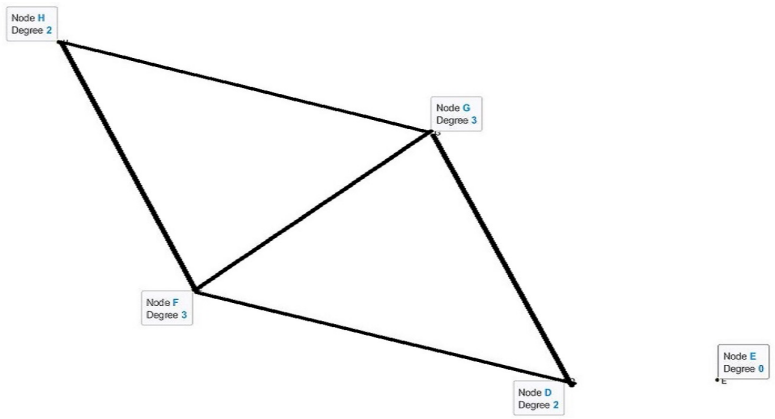


Figure 4: The graph $(\overline{S}, g)$ and the connected components.

DMUs for inefficient DMUs has increased compared to the reference frontier share of the sample $S = \emptyset$. the total share for DMU_A, DMU_D, DMU_E, DMU_F, DMU_G, and DMU_H are 1.2545, 3.15, 2, 1.75, 1.2667, and 3.1212, respectively.

The modified total share has decreased compared to the total share of the sample $S = \emptyset$ because of removing the efficient B and C and appearing DMU_E to be efficient. Therefore, the characteristic value of coalition S in the cooperative game is 4.7255. This value means that removing DMUs B and C from the sample leads to a reduction in the sum of the total share of DMUs. The Myerson value of each DMU and their ranking are shown in Table 7. Efficient and inefficient units are ranked separately. The higher the Myerson value in the efficient units, the more their efficiency. The higher the Myerson value in the inefficient units, the less their efficiency.

Table 7: The Spearman rank correlation coefficient: 0.8095

DMUs	A	B	C	D	E	F	G	H
Myerson value	1.179	1.625	3.259	1.067	2.220	0.189	0.118	4.170
Ranking	3	2	1	6	7	4	5	8
super-efficiency	3	2	1	8	6	5	4	7

5 Numerical illustrations

The proposed approach will be illustrated using a data set from the literature in this section. Then we compare our results with the results obtained by some approaches related to Section 2. We use the Spearman rank correlation coefficient for every four cases.

Example 2. [19] Table 8 shows the dataset related to six DMUs ($M = \{A, B, C, D, E, F\}$) that include two inputs and one output. The total share $\overline{\lambda}_j(M)$ for the DMUs is shown in Table 8. Table 9 shows the result of ranking approaches.

Table 8: Dataset and the total share of all DMUs

DMUs	A	B	C	D	E	F
x_1	13	6	2	1	9	4
x_2	1	3	6	10	5	8
y_1	1	1	1	1	1	1
$\overline{\lambda}_j(M)$	0.4286	1.5000	1.6667	0.5000	2.0952	2.0000
Ranking	4	2	1	3	6	5

Table 9: The proposed ranking of all DMUs. The Spearman rank correlation coefficient: 0.8

DMUs	A	B	C	D	E	F
Myerson value	0.0390	0.8258	1.3425	0.0153	3.2465	2.0235
Ranking	3	2	1	4	6	5
Jahanshahloo	4	2	1	3	–	–

Example 3. [19] We consider the dataset containing eight DMUs listed in Table 10 adapted from Cooper et al. by adding two inefficient DMUs. Results are shown in Table 11.

Table 10: Dataset and the total share of all DMUs

DMUs	A	B	C	D	E	F	G	H
x_1	2	2	5	10	3	10	4	5
x_2	7	12	5	4	6	6	12	11
y_1	3	4	2	2	1	1	2.5	3.5
y_2	1	1	1	1	1	1	1	1
$\overline{\lambda}_j(M)$	2.166	1.8	1.5	1.25	1.25	3.666	3.00	1.3
Ranking	1	2	3	4	5	8	7	6

Table 11: The proposed ranking of all DMUs. The Spearman rank correlation coefficient: 0.6 and 0.7, respectively

DMUs	A	B	C	D	E	F	G	H
Myerson value	1.952	0.814	0.989	0.254	0.199	2.562	3.443	5.372
Ranking	1	3	2	4	5	6	7	8
super-efficiency	1	2	4	3	5	6	8	7
Li (2016)	1	2	3	4	5	–	–	–

Example 4. [24] Table 12 shows the dataset belonging to twelve DMUs with three inputs and two outputs. Moreover, $DMU_4, DMU_5, DMU_6, DMU_8, DMU_9$ and DMU_{10} are efficient DMUs. Table 13 compares the ranking of the proposed approach with other ranking approaches.

Example 5. [28] Table 14 shows the dataset of 18 city commercial bank branches in China with three inputs and three outputs. Feng et al. integrated cooperative game theory and the cross-efficiency method to develop a DEA-game cross-efficiency approach to generate a unique and fair allocation plan. Table 15 compares the ranking of the proposed approach with the DEA-game cross-efficiency approach introduced by Feng et al..

To check the compatibility of the three ranking approaches of Arbitrary, Benevolent, and Aggressive with our approach, we compare their results. By comparing the results of the four ranking methods, we conclude that there are similarities between the three ranking approaches and our approach. The

Table 12: Dataset and the total share of all DMUs

DMUs	x_1	x_2	x_3	y_1	y_2	$\overline{\lambda}_j(M)$	Ranking
1	350	39	9	67	751	3.9	12
2	298	26	8	73	611	2.58	9
3	422	31	7	75	584	3.57	11
4	281	16	9	70	665	3.03	3
5	301	16	6	75	445	3.07	2
6	360	29	17	83	1070	0.92	6
7	540	18	10	72	457	2.36	8
8	276	33	5	74	590	2.42	4
9	323	25	5	75	1074	4.66	1
10	444	64	6	74	1072	2.6	10
11	323	25	5	25	350	1	7
12	444	64	6	104	1199	1.89	5

Table 13: The proposed ranking of all DMUs. The Spearman rank correlation coefficient: 0.8671 and 0.8462, respectively

DMUs	1	2	3	4	5	6	7	8	9	10	11	12
The proposed approach	11	12	9	2	3	6	7	5	1	8	10	4
super-efficiency	10	7	11	3	2	6	8	5	1	9	12	4
Khodabakhshi and Aryavash approach	11	7	9	4	2	6	10	5	1	8	12	3

mean Pearson correlation coefficient of Arbitrary, Benevolent, and Aggressive with our approach equals 0.75439. As a result, this Pearson correlation coefficient confirms the similarity of ranking results between our method and Arbitrary, Benevolent, and Aggressive. Therefore, according to the obtained results, the proposed approach can give an acceptable ranking of DMUs.

In Examples 3 and 4, the Pearson correlation coefficient shows that the ranking approaches coincide with the ranking obtained in our approach. Our approach is the rationale based on the theoretical basis of the reference fron-

Table 14: The dataset of 18 branches of the city commercial bank

DMUs	x_1	x_2	x_3	y_1	y_2	y_3
1	62	1822	1361	140,117	130,288	5260
2	80	1833	1565	213,774	145,761	10,773
3	129	3595	1378	194,084	130,556	8006
4	62	1978	333	87,876	49,454	4479
5	89	2138	549	107,091	60,872	5897
6	84	1910	704	97,472	94,310	3849
7	36	1234	840	114,001	80,019	5292
8	172	4348	959	366,423	306,926	12,479
9	62	879	1253	107,393	86,485	5132
10	53	2566	483	69,691	43,907	3869
11	92	1348	419	148,458	87,193	7234
12	39	1229	513	83,752	40,046	3984
13	144	4640	1323	223,539	211,466	10,655
14	47	2248	670	70,555	65,110	2205
15	39	1571	362	99,143	66,736	5271
16	56	1635	669	112,513	79,366	5202
17	34	939	867	87,660	56,157	3000
18	58	1807	419	88,334	67,160	4171

tier share model and the communication game theory. Based on the comparison and analysis results of the empirical application, the proposed approach will give a unique and acceptable to all DMUs.

6 Conclusion

Discriminating among all DMUs is an important issue in DEA. The lack of discrimination power between the efficient DMUs leads to the major drawbacks of the traditional DEA models. In this paper, a new approach to DEA is proposed for improving the discrimination power of the traditional DEA model. This is based on the Myerson value, which is a well-known allocation

Table 15: The proposed ranking of all DMUs. The mean Spearman rank correlation coefficient: 0.75439

DMUs	Myerson value	Arbitrary	Benevolent	Aggressive
1	0.8259	0.7631	0.9906	0.6858
2	0.8349	0.8429	0.9998	0.7967
3	0.8130	0.7092	0.9980	0.6491
4	0.8173	0.7366	0.9953	0.7216
5	0.8154	0.7228	0.9974	0.6948
6	0.8104	0.7022	0.9874	0.6412
7	0.8268	0.8251	0.9915	0.7686
8	0.84165	0.8521	1.0000	0.7959
9	0.8267	0.7683	0.9933	0.7209
10	0.8272	0.6861	0.9982	0.6606
11	0.8392	0.8413	0.9941	0.8271
12	0.8239	0.7537	0.9938	0.7025
13	0.8251	0.7534	0.9983	0.7014
14	0.8285	0.6651	0.9870	0.6076
15	0.8375	0.8351	0.9894	0.8114
16	0.8369	0.7732	0.9915	0.7153
17	0.8201	0.7469	0.9922	0.6742
18	0.8183	0.7447	0.9903	0.7047

rule of the communication game theory. The players of the proposed TU communication game are the DMUs. The characteristic function measures the payoff of a coalition in terms of the amount of the total share obtained by the removal of the corresponding DMUs in the modified sample. If some efficient DMUs are not present in the sample, then the total share of other efficient DMUs increases or decreases. If the inefficient DMUs do not cooperate, then the total share of efficient DMUs decreases. As the efficient and inefficient DMUs cooperate and are present in the sample, the modified total share of the efficient DMUs becomes lower and closer to the initial total share and the modified total share of the inefficient DMUs becomes higher and closer to the initial total share. The proposed approach on some datasets from the

literature has been tested and used the Spearman rank correlation coefficient for comparison with the results of other methods in each case. Our approach is the rationale based on the theoretical basis of the reference frontier share model and communication game theory. It is based on the participation of each efficient DMU in improving the efficiency of the inefficient DMUs and the participation of each inefficient DMU in creating the reference frontier and defining the total share for the efficient DMUs. Therefore, the proposed approach can be accordingly credited.

As topics for further research, we can consider other TU communication game allocation rules, such as the position value. We can also define the games with particular coalition structures or partition of the set of players and then investigate allocation rules such as Owen value.

References

[1] Abing, S.L.N., Barton, M.G.L., Dumdum, M.G.M., Bongo, M.F., and Ocampo, L.A. *Shapley value-based multi-objective data envelopment analysis application for assessing academic efficiency of university departments,* J. Ind. Eng. Int. 14(4) (2018) 733–746.

[2] Allevi, E., Basso, A., Bonenti, F., Oggioni, G., and Riccardi, R. *Measuring the environmental performance of green SRI funds: A DEA approach,* Energy Econ. 79 (2019) 32–44.

[3] An, Q., Wang, P., Zeng, Y., and Dai, Y. *Cooperative social network community partition: A data envelopment analysis approach,* Comput. Ind. Eng. 172 (2022) 108658.

[4] Andersen, P., and Petersen, N.C. *A procedure for ranking efficient units in data envelopment analysis,* Manag. Sci. 39 (1993) 1261–1264.

[5] Ang, S., Zheng, R., Wei, F., and Yang, F. *A modified DEA-based approach for selecting preferred benchmarks in social networks,* J. Oper. Res. Soc. 72(2) (2021) 342–353.

[6] Behmanesh, R., Rahimi, I., and Gandomi, A.H. *Evolutionary many-objective algorithms for combinatorial optimization problems: A comparative study,* Arch. Comput. Methods Eng. 28(2) (2021) 673–688.

[7] Bergendahl, G. *DEA and benchmarks–an application to Nordic banks,* Ann. Oper. Res. 82 (1998) 233–250.

[8] Boljunčić, V. *Sensitivity analysis of an efficient DMU in DEA model with variable returns to scale (VRS),* J. Product. Anal. 25(1) (2006) 173–192.

[9] Caulier, J.F., Skoda, A., and Tanimura, E. *Allocation rules for networks inspired by cooperative game-theory,* Rev. Econ. Polit. 127(4) (2017) 517–558.

[10] Chang, T.S., Lin, J.G., and Ouenniche, J. *DEA-based Nash bargaining approach to merger target selection,* Eur. J. Oper. Res. 305(2) (2023) 930–945.

[11] Charnes, A., Cooper, W., and Rhodes, E. *Measuring the efficiency of decision-making units,* Eur. J. Oper. Res. 2 (1978) 429–444.

[12] Davtalab-Olyaie, M., Ghandi, F., and Asgharian, M. *On the spectrum of achievable targets in cross-efficiency evaluation and the associated secondary goal models,* Expert Syst. Appl. 177 (2021) 114927.

[13] Doyle, J. and Green, R., *Efficiency and cross efficiency in DEA: Derivations, meanings and the uses,* J. Oper. Res. Soc. 45(5) (1994) 567–578.

[14] Du, J., Liang, L., Yang, F., Bi, G.B., and Yu, X.B. *A new DEA-based method for fully ranking all decision-making units,* Expert Syst. 27(5) (2010) 363–373.

[15] Ekiz, M.K. and Tuncer Şakar, C. *A new DEA approach to fully rank DMUs with an application to MBA programs,* Int. Trans. Oper. Res. 27(4) (2020) 1886–1910.

[16] Fallahnejad, R., Asadi Rahmati, S., and Moradipour, K. *An entropy based Shapley value for ranking in data envelopment analysis,* Iran. J. Optim. 14(1) (2022) 39–49.

[17] Ghaeminasab, F., Rostamy-Malkhalifeh, M., Hosseinzadeh Lotfi, F., Behzadi, M.H., and Navidi, H. *Equitable Resource Allocation Combining Coalitional Game and Data Envelopment Analysis,* J. Appl. Res. Ind. Eng. (2022).

[18] Ghiyasi, M. *Full ranking of efficient and inefficient DMUs with the same measure of efficiency in DEA,* Int. J. Bus. Perform. Supply Chain Model. 10(3) (2019) 236–252.

[19] Hinojosa, M.A., Lozano, S., Borrero, D.V., and Marmol, A.M. *Ranking efficient DMUs using cooperative game theory,* Expert Syst. Appl. 80 (2017) 273–283.

[20] Hosseinzadeh Lotfi, F., Noora, A.A., Jahanshahloo, G.R., and Reshadi, M. *One DEA ranking method based on applying aggregate units,* Expert Syst. Appl. 38 (2011) 13468–13471.

[21] Izadikhah, M. and Farzipoor Saen, R. *A new data envelopment analysis method for ranking decision making units: an application in industrial parks,* Expert Syst. 32 (5) (2015) 596–608.

[22] Jahanshahloo, G.R., Vieira Junior, H., Hosseinzadeh Lofti, F., and Akbarian, D. *A new DEA ranking system based on changing the reference,* Eur. J. Oper. Res. 181 (2007) 331–337.

[23] Khmelnitskaya, A., Selçuk, O., and Talman, D. *The average covering tree value for directed graph games,* J. Comb. Optim. 39 (2020) 315–333.

[24] Khodabakhshi, M. and Aryavash, K. *Ranking all units in data envelopment analysis,* Appl. Math. Lett. 25(12) (2012) 2066–2070.

[25] Kiaei, H. and Nasseri, S.H. *Allocation of Weights Using Simultaneous Optimization of Inputs and Outputs' Contribution in Cross-efficiency Evaluation of DEA*. Yugoslav J. Oper. Res. 28(4) (2018) 521–538.

[26] Li, D.L., and Shan, E. *The Myerson value for directed graph games,* Oper. Res. Lett. 48(2) (2020) 142–146.

[27] Li, Y., Xie, J., Wang, M., and Liang, L. *Super-efficiency evaluation using a common platform on a cooperative game,* Eur. J. Oper. Res. 255(3) (2016) 884–892.

[28] Li, F., Zhu, Q., and Liang, L. *Allocating a fixed cost based on a DEA-game cross efficiency approach,* Expert Syst. Appl. 96 (2018) 196–207.

[29] Liang, L., Wu, J., Cook, W.D., and Zhu, J. *Alternative secondary goals in DEA cross-efficiency evaluation,* Int. J. Prod. Econ. 113 (2008) 1025–1030.

[30] Liang, L., Wu, J., Cook, W.D., and Zhu, J. *The DEA game cross-efficiency model and its Nash equilibrium,* Oper. Res. 56 (5), (2008) 1278–1288.

[31] Liang, L., Wu, J., Cook, W.D., and Zhu, J. *The DEA game cross-efficiency model and its Nash equilibrium,* Oper. Res. 56(5) (2008) 1278–1288.

[32] Liu, P., Wang, L.F., and Chang, J. *A revised model of the neutral DEA model and its extension,* Math. Probl. Eng. Res. (2017) 2017.

[33] Lozano, S. *Bargaining approach for efficiency assessment and target setting with fixed-sum variables,* Omega, 114 (2023) 102728.

[34] Ma, X., Liu, Y., Wei, X., Li, Y., Zheng, M., Li, Y., Cheng, C., Wu, Y., Liu, Z., and Yu, Y. *Measurement and decomposition of energy efficiency of Northeast China–based on super efficiency DEA model and Malmquist index,* Environ. Sci. Pollut. Res. 24(24) (2017) 19859–19873.

[35] Myerson, R.B. *Graphs and cooperation in games,* Math. Oper. Res. 2 (1977) 225–229.

[36] Nakabayashi, K. and Tone, K. *Egoist's dilemma: a DEA game,* Omega, 34(2) (2006) 135–148.

[37] Omrani, H., Beiragh, R.G., and Kaleibari, S.S. *Performance assessment of Iranian electricity distribution companies by an integrated cooperative game data envelopment analysis principal component analysis approach,* Int J. Electr. Power Energy Syst. 64 (2015) 617–625.

[38] Omrani, H., Fahimi, P., and Mahmoodi, A. *A data envelopment analysis game theory approach for constructing composite indicator: An application to find out development degree of cities in West Azarbaijan province of Iran,* Socio-Econ. Plan. Sci. 69 (2020) 100675.

[39] Ramón, N., Ruiz, J.L., and Sirvent, I. *Reducing differences between profiles of weights: A "peer-restricted" cross-efficiency evaluation,* Omega, 39(6) (2011) 634–641.

[40] Ramón, N., Ruiz, J.L., and Sirvent, I. *Cross-benchmarking for performance evaluation: Looking across best practices of different peer groups using DEA,* Omega, 92 (2020) 102169.

[41] Rezaee, M.J., Izadbakhsh, H., and Yousefi, S. *An improvement approach based on DEA-game theory for comparison of operational and spatial efficiencies in urban transportation systems,* KSCE J. Civ. Eng., 20(4) (2016) 1526–1531.

[42] Rezaeiani, M.J., and Foroughi, A.A. *Ranking efficient decision making units in data envelopment analysis based on reference frontier share,* Eur. J. Oper. Res. 264(2) (2018) 665–674.

[43] Roshdi, I., Mehdiloozad, M., and Margaritis, D. *A linear programming based approach for determining maximal closest reference set in DEA,* arXiv preprint arXiv:1407.2592 (2014).

[44] Sexton, T.R., Silkman, R.H., and Hogan, A.J. *Data envelopment analysis: Critique and extensions,* New Dir. Program Eval. 32 (1986) 73–105.

[45] Sojoodi, S., Dastmalchi, L., and Neshat, H. *Efficiency ranking of different types of power plants in Iran using super efficiency method,* Energy, (2021) 121104.

[46] Soofizadeh, S. and Fallahnejad, R. *Evaluation of groups using cooperative game with fuzzy data envelopment analysis,* AIMS Math. 8(4) (2023) 8661–8679.

[47] Sun, J., Li, G., and Wang, Z. *Technology heterogeneity and efficiency of China's circular economic systems: A game meta-frontier DEA approach,* Resour. Conserv. Recycl. 146 (2019) 337–347.

[48] Wang, Y.M. and Chin, K.S. *Some alternative models for DEA cross-efficiency evaluation,* Int. J. Prod. Econ. 128 (2010) 332–338.

[49] Wang, Y. M., Chin, K.S., and Jiang, P. *Weight determination in the cross-efficiency evaluation,* Comput. Ind. Eng. 61(3) (2011) 497–502.

[50] Wang, Y.M., Chin, K.S., and Luo, Y. *Cross-efficiency evaluation based on ideal and anti-ideal decision making units,* Expert Syst. Appl. 38(8) (2011) 10312–10319.

[51] Wu, J., Liang, L., Feng, Y., and Hong, Y. *Bargaining game model in the evaluation of decision making units,* Expert Syst. Appl. 36 (3) (2009) 4357–4362.

[52] Xie, Q., Zhang, L.L., Shang, H., Emrouznejad, A., and Li, Y. *Evaluating performance of super-efficiency models in ranking efficient decision-making units based on Monte Carlo simulations,* Ann. Oper. Res. 305(1) (2021) 273–323.

[53] Yu, M.M. and Rakshit, I. *Target setting for airlines incorporating CO2 emissions: The DEA bargaining approach,* J. Air Transp. Manag. 108 (2023) 102376.

[54] Zhu, Q., Song, M., and Wu, J. *Extended secondary goal approach for common equilibrium efficient frontier selection in DEA with fixed-sum outputs,* Comput. Ind. Eng. 144 (2020) 106483.

Iranian Journal of Numerical Analysis and Optimization
Vol. 14, No. 1, 2024, pp 77–106
https://doi.org/10.22067/ijnao.2023.82352.1259
https://ijnao.um.ac.ir/

CC BY NC

Research Article

Efficient numerical methods on modified graded mesh for singularly perturbed parabolic problem with time delay

A. Kumar and S. Gowrisankar*,

Abstract

In this article, we develop an efficient numerical method for one-dimensional time-delayed singularly perturbed parabolic problems. The proposed numerical approach comprises an upwind difference scheme with modified graded mesh in the spatial direction and a backward Euler scheme on uniform mesh in the temporal direction. In order to capture the local behavior of the solutions, stability and error estimations are obtained with respect to the maximum norm. The proposed numerical method converges uniformly with first-order up to logarithm in the spatial variable and also first-order in the temporal variable. Finally, the outcomes of the numerical experiments are included for two test problems to validate the theoretical findings.

*Corresponding author

Received 11 May 2023; revised 18 August 2023; accepted 19 August 2023

Ankesh Kumar
Department of Mathematics, National Institute of Technology Patna, Patna - 800005, India. e-mail: ankeshk.phd19.ma@nitp.ac.in

S. Gowrisankar
Department of Mathematics, National Institute of Technology Patna, Patna - 800005, India. e-mail: s.gowri@nitp.ac.in

How to cite this article
Kumar, A. and Gowrisankar, S., Efficient numerical methods on modified graded mesh for singularly perturbed parabolic problem with time delay. *Iran. J. Numer. Anal. Optim.*, 2024; 14(1): 77-106. https://doi.org/10.22067/ijnao.2023.82352.1259

AMS subject classifications (2020): 65M06, 65M15, 65M22.

Keywords: Singular perturbation problem; Finite difference methods; Modified graded mesh; Boundary layers; Uniform convergence.

1 Introduction

Singularly perturbed problems are widely applicable in different fields of science and engineering. In particular, singularly perturbed convection-diffusion parabolic equations, which contain a small parameter and a delay term, are effective in the simulation of reservoirs subsurface oil extraction, convective heat transfer problems with high péclet numbers, fluid flows, mathematical biology, and so on. Problems with delay have also appeared in the study of tumor growth, neural networks, and also in problems related to the respiratory system. The dynamics of the delay differential equations may become complicated due to time delay as it alters the stability equilibrium of the system. In epidemiology, to denote the incubation time or the host infection time, delay can be used. From the statistical study of ecological data, it is revealed that delay effects are seen in many species' population dynamics. An example of a delay partial differential equation (DPDE) that is used in a furnace to process metal sheets is as follows:

$$\frac{\partial \hat{w}(s,\tau)}{\partial \tau} - \varepsilon \frac{\partial^2 \hat{w}(s,\tau)}{\partial s^2} = v(f(\hat{w}(s,\tau-\xi)))\frac{\partial \hat{w}(s,\tau)}{\partial s} + c[g(\hat{w}(s,\tau-\xi)) - \hat{w}(s,\tau)].$$

For detailed literature on the above delay model, the reader can refer to [32]. DPDEs are closer to real-world phenomena as the solution does not only depend on the solution at a current stage but also at an earlier stage. The highest order derivative of the singularly perturbed delay differential equations (SPPDEs) is multiplied with a small parameter ε, with at least one term included with negative/positive shifts. There are a lot of manuscripts available in the literature dealing with DPDEs. For instance, see [14, 29, 22]. It is challenging to solve the SPPDEs analytically. Additionally, the presence of a small parameter ε leads to a sudden change in solution, resulting

in boundary and interior layers. Moreover, due to the requirements of an unexpectedly large number of mesh points on uniform mesh, a classical numerical method is not practical to accurately capture the layer in the solution. In this sense, the scheme mentioned above fails. So, a singularly perturbed DPDE numerical solution requires special treatment, which leads to the development of a uniform convergent method. Various articles discuss the numerical as well as the analytical aspects of singularly perturbed differential equations via finite difference and finite element methods. For example, see [2, 7, 19, 28, 24, 4, 23, 25]. However, only a few articles are available on partial differential equations with delay arguments in which analytical and numerical solutions are discussed. Ansari, Bakr, and Shishkin [1] used a finite difference method (FDM). Their method is implicit in time and centered in space on a piecewise uniform mesh. They achieved an order of convergence $\mathcal{O}((\ N^{-1}\ln N)^2 + M^{-1})$ for the singularly perturbed delay parabolic reaction-diffusion problem. Gowrisankar and Natesan [10] solved SPPDEs of a convection-diffusion type by virtue of the FDM. In this article, the spatial domain through the adaptive mesh and temporal variable using implicit Euler has been discretized, and the method converges with order $\mathcal{O}((N^{-1}\ln N) + \Delta t)$. Gowrisankar and Natesan [21] established uniform convergence for a singularly perturbed convection-diffusion problem with a convergence of order $\mathcal{O}(\Delta t + N^{-1+\bar{q}})$, where $0 < \bar{q} < 1$. Kumar and Kumari [15] constructed an ε-uniform convergent FDM for the reaction-diffusion initial-boundary value problem with a delay in the temporal direction. The authors have confirmed the order of convergence for a fully discretized scheme is $\mathcal{O}(\Delta t + N^{-2}\ln^3 N)$ utilizing extended cubic B-spline on the layer-adapted mesh. Kumar et al. [16] have explored uniform convergence of time-delayed parabolic partial differential equations (PDEs) on graded mesh generation algorithms based on entropy function. Das, Govindarao, and Mohapatra [3] proposed a weighted monotone numerical scheme comprised of Crank-Nicolson in time and a weighted monotone hybrid scheme for spatial derivatives on Shishkin mesh, which is parameter uniform converges with order $\mathcal{O}((N^{-1}\ln N)^2 + (\Delta t)^2)$. Gartland [8] studied boundary value problems on the graded mesh in which the scheme converges of order $\mathcal{O}(h^k)$, where k times as many points are available inside the layer as outside. The con-

vergence of order two in both the spatial and temporal variables has been established by Gupta, Kadalbajoo, and Dubey [11]. The difference scheme on the Bakhvalov–Shishkin mesh is used in [12] to explore the first-order convergence. Indeed in all these works, authors have described delay/nondelay problems on layer-adapted mesh only. There are various numerical methods used in different branches of science and engineering; for reference, see [26, 27, 5, 30, 31, 6, 33]. Any work related to the convergence of difference schemes on the modified graded mesh has not been noticed yet. Therefore, we are now in a position to establish a different scheme on the modified graded mesh. Motivated by the work of [1, 10, 4, 23], we proposed an upwind FDM on a modified graded mesh for a time-delayed convection-diffusion parabolic problem. We also establish that the method is almost first-order convergence without a logarithmic factor in the mesh parameter N.

Many articles discuss convection-diffusion problems involving boundary layers with layer adaptive meshes like Shishkin, Bekhvalov, and adaptive meshes. Gowrisankar and Natesan [10] solved SPPDEs of convection-diffusion type by the FDM on piecewise uniform mesh, which converges uniformly with an order $\mathcal{O}((N^{-1}\ln N) + \Delta t)$. Kumar et al. [17] analyzed convection-diffusion problems on the adaptive mesh via equidistribution of monitor function and obtained first-order convergence. Generating adaptive meshes for the parabolic problems requires the iterative process to obtain the grid at each time level. So, it requires high computation in comparison to other meshes discussed above. It is clear from the numerical findings that the modified graded mesh proposed in this work provides the equivalent rate of convergence as the adaptive meshes with negligible computation in comparison with the adaptive algorithm. Also, the modified graded mesh helps improve the accuracy and efficiency of the method.

The description of the problem is as follows: Let $\aleph = \Lambda \times (0, T]$, where $\Lambda = (0, 1)$ and $\Upsilon = \Upsilon_l \cup \Upsilon_b \cup \Upsilon_r$, in which Υ_r and Υ_l are, respectively, the right and left side of the rectangular domain $\aleph$ and $\Upsilon_b = [0, 1] \times [-\xi, 0]$. The following problem on $\aleph = \Lambda \times (0, T)$ is examined in this paper:

$$\begin{aligned} L_s^\tau \hat{w}(s,\tau) &= -\hat{d}(s,\tau)\hat{w}(s,\tau-\xi) + g(s,\tau), \quad \text{for all} \quad (s,\tau) \in \aleph, \qquad (1)\\ \hat{w}(s,\tau) &= \Psi_l(\tau), \quad (s,\tau) \in \{0\} \times \Upsilon_l = \{(0,\tau), 0 < \tau \le T\}, \end{aligned}$$

$$\hat{w}(s,\tau) = \Psi_r(\tau), \quad \hat{w}(s,\tau) \in \{1\} \times \Upsilon_r = \{(1,\tau), 0 < \tau \leq T\},$$
$$\hat{w}(s,\tau) = \Psi_b(s,\tau), \quad (s,\tau) \in \Upsilon_b = [0,1] \times [-\xi, 0],$$

where

$$L_s^\tau \hat{w}(s,\tau) := \Big(\frac{\partial}{\partial \tau} - \varepsilon \frac{\partial^2}{\partial s^2} + \hat{a}(s,\tau)\frac{\partial}{\partial s} + \hat{b}(s,\tau)\Big)\hat{w}(s,\tau). \tag{2}$$

The above-considered singular perturbation parameter is $\varepsilon \in (0,1]$. Additionally, $\xi \geq 0$, a delay parameter, satisfies the equation $T = q\xi$ for some positive integer q. Here, the above considered functions $\hat{a}(s,\tau)$, $\hat{b}(s,\tau)$, $\hat{d}(s,\tau)$, $g(s,\tau)$, Ψ_b, Ψ_r, Ψ_l, and Ψ_b are sufficiently smooth, bounded and also fulfill the following conditions:

$$\hat{a}(s,\tau) \geq \delta > 0, \quad \hat{b}(s,\tau) \geq 0, \quad \text{and} \quad \hat{d}(s,\tau) \geq \beta > 0.$$

The reduced problem corresponding to (1) is

$$\begin{cases} \frac{\partial \hat{w}_0(s,\tau)}{\partial \tau} + \hat{a}(s,\tau)\frac{\partial \hat{w}_0(s,\tau)}{\partial s} + \hat{b}(s,\tau)\hat{w}_0(s,\tau) & \\ \quad = -\hat{d}(x,\tau)\hat{w}_0(s,\tau-\xi) + g(s,\tau), & \text{for all } (s,\tau) \in \aleph, \\ \hat{w}_0(s,\tau) = \Psi_b(s,\tau), & (s,\tau) \in \Upsilon_b = [0,1] \times [-\xi, 0], \\ \hat{w}_0(s,\tau) = \Psi_r(s), & (s,\tau) \in \Upsilon_r. \end{cases} \tag{3}$$

By considering s to be constant, the solution to the reduced problem (3) is vertical lines, which ensures that the boundary layer that arises will be of the parabolic type. The compatibility condition for the initial functions $\Psi_b(s,\tau)$ are also satisfied at the corner points $(0,0)$, $(1,0)$, $(0,-\xi)$, and $(1,-\xi)$,

$$\begin{aligned} &\Psi_b(0,0) = \Psi_l(0), \\ &\Psi_b(1,0) = \Psi_r(0), \\ &\frac{d\Psi_l(0)}{d\tau} - \varepsilon\frac{\partial^2 \Psi_b(0,0)}{\partial s^2} + \hat{a}(0,0)\frac{\partial \Psi_b(0,0)}{\partial s} + \hat{b}(0,0)\Psi_b(0,0) \\ &\qquad = -\hat{d}(0,0)\Psi_b(0,-\xi) + g(0,0), \\ &\frac{d\Psi_r(0)}{d\tau} - \varepsilon\frac{\partial^2 \Psi_b(1,0)}{\partial s^2} + \hat{a}(1,0)\frac{\partial \Psi_b(1,0)}{\partial s} + \hat{b}(1,0)\Psi_b(1,0) \\ &\qquad = -\hat{d}(1,0)\Psi_b(1,-\xi) + g(1,0). \end{aligned}$$

Notation: Assume that C is used as a generic constant throughout the paper. Also, C is independent of the perturbation parameter ε and the mesh point N. We use discrete maximum norm to study convergence, which is defined as

$$\|\hat{w}\|_{\Lambda} = \max_{s\in\Lambda} |\hat{w}(s)|.$$

The Paper is presented as follows. Continuous solutions and their derivatives are covered in Section 2. The constructed numerical method is presented in Section 3. Section 4 studies the error analysis of the proposed method. Numerical examples and results support the theoretical findings in Section 5. The paper ends with a conclusion presented in Section 6.

2 Bounds for continuous solution and its derivatives

This section discusses the bounds for the solution and its partial derivatives across the prescribed domain. Furthermore, for the proof of ε-uniform convergence in a later section, we split the analytical solution $\hat{w}$ of (1). Also, it establishes strong bounds on the layer and smooth component, which we obtain by splitting the analytical solution. The operator L_s^{τ} defined in (2) satisfies the following maximum principle.

Lemma 1 (Maximum principle)**.** [15] Assume that $\xi(s,\tau) \geq 0$ holds true for all values of $(s,\tau) \in \Upsilon$. If $L_s^{\tau}\xi(s,\tau) \geq 0$ for all $(s,\tau) \in \aleph$, then $\xi(s,\tau) \geq 0$ for all $(s,\tau) \in \overline{\aleph}$.

Proof. Suppose $(s^*,\tau^*) \in \overline{\aleph}$ such that

$$\xi(s^*,\tau^*) = \min \xi(s,\tau),$$

and assume $\xi(s^*,\tau^*) < 0$. Also, we have $\frac{\partial \xi(s^*,\tau^*)}{\partial s} = 0$, $\frac{\partial \xi(s^*,\tau^*)}{\partial \tau} = 0$, and $\frac{\partial^2 \xi(s^*,\tau^*)}{\partial \tau^2} \geq 0$. Then

$$L_s^{\tau}\xi(s^*,\tau^*) < 0,$$

which is contrary to our supposition. It follows that $\xi(s^*,\tau^*) \geq 0$ and so $\xi(s,\tau) \geq 0$ for all $(s,\tau) \in \overline{\aleph}$. □

Lemma 2 (Uniform stability bounds)**.** [13, 1] For the specified function η in the defined domain of the differential operator L_s^{τ} in (2), we have

$$\|\eta\| \leq (1+\alpha T)\max\{\|L_s^\tau \eta\|, \|\eta_l\|, \|\eta_r\|, \|\eta_b\|\},$$

and for any solution $\hat{w}(s,\tau)$ of (1), we have

$$\|\hat{w}(s,\tau)\| \leq (1+\alpha T)\max\{\|g\|, \|\Psi_l\|, \|\Psi_r\|, \|\Psi_b\|\},$$

where $\alpha = \max\{0, 1-\delta\}$.

Proof. We define barrier function as

$$\psi^{\pm}(s,\tau) = (1+\alpha T)\max\{\|g\|, \|\Psi_l(\tau)\|, \|\Psi_r(\tau)\|, \|\Psi_b(\tau)\|\} \pm \hat{w}(s,\tau), \quad \text{for all } (s,\tau) \in \overline{\aleph}.$$

Then, at initial condition

$$\psi^{\pm}(s,0) = (1+\alpha T)\max\{\|g\|, \|\Psi_l(0)\|, \|\Psi_r(0)\|, \|\Psi_b(0)\|\} \pm \hat{w}(s,0),$$

we obtain

$$\psi^{\pm}(s,0) \geq 0, \quad \text{for all } \ (s,\tau) \in \Upsilon_b.$$

At boundary points, we have

$$\psi^{\pm}(0,\tau) = (1+\alpha T)\max\{\|g\|, \|\Psi_l(\tau)\|, \|\Psi_r(\tau)\|, \|\Psi_b(\tau)\|\} \pm \hat{w}(0,\tau),$$

which gives

$$\psi^{\pm}(0,\tau) \geq 0, \quad \text{for all } (s,\tau) \ \in \Upsilon_l.$$

Similarly, we have

$$\psi^{\pm}(1,\tau) \geq 0, \quad \text{for all } (s,\tau) \in \Upsilon_r.$$

The differential operator L_s^τ satisfies

$$\begin{aligned} L_s^\tau \psi^{\pm}(s,\tau) &= (\pm\hat{w}_\tau(s,\tau) - \varepsilon(\pm\hat{w}_{ss}(s,\tau)) + \hat{a}(s,\tau)(\pm\hat{w}_s(s,\tau)), \\ &\quad + \hat{b}(s,\tau)((1+\alpha T)\max\{\|g\|, \|\Psi_l(\tau)\|, \|\Psi_r(\tau)\|, \|\Psi_b(\tau)\|\} \pm \hat{w}(s,\tau))) \\ &= (\pm\hat{w}_\tau(s,\tau) \mp \varepsilon(\hat{w}_{ss}(s,\tau)) \pm \hat{a}(s,\tau)\hat{w}_s(s,\tau) \pm \hat{b}(s,\tau)\hat{w}(s,\tau), \\ &\quad + \hat{b}(s,\tau)((1+\alpha T)\max\{\|g\|, \|\Psi_l(\tau)\|, \|\Psi_r(\tau)\|, \|\Psi_b(\tau)\|\} \pm \hat{w}(s,\tau))) \\ &= L_s^\tau \hat{w}(s,\tau) + \hat{b}(s,\tau)((1+\alpha T)\max\|g\|, \|\Psi_l(\tau)\|, \|\Psi_r(\tau)\|, \|\Psi_b(\tau)\|) \\ &\geq 0. \end{aligned}$$

Thus, the maximum principle asserts that

$$\psi^{\pm}(s,\tau) \geq 0, \quad \text{for all } (s,\tau) \in \overline{\aleph}.$$

Finally, we get

$$\|\hat{w}(s,\tau)\| \leq (1+\alpha T)\max\{\|g\|, \|\Psi_l\|, \|\Psi_r\|, \|\Psi_b\|\}.$$

□

Theorem 1. Let the data $\hat{a} \in C^{(2+\gamma,1+\gamma/2)}(\overline{\Lambda})$, $\hat{b}$, $\hat{d}$, $g \in C^{(2+\gamma,1+\gamma/2)}$ $(\overline{\aleph})$, $\Psi_l \in C^{2+\gamma/2}([0,T])$, $\Psi_r \in C^{2+\gamma/2}([0,T])$, $\Psi_b \in C^{(4+\gamma,2+\gamma/2)}(\Upsilon_b)$, $\gamma \in (0,1)$, meet the necessary compatibility requirement on the corner. Consequently, (1) has a unique solution $\hat{w}$ and $\hat{w} \in C^{(4+\gamma,2+\gamma/2)}$ $(\overline{\aleph})$. Additionally, bounds on derivatives of solution $\hat{w}(s,\tau)$ is satisfied for all nonnegative integers i and j, where $0 \leq i+2j \leq 4$,

$$\left\| \frac{\partial^{i+j}\hat{w}}{\partial s^i \partial \tau^j} \right\| \leq C\varepsilon^{-i}.$$

Proof. The proof of existence and uniqueness can be found in [18]. To establish the derivative bound, the variable s is replaced to $\varepsilon\vartheta = s$ and follows the approach given in [1]. □

In the above Theorem 1, we establish the bound for exact solution $\hat{w}(s,\tau)$, which is not sufficient to establish the ε-uniform convergence of the method. In order to get strong bounds on the derivatives of the exact solution $\hat{w}$, we decompose the above $\hat{w}$ into smooth and layer component,

$$\hat{w}(s,\tau) = \hat{v}(s,\tau) + w(s,\tau), \quad (s,\tau) \in \bar{\aleph},$$

where the following differential equations are satisfied by the smooth component, $\hat{v}(s,\tau)$,

$$L_s^{\tau}\hat{v}(s,\tau) = -\hat{d}(s,\tau)\hat{v}(s,\tau-\xi) + g(s,\tau), \quad (s,\tau) \in \aleph,$$

with initial condition

$$\hat{v}(s,\tau) = \hat{w}(s,\tau), \quad (s,\tau) \in \Upsilon_b,$$

and the boundary condition

$$\hat{v}(0,\tau) = \hat{w}(0,\tau),\ \hat{v}(1,\tau) = \hat{w}(1,\tau), \quad 0 \leq \tau \leq T.$$

The smooth component $\hat{v}(s,\tau)$ is further decomposed as

$$\hat{v}(s,\tau) = \hat{v}_0(s,\tau) + \varepsilon\hat{v}_1(s,\tau), \qquad (s,\tau) \in \bar{\aleph},$$

where $\hat{v}_0(s,\tau)$ is the solution of reduced problem

$$\begin{aligned}
&\frac{\partial \hat{v}_0(s,\tau)}{\partial \tau} + \hat{a}(s,\tau)\frac{\partial \hat{v}_0(s,\tau)}{\partial s} + \hat{b}(s,\tau)\hat{v}_0(s,\tau) \\
&\qquad = -\hat{d}(s,\tau)\hat{v}_0(s,\tau-\xi) + g(s,\tau), \quad \text{for all } (s,\tau) \in \aleph, \\
&\hat{v}_0(s,\tau) = \Psi_b(s,\tau), \qquad (s,\tau) \in \Upsilon_b,
\end{aligned}$$

and $\hat{v}_1(s,\tau)$ satisfies the following problem,

$$\begin{aligned}
&\frac{\partial \hat{v}_1(s,\tau)}{\partial \tau} + \hat{a}(s,\tau)\frac{\partial \hat{v}_1(s,\tau)}{\partial s} + \hat{b}(s,\tau)\hat{v}_1(s,\tau) \\
&\qquad = -\hat{d}(s,\tau)\hat{v}_1(s,\tau-\xi) + \frac{\partial^2 \hat{v}_0(s,\tau)}{\partial s^2}, \quad (s,\tau) \in \aleph, \\
&\hat{v}_1(s,\tau) = 0, \quad (s,\tau) \in \Upsilon_b.
\end{aligned}$$

Furthermore, $\hat{v}$ satisfies

$$\begin{aligned}
&L_s^\tau \hat{v}(s,\tau) = -\hat{d}(s,\tau)\hat{v}(s,\tau-\tau) + g(s,\tau), \qquad (s,\tau) \in \aleph, \\
&\hat{v}(s,\tau) = \hat{w}(s,\tau), \qquad (s,\tau) \in \Upsilon_b, \\
&\hat{v}(0,\tau) = \hat{v}_0(0,\tau), \quad \hat{v}(1,\tau) = \hat{v}_0(1,\tau) + \varepsilon\hat{v}_1(1,\tau), \qquad \tau \in [0,T],
\end{aligned}$$

and the singular components w satisfies

$$\begin{aligned}
&L_s^\tau w(s,\tau) = -\hat{d}w(s,\tau-\xi), \qquad (s,\tau) \in \aleph, \\
&w(s,\tau) = 0, \qquad (s,\tau) \in \Upsilon_b, \\
&w(0,\tau) = 0, \qquad w(1,\tau) = \hat{w}(1,\tau) - \hat{v}(1,\tau), \qquad \tau \in [0,T],
\end{aligned}$$

and the estimation of singular and the smooth component are discussed in Theorem 2 given below.

Theorem 2. Assume that $\hat{a} \in C^{(4+\gamma,2+\gamma/2)}(\bar{\Lambda})$, $\hat{b}, \hat{d}, g \in C^{(4+\gamma,2+\gamma/2)}(\bar{\aleph})$, $\Psi_l \in C^{3+\gamma/2}([0,T])$, $\Psi_r \in C^{3+\gamma/2}([0,T])$, $\Psi_b \in C^{(6+\gamma,3+\gamma/2)}(\Upsilon_b)$, $\gamma \in (0,1)$, fulfill the necessary compatibility condition on the corner. Then, for all nonnegative integers i,j such that $0 \le i+2j \le 4$, we have the following bounds for the smooth $\hat{v}$ and the layer part w in the decomposition of solution $\hat{w}$ is

$$\left\| \frac{\partial^{i+j}\hat{v}}{\partial s^i \partial \tau^j} \right\| \leq C(1+\varepsilon^{1-i}),$$

and

$$\left\| \frac{\partial^{i+j} w}{\partial s^i \partial \tau^j} \right\| \leq C\varepsilon^{-i} e^{-\gamma s/\varepsilon}.$$

Proof. For the proof of theorem, one may refer [20]. □

3 Grid construction and numerical discretization

In this section, grids are developed to discretize the problem in both spatial and temporal directions. Later, we use a backward Euler in the temporal direction and an upwind method on the spatial derivative for the discretization of the problem (1).

3.1 Temporal discretization

To establish the convergence of (1) at each instance, we use the uniform time grid.

$$\Omega^M = \{\tau_k = k\Delta\tau, \quad k = 0,\ 1,\ \ldots,\ M, \quad \Delta\tau M = T\}.$$

Here M represents the grid points in the temporal direction.

3.2 Spatial discretization

We generate a modified graded mesh, Λ_s^N in the interval $[0,1]$ as follows:

$$\begin{cases} \mu_0 = 0, & \\ \mu_j = 2\varepsilon \frac{j}{N}, & 1 \leq j \leq \frac{N}{2}, \\ \mu_{j+1} = \mu_j(1+\rho h), & \frac{N}{2} \leq j \leq N-2, \\ \mu_N = 1, & \end{cases} \tag{4}$$

where the parameter h satisfies the following nonlinear equation:

$$\ln(1/\varepsilon) = (N/2)\ln(1+\rho h). \tag{5}$$

The above selection of the parameter h ensures that there are $N/2$ grid points in the interval $[\varepsilon, 1]$, which are distributed gradely in the interval $[\varepsilon, 1]$. Numerical verification stimulates that in comparison to (μ_{N-2}, μ_{N-1}), the interval $(\mu_{N-1}, 1)$ is not too small. In the subinterval $[0, \varepsilon]$, we distribute $N/2$ points with uniform step length $2\varepsilon/N$, while in the subinterval $[\varepsilon, 1]$, we first find h for some fix N by means of the nonlinear equation (5), and corresponding to that h, we distribute $N/2$ points in $[\varepsilon, 1]$.
The mesh length is denoted by $h_j = \mu_j - \mu_{j-1}$ for $j = 1, 2, \ldots, N$.

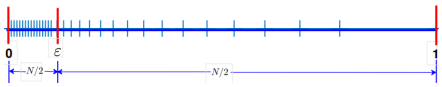


Figure 1: Modified graded mesh for $N = 32$ and $\varepsilon = 10^{-1}$.

Remark 1. The mesh size in piecewise uniform and the modified graded region is given by

$$h_j = \begin{cases} 2\varepsilon/N & \text{for } j = 1,\ 2,\ \ldots,\ N/2, \\ \rho h \mu_{j-1} & \text{for } j = N/2+1,\ N/2+2,\ \ldots,\ N. \end{cases}$$

Lemma 3. The mesh defined in (4) satisfies the following estimates:

$$|h_{j+1} - h_j| \leq \begin{cases} 0 & for\ j = 1,\ 2,\ \ldots,\ N/2, \\ Ch & for\ j = N/2+1,\ N/2+2,\ \ldots,\ N. \end{cases}$$

Proof. Initially, we consider $j = 1,\ 2,\ \ldots,\ N/2$. As the mesh is uniform in this portion, so nothing to prove.
For $j = N/2+1,\ N/2+2,\ \ldots,\ N$. We have

$$\begin{aligned} |h_{j+1} - h_j| &= |\rho h \mu_j - \rho h \mu_{j-1}| \\ &= \rho h |\mu_j - \mu_{j-1}| \\ &= \rho^2 h^2 \mu_{j-1} \\ &\leq Ch. \end{aligned}$$

Here, we have taken $0 < \rho, h < 1$. □

Lemma 4. For the modified graded mesh defined in (4), the parameter h satisfies the following bound,

$$h \leq CN^{-1}\ln(1/\varepsilon).$$

Proof. Let K_1 denote the number of partitioned points in (4) such that $\mu_j \leq \varepsilon$, for $j = 1,\ 2,\ \ldots,\ N/2$. Clearly, $K_1 \leq C/h$. Let K_2 be the number of points in the partition (4) such that $\mu_j > \varepsilon$. Let $\mu_{N/2+1}$ be the smallest point such that $\mu_j > \varepsilon$. We have to estimate the bound for K_2. Assuming $\rho h \leq 1$, we have

$$\begin{aligned}
K_2 &= \sum_{N/2+1}^{N} 1 = \sum_{N/2+1}^{N} (\mu_{j+1}-\mu_j)^{-1}\int_{\mu_j}^{\mu_{j+1}} d\mu \\
&= \sum_{N/2+1}^{N} (h_{j+1})^{-1}\int_{\mu_j}^{\mu_{j+1}} d\mu \\
&= \sum_{N/2+1}^{N} (h\rho\mu_j)^{-1}\int_{\mu_j}^{\mu_{j+1}} d\mu \\
&\leq \sum_{N/2+1}^{N} (2/\rho h\mu_{j+1})^{-1}\int_{\mu_j}^{\mu_{j+1}} d\mu,
\end{aligned}$$

because $\mu_{j+1} < 2\mu_j$. For any $\mu \in [\mu_j, \mu_{j+1}]$, we have

$$\begin{aligned}
K_2 &\leq \sum_{N/2+1}^{N} 2(\rho h)^{-1}\int_{\mu_j}^{\mu_{j+1}} \frac{1}{\mu} d\mu \\
&\leq 2(\rho h)^{-1}\int_{\varepsilon}^{1} \frac{1}{\mu} d\mu \\
&\leq 2(\rho h)^{-1}\ln(1/\varepsilon).
\end{aligned}$$

Recalling $N = K_1 + K_2$, we have

$$\begin{aligned}
N &\leq C/\rho h + 2(\rho h)^{-1}\ln(1/\varepsilon) \\
N &\leq 1/h(\rho C + 2\rho\ln(1/\varepsilon)) \\
N &\leq 1/h(C\ln(1/\varepsilon)).
\end{aligned}$$

Finally, we get

$$h \leq CN^{-1}\ln(1/\varepsilon),$$

where N represents grid points in the s-direction. □

Now, we systematically discretize our domain. For this, a non-uniform grid Λ_s^N on Λ with N mesh points is obtained by distributing $N/2$ points uniformly in the layer part and $N/2$ points nonuniformly outside the layer part. Here, we consider an equidistant grid Ω^M and Ω^l with uniform step-length $\Delta\tau$ on $[0,T]$ and $[-\xi,0]$ with M and l grid points, respectively. Then, the discretized domain will define as

$$\aleph^N = \Lambda_s^N \times \Omega^M, \quad \Upsilon_b^N = \Lambda_s^N \times \Omega^l,$$

and the boundary points Υ^N of $\aleph^N$ are $\Upsilon^N = \bar{\aleph}^N \cap \Upsilon$. Additionally, $\Upsilon_l^N = \aleph^N \cap \Upsilon_l$ and $\Upsilon_r^N = \aleph^N \cap \Upsilon_r$ define the left and right boundary points, respectively. Also, $\aleph_i^N = \Lambda_s^N \times \Omega_{i,\tau}^l$, where $\Omega_{i,\tau}^l$ is the set of uniform mesh in $[(i-1)\xi, i\xi]$. Now, we discretize our problem with the above-discretized mesh. We apply an upwind difference method and backward Euler for spatial and temporal derivatives. Then

$$\begin{cases} (D_\tau^- - \varepsilon\frac{D_s^+ - D_s^-}{\widehat{h_j}} + \hat{a}(j,k+1)\ D_s^+)S_{j,k+1} \\ \qquad +\hat{b}_{j,k+1}S_{j,k+1} = -\hat{d}(j,k+1)S_{j,k-l} + g_{j,k+1}, \\ S_{0,k+1} = \Psi_l(\tau_{k+1}), \\ S_{N,k+1} = \Psi_r(\tau_{k+1}), \\ S_{j,-p} = \Psi_b(s_j, -\tau_p), \qquad j = 1,\ 2,\ \ldots,\ N-1,\ p = 0,\ 1,\ \ldots,\ l. \end{cases} \tag{6}$$

By arranging (6), we will get a tri-diagonal system

$$(\alpha_{j,k+1})S_{j-1,k+1} + (\beta_{j,k+1})S_{j,k+1} + (\gamma_{j,k+1})S_{j+1,k+1} = h_{j,k}, \tag{7}$$

where

$$\begin{aligned} \alpha_{j,k+1} &= \frac{-\varepsilon\Delta\tau}{h_j\hat{h}_j}, \\ \beta_{j,k+1} &= 1 + \frac{\varepsilon\Delta\tau}{h_{j+1}\hat{h}_j} + \frac{\varepsilon\Delta\tau}{h_j\hat{h}_j} - \frac{\hat{a}(j,k+1)\Delta\tau}{h_{j+1}} + \hat{b}_{j,k+1}\Delta\tau, \\ \gamma_{j,k+1} &= \frac{-\varepsilon\Delta\tau}{h_{j+1}\hat{h}_j} + \frac{a(j,k+1)\Delta\tau}{h_{j+1}}, \\ h_{j,k} &= -\hat{d}(j,k+1)S_{j,k-l}\Delta\tau + g_{j,k+1}\Delta\tau + S_{j,k}, \end{aligned}$$

$$\hat{h}_j = \frac{h_{j+1} + h_j}{2},$$

and the step length $\Delta\tau$ satisfies $l\Delta\tau = \xi$, and also $\tau_p = p\Delta\tau,\ p \geq -l$, and l is a positive integer and the mesh function $\widetilde{v}(s_j, \tau_k) = \widetilde{v}_{j,k}$ is defined by

$$D_s^+ \widetilde{v}_{j,k} = \frac{\widetilde{v}_{j+1,k} - \widetilde{v}_{j,k}}{h_{j+1}}, \quad D_s^- \widetilde{v}_{j,k} = \frac{\widetilde{v}_{j,k} - \widetilde{v}_{j-1,k}}{h_j},$$

$$D_\tau^- \widetilde{v}_{j,k} = \frac{\widetilde{v}_{j,k} - \widetilde{v}_{j,k-1}}{h_{j+1}},$$

and

$$\delta_s^2 \widetilde{v}_{j,k} = \frac{(D_s^+ - D_s^-)\widetilde{v}_{j,k}}{\hat{h}_j}.$$

3.3 Numerical algorithm

The following algorithm provides the grid construction and the corresponding numerical solution:

Step 1. Given the number of mesh points in the temporal and the spatial direction, M and N, respectively, take uniform mesh points in the temporal direction, $\{\tau_k\}_{k=0}^{M}$.

Step 2. For the finer part in the spatial direction (i.e., $[0, \varepsilon]$), we have the uniform mesh $\{\mu_j\}_{j=0}^{N/2}$.

Step 3. For the coarser part (i.e., $[\varepsilon, 1]$), the graded mesh parameter h is obtained by solving the nonlinear equation (5) by the bisection method.

Step 4. Using the graded mesh parameter h, obtain the graded mesh in the interval $[\varepsilon, 1]$ from (4).

Step 5. Set $k = 1$.

Step 6. For the value of k, solve the tridiagonal system (7) to obtain the solution for the time level $t = k$.

Step 7. $k = k + 1$ goto Step 6.

Step 8. If $k = M$, then stop and mark $S_{j,k}$, as the required solution.

4 Error analysis

Lemma 5 (Discrete maximum principle)**.** [9] Assume that the mesh function $\xi(s_j,\tau_k)$ satisfies $\xi(s_j,\tau_k)\geq 0$ on Υ^N. If $L_s^{\tau,N}\xi(s_j,\tau_k)\geq 0$ on $(s_j,\tau_k)\in\aleph^N$, then $\xi(s_j,\tau_k)\geq 0$ on $\overline{\aleph}^N$.

Lemma 6. For any solution $S(s_j,\tau_k)$ of (6), we have

$$\|S(s_j,\tau_k)\|\leq(1+\alpha T)\max(\|L_s^{\tau,N}\|\|,\|\Psi\|_{\Upsilon^N}). \tag{8}$$

Proof. By constructing the barrier function, we have

$$\hat{S}^{\pm}(s_j,\tau_k)=(1+\alpha T)\max(\|L_s^{\tau,N}\|\|,\|\Psi\|_{\Upsilon^N})\pm S(s_j,\tau_k). \tag{9}$$

We can obtain (8) by applying (5) on (9). □

Theorem 3. Let $\hat{w}$ and S be the solution of continuous problem (1) and discretized problem (6), respectively. Furthermore, the both solutions meet the corners compatibility requirements. Then, the error estimate is given by

$$\max|(\hat{w}-S)(s_j,\tau_k)|\leq C[\Delta\tau+N^{-1}\ln(1/\varepsilon)],\quad (s_j,\tau_k)\in\aleph^N. \tag{10}$$

Here, constant C is free from N, $\Delta\tau$, and ε.

Proof. The proof is similar to the proof given in [10, 1]. We derive it by using our mesh briefed in sections 3.1 and 3.2 and (5).
To prove the theorem for different time levels, we first divide the domain $\aleph$ into $\aleph=\aleph_1\cup\aleph_2$, where $\aleph_1=\Lambda\times[0,\xi]$ and $\aleph_2=\Lambda\times[\xi,2\xi]$. The discretized domain $\aleph^N=\aleph_1^N\times\aleph_2^N$, where $\aleph_1^N=\Lambda_s^N\times\Omega^l$ and $\aleph_2^N=\Lambda_s^N\times\Omega^M$.
First, for $\tau\in[0,\xi]$, the right-hand side of (1) becomes $-\hat{d}(s,\tau)\hat{w}(s,\tau-\xi)+g(s,\tau)$, which is free from ε. Thus the results in [10] is applicable, and we obtain

$$\max|(\hat{w}-S)(s_j,\tau_k)|\leq C[\Delta\tau+N^{-1}\ln(1/\varepsilon)]. \tag{11}$$

For $\tau\geq\xi$, the term $\hat{w}(s,\tau-\xi)$ is not free from ε. So, we are required to examine the estimates between the numerical solution S and the analytical solution $\hat{w}$ over the interval $[\xi,2\xi]$. The following SPPDEs is considered:

$$\Big(\frac{\partial}{\partial\tau}-\varepsilon\frac{\partial^2}{\partial s^2}+\hat{a}(s,\tau)\frac{\partial}{\partial s}+\hat{b}(s,\tau)\Big)\hat{w}(s,\tau)$$

$$
\begin{aligned}
&= -\hat{d}(s,\tau)\hat{w}(s,\tau-\xi) + g(s,\tau), \quad (s,\tau) \in \aleph_2^N, \qquad (12)\\
\hat{w}(s,\tau) &= \hat{w}(s,\tau_l), \quad s \in (0,1),\\
\hat{w}(0,t) &= \Psi_0(\tau), \quad \hat{w}(1,\tau) = \Psi_1(\tau), \quad \tau \in [\xi, 2\xi].
\end{aligned}
$$

To determine the numerical solution, we discretize (12) by using upwind finite difference for spatial derivatives and the backward-Euler for time derivative,

$$
\begin{aligned}
L_s^\tau S(s_j,\tau_k) &\equiv D_\tau^- S_{j,k} - \varepsilon\delta_s^2 S_{j,k} + \hat{a}_{j,k} D_s^- Sj,k + \hat{b}_{j,k} S_{j,k}\\
&= -\hat{d}_{j,k} S_{j,k-l} + g(s_j,\tau_k), \quad (s_j,\tau_k) \in \aleph_2^N, \qquad (13)\\
S(s_j,\tau_k) &= S_1(s_j,\tau_k), \quad (s_j,\tau_k) \in \aleph^N,\\
S(0,\tau_k) &= \Psi_0(\tau_k),\\
S(1,\tau_k) &= \Psi_1(\tau_k), \quad \tau_k \in \Omega_{2,\tau}^l,
\end{aligned}
$$

where $S_1(s_j,\tau_k)$ is the approximate solution obtained over the interval $\aleph_1^N$. Now, we divide the solution $\hat{w}$ of (1) into its regular and singular components as $\hat{w} = \hat{y} + \hat{z}$, and furthermore, $\hat{y} = y_0 + \varepsilon y_1$, where y_0 is the solution to the reduced problem,

$$
\begin{aligned}
&\Big(\frac{\partial}{\partial\tau} + \hat{a}(s,\tau)\frac{\partial}{\partial s} + \hat{b}(s,\tau)\Big) y_0(s,\tau)\\
&\qquad = -\hat{d} y_0(s,\tau-\xi) + g(s,\tau), \quad (s,\tau) \in (0,1)\times(\xi,2\xi),\\
&y_0(s,\tau) = \hat{w}(s,\tau), \quad (s,\tau) \in (0,1)\times[0,\xi],\\
&y_0(0,\tau) = \hat{w}(0,\tau), \quad \tau \in [\xi,2\xi],
\end{aligned}
$$

and

$$
\begin{aligned}
L_s^\tau y_1(s,\tau) &= -\hat{d}(s,\tau) y_1(s,\tau-\xi) + \frac{\partial^2 y_0(s,\tau)}{\partial s^2},\\
y_1(s,\tau) &= 0, \quad (s,\tau) \in (0,1)\times[0,\xi],\\
y_1(0,\tau) &= y_1(1,\tau) = 0, \quad \tau \in [\xi,2\xi].
\end{aligned}
$$

Furthermore, $\hat{y}$ satisfies

$$
\begin{aligned}
L_s^\tau \hat{y}(s,\tau) &= -\hat{d}(s,\tau)\hat{y}(s,\tau-\xi) + g(s,\tau), \quad (s,\tau) \in (0,1)\times[\xi,\xi],\\
\hat{y}(s,\tau) &= \hat{w}(s,\tau), \quad (s,\tau) \in (0,1)\times[0,\xi],\\
\hat{y}(0,\tau) &= y_0(0,\tau), \quad \hat{y}(1,\tau) = y_0(1,\tau), \quad \tau \in [\xi,2\xi].
\end{aligned}
$$

The singular components z satisfies

$$\begin{aligned} L_s^\tau \hat{z}(s,\tau-\xi) &= -\hat{d}\hat{z}(s,\tau-\xi), \quad (s,\tau)\in(0,1)\times(\xi,2\xi),\\ \hat{z}(s,\tau) &= 0, \quad (s,\tau)\in(0,1)\times[0,\xi],\\ \hat{z}(1,\tau) &= 0, \quad \hat{z}(0,\tau) = \Psi_l(\tau) - y_0(0,\tau), \quad \tau\in[\xi,2\xi]. \end{aligned}$$

Now, we decompose the numerical solution S of (13) into $S = Y + Z$. Here, Z represents the singular component of the decomposition, and Y, the regular component, is the solution to the following non-homogeneous problem:

$$\begin{aligned} L_s^{\tau,N} Y &= -\hat{d}Y(s_j,\tau_{k-l}) + g(s_j,\tau_k), \quad (s_j,\tau_k)\in\aleph_2^N,\\ Y(s_j,\tau_k) &= S_1(s_j,\tau_k), \quad (s_j,\tau_k)\in(0,1)\times(0,\xi),\\ Y(0,\tau_k) &= \hat{y}(0,\tau_k), \quad Y(1,\tau_k) = \hat{y}(1,\tau_k), \quad \tau_k\in\Omega_{2,\tau}^l, \end{aligned}$$

and the singular component Z must satisfy

$$\begin{aligned} L_s^{\tau,N} Z &= -\hat{d}Z(s_j,\tau_{k-l}), \quad (s_j,\tau_k)\in\aleph_2^N,\\ Z(s_j,\tau_k) &= 0, \quad (s_j,\tau_k)\in\aleph_1^N,\\ Z(1,\tau_k) &= 0, \quad Z(0,\tau_k) = \Psi_l(\tau_k) - \hat{y}(0,\tau_k), \quad \tau_k\in\Omega_{2,\tau}^l. \end{aligned}$$

Therefore, the error can be written in the form

$$S - \hat{w} = (Y - \hat{y}) + (Z - \hat{z}).$$

Now, we will establish the bounds for the smooth and the layer components. To establish the bound for the smooth component, we use the classical argument. The smooth error component can be expressed as

$$\begin{aligned} L_s^{\tau,N}(Y-\hat{y}) &= -\hat{d}Y(s_j,\tau_{k-l}) + g(s_j,\tau_k) - L_s^{\tau,N}\hat{y}\\ &= -\hat{d}Y(s_j,\tau_{k-l}) + L_s^\tau\hat{y} + \hat{d}\hat{y}(s_j,\tau_{k-l}) - L_s^{\tau,N}\hat{y}\\ &= \hat{d}(\hat{y}(s_j,\tau_{k-l}) - Y(s_j,\tau_{k-l})) + L_s^{\tau,N}\hat{y}. \end{aligned}$$

Thus we obtain

$$\begin{aligned} L_s^{\tau,N}(Y-\hat{y}) = \hat{d}(\hat{y}(s_j,\tau_{k-l}) - Y(s_j,\tau_{k-l})) - \varepsilon(\frac{\partial^2}{\partial s^2} - \delta_s^2)\hat{y}\\ + (\frac{\partial}{\partial\tau} - \delta_\tau)\hat{y} + \hat{a}(s_j,\tau_k)(\frac{\partial}{\partial s} - \delta_s^-)\hat{y}. \end{aligned}$$

When absolute values are taken on both sides and used with (10), the above inequality becomes

$$|L_s^{\tau,N}(Y-\hat{y})| \le C(\Delta\tau + N^{-1}\ln(1/\varepsilon)) + \varepsilon|(\frac{\partial^2}{\partial s^2} - \delta_s^2)\hat{y}| + |(\frac{\partial}{\partial s} - \delta_s^-)\hat{y}| + |(\frac{\partial}{\partial \tau} - \delta_\tau)\hat{y}|.$$

Using the Taylor series expansion, we have

$$|L_s^{\tau,N}(Y-\hat{y})| \le C(\Delta\tau + N^{-1}\ln(1/\varepsilon)) + \frac{\varepsilon}{12}(s_{j+1}-s_{j-1})^2\left\|\frac{\partial^4\hat{y}}{\partial s^4}\right\| + (s_j - s_{j-1})\left\|\frac{\partial^2\hat{y}}{\partial s^2}\right\| + \frac{(\tau_k - \tau_{k-1})}{2}\left\|\frac{\partial^2\hat{y}}{\partial \tau^2}\right\|.$$

On using the estimates of derivatives, bounds of mesh length, and discrete maximum principle, we have

$$|(Y-\hat{y})(s_j,\tau_k)| \le C[\Delta\tau + N^{-1}\ln(1/\varepsilon)]. \tag{14}$$

Like continuous z, Z is discretized for estimating the singular component. Thus

$$\begin{aligned} L_s^{\tau,N}Z &= -\hat{d}Z(s_j,\tau_{k-l}), \quad (s_j,\tau_k) \in \aleph_2^N, \\ Z(s_j,\tau_k) &= 0, \quad (s_j,\tau_k) \in \aleph_2^N, \\ Z(1,\tau_k) &= 0, \quad Z(0,\tau_k) = \Psi_l(\tau_k) - \hat{y}(0,\tau_k), \quad \tau_k \in \Omega_{2,\tau}^l. \end{aligned}$$

The singular component error can be expressed as

$$\begin{aligned} L_s^{\tau,N}(Z-\hat{z}) &= L_s^{\tau,N}Z - L_s^{\tau,N}\hat{z} \\ &= -\hat{d}Z(s_i,\tau_{n-l}) - L_s^{\tau,N}\hat{z} \\ &= (L_s^{\tau} - L_s^{\tau,N})\hat{z} \\ &= -\varepsilon(\frac{\partial^2}{\partial s^2} - \delta_s^2)\hat{z} + (\frac{\partial}{\partial \tau} - \delta_\tau)\hat{z}. \end{aligned}$$

Then, the classical estimates gives

$$|L_s^{\tau,N}(Z-\hat{z})(s_j,\tau_k)| \le C[\Delta\tau + N^{-1}\ln(1/\varepsilon)].$$

The discrete maximum principle is satisfied by the operator $L_s^{\tau,N}$ and also, due to the uniform boundedness of the inverse operator, the above inequality

reduces to

$$|(Z-\hat{z})(s_j,\tau_k)| \le C[\Delta\tau + N^{-1}\ln(1/\varepsilon)]. \tag{15}$$

Combining (14) and (15) completes the proof on $[\xi, 2\xi]$. In the similar fashion, we estimate the error in the successive interval in time by using induction. □

5 The numerical result with examples

This section consists of two test examples with boundary layers to illustrate the numerical method discussed above. Using tables and graphs, we present the findings of numerical methods. All the results are performed by taking $\rho = 0.9$. Numerical results confirm our theoretical findings.

Example 1. Consider the following problem:

$$\begin{cases} \hat{w}_\tau(s,\tau) - \varepsilon\hat{w}_{ss}(s,\tau) - \hat{w}_s(s,\tau) & \\ \quad = \hat{w}(s,\tau-1) + g(s,\tau), & (s,\tau)\in(0,1)\times(0,2], \\ \hat{w}(s,0) = \hat{w}_0(s,\tau), & (s,\tau)\in(0,1)\times[-1,0], \\ \hat{w}(0,\tau) = \tau, & \hat{w}(1,\tau) = 0, \quad \tau\in[0,2]. \end{cases}$$

Using the exact solution, $\hat{w}(s,\tau) = \frac{e^{-s/\varepsilon}-e^{-1/\varepsilon}}{1-e^{-1/\varepsilon}}\tau + 2\tau s\cos((\pi s)/2)$, we will obtain $\hat{w}_0(s,\tau)$ and $g(s,\tau)$. Additionally, the maximum point-wise error for each ε is defined by

$$e_\varepsilon^{N,\Delta\tau} = \max|(\hat{w}-S)(s_j,\tau_k)| \quad (s_j,\tau_k)\in\aleph^N,$$

where $\hat{w}$ and S are exact and approximate solution, respectively. The order of convergence is computed by

$$p_\varepsilon^{N,\Delta\tau} = \frac{\log(e^{N,\Delta\tau}/e^{2N,\Delta\tau/2})}{\log 2}.$$

For Example 1, the computed maximum pointwise error in the layer region is shown in Table 1, and the associated convergence order is in Table 2. Additionally, in the smooth region, Table 3 shows the maximum pointwise error, and Table 4 presents the related order of convergence. The plots for the numerical solution of Example 1 for the parameters $N = 128$, $\varepsilon = 10^{-2}$

and $\varepsilon = 10^{-8}$ are shown in Figures 2a and 2b. At $s = 0$, the figure shows the existence of a boundary layer. Error plot of Example 1 for $N = 128$, $\varepsilon = 10^{-2}$, and $\varepsilon = 10^{-8}$ is presented in Figure 3a and 3b, respectively. Figure 6a is a plot of Example 1 on the log-log scale for maximum pointwise error. From Table 5 and Figure 6a, it is clear that the applied scheme is first-order uniformly convergent.

Table 1: Maximum error on the modified graded mesh in the layer region for Example 1

	Number of intervals $N/\Delta\tau$					
ε	$128/\frac{1}{10}$	$256/\frac{1}{20}$	$512/\frac{1}{40}$	$1024/\frac{1}{80}$	$2048/\frac{1}{160}$	$4096/\frac{1}{320}$
10^{-2}	2.2150E-01	1.0252E-01	4.8929E-02	2.3896E-02	1.1828E-02	5.8921E-03
10^{-4}	5.0063E-01	2.3021E-01	1.0696E-01	5.1049E-02	2.4870E-02	1.2266E-02
10^{-6}	7.6497E-01	3.6319E-01	1.6700E-01	7.8479E-02	3.7821E-02	1.8535E-02
10^{-8}	9.9761E-01	4.9961E-01	2.3016E-01	1.0699E-01	5.1072E-02	2.4881E-02

Table 2: Convergence rate for Example 1 using a modified graded mesh in the layer region

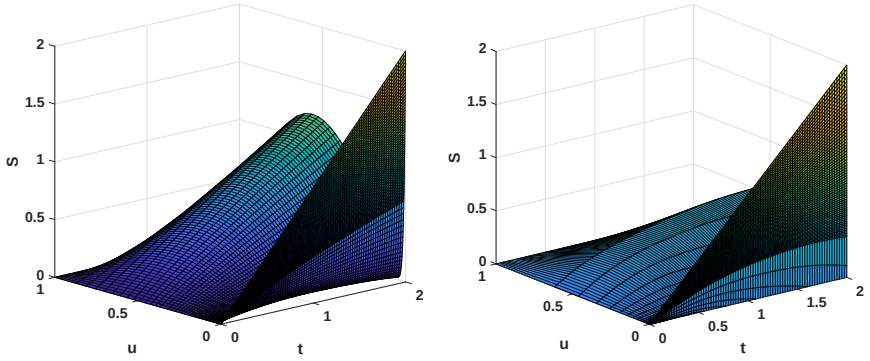
	Number of Intervals $N/\Delta\tau$				
ε	$128/\frac{1}{10}$	$256/\frac{1}{20}$	$512/\frac{1}{40}$	$1024/\frac{1}{80}$	$2048/\frac{1}{160}$
10^{-2}	1.1158	1.0731	1.0373	1.0158	1.0053
10^{-4}	1.1057	1.1067	1.0703	1.0394	1.0207
10^{-6}	1.0447	1.1156	1.0919	1.0556	1.0302
10^{-8}	0.9509	1.1049	1.1060	1.0697	1.0391

Table 3: Maximum error on the modified graded mesh in the smooth region for Example 1

	Number of Intervals $N/\Delta\tau$					
ε	$128/\frac{1}{10}$	$256/\frac{1}{20}$	$512/\frac{1}{40}$	$1024/\frac{1}{80}$	$2048/\frac{1}{160}$	$4096/\frac{1}{320}$
10^{-2}	2.2150E-01	1.0252E-01	4.8929E-02	2.3896E-02	1.1828E-02	5.8921E-03
10^{-4}	5.0063E-01	2.3021E-01	1.0696E-01	5.1049E-02	2.4870E-02	1.2266E-02
10^{-6}	7.6497E-01	3.6319E-01	1.6700E-01	7.8479E-02	3.7821E-02	1.8535E-02
10^{-8}	9.9761E-01	4.9961E-01	2.3016E-01	1.0699E-01	5.1072E-02	2.4881E-02

Table 4: Convergence rate for Example 1 using a modified graded mesh in the smooth region

	Number of Intervals $N/\Delta\tau$				
ε	$128/\frac{1}{10}$	$256/\frac{1}{20}$	$512/\frac{1}{40}$	$1024/\frac{1}{80}$	$2048/\frac{1}{160}$
10^{-2}	1.1113	1.0672	1.0339	1.0146	1.0053
10^{-4}	1.1208	1.1058	1.0671	1.0374	1.0198
10^{-6}	1.0747	1.1209	1.0894	1.0531	1.0289
10^{-8}	0.9977	1.1182	1.1051	1.0669	1.0375

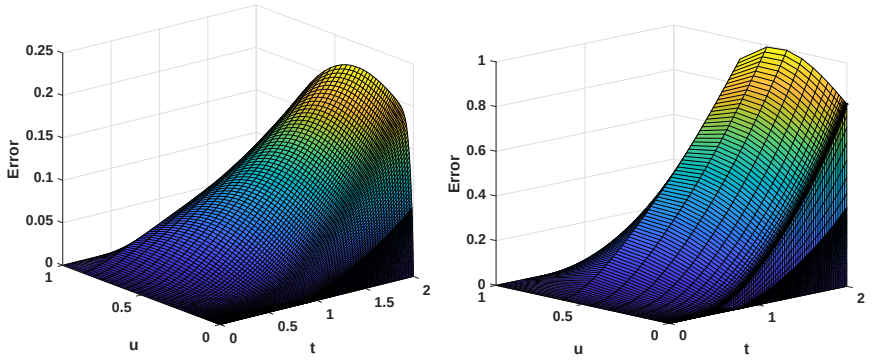


(a) $N = 128,\ \varepsilon = 10^{-2}$ (b) $N = 128,\ \varepsilon = 10^{-8}$

Figure 2: Solution profile for Example 1 for various values of N and ε

Table 5: Maximum error and rate of convergence on a modified graded mesh for Example 1

	Number of Intervals $N/\Delta\tau$					
ε	$128/\frac{1}{10}$	$256/\frac{1}{20}$	$512/\frac{1}{40}$	$1024/\frac{1}{80}$	$2048/\frac{1}{160}$	$4096/\frac{1}{320}$
10^{-2}	2.2150E-01	1.0252E-01	4.8929E-02	2.3896E-02	1.1828E-02	5.8921E-03
	1.1113	1.0672	1.0339	1.0146	1.0053	
10^{-4}	5.0063E-01	2.3021E-01	1.0696E-01	5.1049E-02	2.4870E-02	1.2266E-02
	1.1208	1.1058	1.0671	1.0374	1.0198	
10^{-6}	7.6497E-01	3.6319E-01	1.6700E-01	7.8479E-02	3.7821E-02	1.8535E-02
	1.0747	1.1209	1.0894	1.0531	1.0289	
10^{-8}	9.9761E-01	4.9961E-01	2.3016E-01	1.0699E-01	5.1072E-02	2.4881E-02
	0.9977	1.1182	1.1051	1.0669	1.0375	



(a) $N = 128,\ \varepsilon = 10^{-2}$ (b) $N = 128,\ \varepsilon = 10^{-8}$

Figure 3: Error profile for Example 1 for various values of N and ε

Example 2. Consider the following problem:

$$\begin{cases} \hat{w}_\tau(s,\tau) - \varepsilon \hat{w}_{ss}(s,\tau) - 20\hat{w}_s(s,\tau) \\ \quad = \hat{w}(s,\tau-1) + 10\tau^2 \exp(-\tau)s(1-s), & (s,\tau) \in (0,1)\times(0,2], \\ \hat{w}(s,\tau) = 0, & (s,\tau) \in (0,1)\times[-1,0], \\ \hat{w}(0,\tau) = 0, & \hat{w}(1,\tau) = 0, \quad \tau \in [0,2]. \end{cases}$$

In Example 2, we do not have an exact solution. To determine the maximum pointwise error and to establish an order of convergence, we employ the double mesh principle to the computed solution. For this, we consider $2M$ and $2N$ mesh interval in the temporal and spatial directions to determine the numerical solution $\widetilde{S}(s_j,\tau_k)$ on the mesh $\bar{\aleph}^{2N} = \bar{\Lambda}_s^{2N} \times \bar{\Omega}^{2M}$, for $j = 1,\ 2,\ \ldots,\ N$. Additionally, the jth mesh point of Λ_s^N and the $2j$th mesh point of $\bar{\Lambda}_s^{2N}$ are identical. Furthermore, for each ε, the maximum pointwise error is defined by

$$E_\varepsilon^{N,\Delta\tau} = \max |(S - \widetilde{S})(s_j,\tau_k)|, \quad (s_j,\tau_k) \in \aleph^N,$$

and the order of convergence is given by

$$\hat{p}_\varepsilon^{N,\Delta\tau} = \frac{\log(e^{N,\Delta\tau}/e^{2N,\Delta\tau/2})}{\log 2}.$$

In Example 2, the computed maximum pointwise error is shown in Table 6, and the associated convergence order is in Table 7. The plots for the numerical solution of Example 2 for the parameters $N = 128$, $\varepsilon = 10^{-2}$ and $\varepsilon = 10^{-8}$ are shown in Figure 4a and Figure 4b. The boundary layer at $s = 0$ can be seen in these figures. Error plot of Example 2 for $N = 128$, $\varepsilon = 10^{-2}$ and $\varepsilon = 10^{-8}$ is presented in Figure 5a and Figure 5b respectively. Figure 6b is a plot of Example 2 on the log-log scale for maximum pointwise error. Table 7 and Figure 6b show that the employed scheme is first-order uniformly convergent.

Table 6: Maximum errors for Example 2 on modified graded mesh

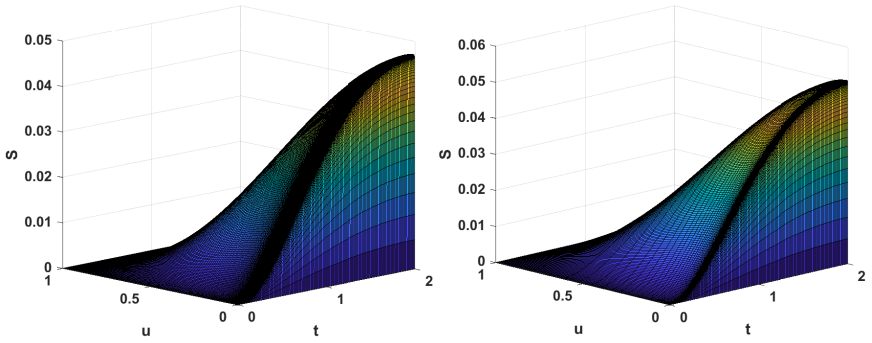
	Number of Intervals $N/\Delta\tau$					
ε	$128/\frac{1}{10}$	$256/\frac{1}{20}$	$512/\frac{1}{40}$	$1024/\frac{1}{80}$	$2048/\frac{1}{160}$	$4096/\frac{1}{320}$
10^{-2}	1.4063E-03	6.3978E-04	2.9302E-04	1.3810E-04	6.6717E-05	3.2746E-05
10^{-4}	2.8608E-03	1.4120E-03	6.4327E-04	2.9485E-04	1.3905E-04	6.7192E-05
10^{-6}	3.8686E-03	2.1730E-03	1.0218E-03	4.6404E-04	2.1513E-04	1.0258E-04
10^{-8}	4.3736E-03	2.8588E-03	1.4119E-03	6.4336E-04	2.9491E-04	1.3906E-04

Table 7: Rate of convergence for Example 2 on a modified graded mesh

	Number of Intervals $N/\Delta\tau$				
ε	$128/\frac{1}{10}$	$256/\frac{1}{20}$	$512/\frac{1}{40}$	$1024/\frac{1}{80}$	$2048/\frac{1}{160}$
10^{-2}	1.1362	1.1266	1.0853	1.0496	1.0267
10^{-4}	1.0187	1.1343	1.1254	1.0844	1.0493
10^{-6}	0.8321	1.0885	1.1389	1.1090	1.0685
10^{-8}	0.6134	1.0178	1.1340	1.1253	1.0846

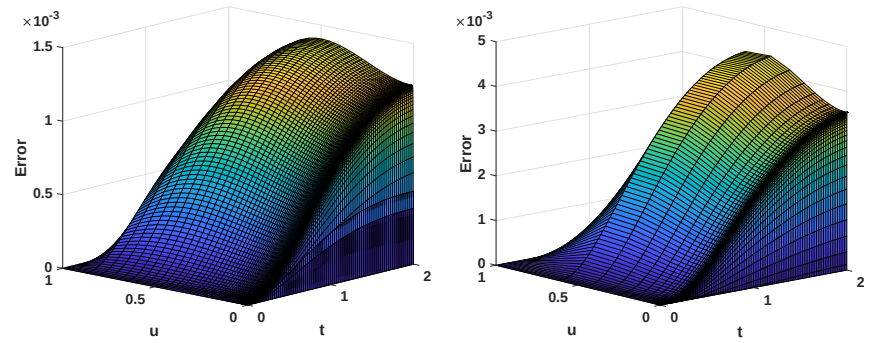
6 Conclusions

This article explored the upwind difference scheme on a modified graded mesh. Optimal error bounds were established in the maximum norm. The



(a) $N = 128,\ \varepsilon = 10^{-2}$ (b) $N = 128,\ \varepsilon = 10^{-8}$

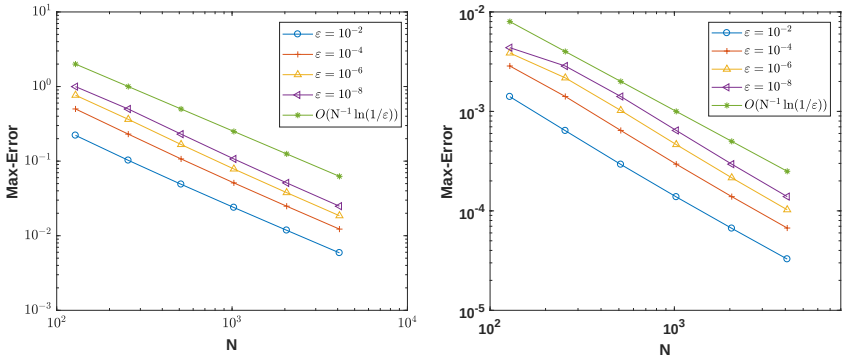
Figure 4: Solution profile for Example 2 for various values of N and ε



(a) $N = 128,\ \varepsilon = 10^{-2}$ (b) $N = 128,\ \varepsilon = 10^{-8}$

Figure 5: Error profile for Example 2 for various values of N and ε

method was proven to converge with first order up to logarithm. It can be observed from the numerical simulation that this logarithm term $\ln(1/\varepsilon)$ is not significant. Analytical estimations were carried out to gain uniform convergence results of computed solutions in the article. Two test problems were included to check the efficiency of the numerical method. Tables have been used to present the maximum pointwise error and order of convergence. Finally, numerical experiments confirmed the theoretical findings.



(a) Log-log plot for Example 1 (b) Log-log plot for Example 2

Figure 6: Log-log plot for Examples 1 and 2

References

[1] Ansari, A.R., Bakr, S.A. and Shishkin, G.I. *A parameter-robust finite difference method for singularly perturbed delay parabolic partial differential equations*, J. Comput. Appl. Math. 205(1) (2007), 552–566.

[2] Clavero, C., Jorge, J.C., Lisbona, F. and Shishkin, G.I.*An alternating direction scheme on a nonuniform mesh for reaction-diffusion parabolic problems*, IMA J. Numer. Anal. 20(2) (2000), 263–280.

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

[4] Durán, R.G. and Lombardi, A.L. *Finite element approximation of convection diffusion problems using graded meshes*, Appl. Numer. Math. 56(10-11) (2006), 1314–1325.

[5] Elboughdiri, N., Ghernaout, D., Muhammad, T., Alshehri, A., Sadat, R., Ali, M.R. and Wakif, A. *Towards a novel EMHD dissipative stagnation point flow model for radiating copper-based ethylene glycol nanoflu-*

ids: An unsteady two-dimensional homogeneous second-grade flow case study, Case Stud. Therm. Eng. 45 (2023), 102914.

[6] Elboughdiri, N., Reddy, C.S., Alshehri, A., Eldin, S.M., Muhammad, T. and Wakif, A. *A passive control approach for simulating thermally enhanced Jeffery nanofluid flows nearby a sucked impermeable surface subjected to buoyancy and Lorentz forces*, Case Stud. Therm. Eng. (2023), 103106.

[7] Farrell, P.A., Hegarty, A.F., Miller, J.J.H., O'Riordan, E. and Shishkin, G.I. *Robust computational techniques for boundary layers*, Applied Mathematics (Boca Raton), 16. Chapman & Hall/CRC, Boca Raton, FL, 2000.

[8] Gartland, E.C. Jr., *Graded-mesh difference schemes for singularly perturbed two-point boundary value problems*, Math. Comp. 51(184) (1988), 631–657.

[9] Gowrisankar, S. and Natesan, S. *A robust numerical scheme for singularly perturbed delay parabolic initial-boundary-value problems on equidistributed grids*, Electron. Trans. Numer. Anal. 41 (2014), 376–395.

[10] Gowrisankar, S. and Natesan, S. *ε-Uniformly convergent numerical scheme for singularly perturbed delay parabolic partial differential equations*, Int. J. Comput. Math. 94(5) (2017), 902–921.

[11] Gupta, V., Kadalbajoo, M.K. and Dubey, R.K. *A parameter-uniform higher order finite difference scheme for singularly perturbed time-dependent parabolic problem with two small parameters*, Int. J. Comput. Math. 96(3) (2019), 474–499.

[12] Jha, A. and Kadalbajoo, M.K. *A robust layer adapted difference method for singularly perturbed two-parameter parabolic problems*, Int. J. Comput. Math. 92(6) (2015), 1204–1221.

[13] Kabeto, M.J. and Duressa, G.F. *Robust numerical method for singularly perturbed semilinear parabolic differential difference equations*, Math. Comput. Simulation 188 (2021), 537–547.

[14] Kuang, Y. *Delay differential equations with applications in population dynamics*, Mathematics in Science and Engineering, 191. Academic Press, Inc., Boston, MA, 1993.

[15] Kumar, D. and Kumari, P. *A parameter-uniform scheme for singularly perturbed partial differential equations with a time lag*, Numer. Methods Partial Differ. Equ. 36(4) (2020), 868–886.

[16] Kumar, K., Podila, P.C., Das, P. and Ramos, H. *A graded mesh refinement approach for boundary layer originated singularly perturbed time-delayed parabolic convection diffusion problems*, Math. Methods Appl. Sci. 44(16) (2021), 12332–12350.

[17] Kumar, S., Sumit, Vigo-Aguiar, J. *A parameter-uniform grid equidistribution method for singularly perturbed degenerate parabolic convection-diffusion problems*, J. Comput. Appl. Math. 404 (2022), Paper No. 113273, 15.

[18] Ladyženskaja, O.A., Solonnikov, V.A. and Ural ceva, N.N. *Linear and quasi-linear equations of parabolic type*, volume 23 of Translations of Mathematical Monographs. American Mathematical Society, 1968.

[19] Miller, J.J., O'riordan, E. and Shishkin, G.I., *Fitted numerical methods for singular perturbation problems: error estimates in the maximum norm for linear problems in one and two dimensions*, World Scientific, 1996.

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

[21] Natesan, S. and Gowrisankar, S. *Robust numerical scheme for singularly perturbed parabolic initial-boundary-value problems on equidistributed Mesh*, CMES - Comput. Model. Eng. Sci. 88(4), (2012), 245–268.

[22] Nelson, P.W. and Perelson, A.S. *Mathematical analysis of delay differential equation models of HIV-1 infection*, Math. Biosci. 179(1) (2002), 73–94.

[23] Radojev, G. and Brdar, M. *A collocation method on a Gartland-type mesh for a singularly perturbed reaction-diffusion problem*, Math. Commun. 24(1) (2019), 19–37.

[24] Rajeev Ranjan, K. and Gowrisankar, S. *Uniformly convergent NIPG method for singularly perturbed convection diffusion problem on Shishkin type meshes*, Appl. Numer. Math. 179 (2022), 125–148.

[25] Rajeev Ranjan, K. and Gowrisankar, S. *NIPG method on Shishkin mesh for singularly perturbed convection-diffusion problem with discontinuous source term*, Int. J. Comput. Methods 20(2) (2023), Paper No. 2250048, 29 pp.

[26] Rasool, G., Wakif, A., Wang, X., Shafiq, A. and Chamkha, A.J., *Numerical passive control of alumina nanoparticles in purely aquatic medium featuring EMHD driven non-Darcian nanofluid flow over convective Riga surface*, Alex. Eng. J. 68 (2023), 747–762.

[27] Sharma, J., Ahammad, N.A., Wakif, A., Shah, N.A., Chung, J.D. and Weera, W. *Solutal effects on thermal sensitivity of casson nanofluids with comparative investigations on Newtonian (water) and non-Newtonian (blood) base liquids*, Alex. Eng. J. 71 (2023), 387–400.

[28] Shishkin, G.I. *Robust novel high-order accurate numerical methods for singularly perturbed convection-diffusion problems*, Math. Model. Anal. 10(4) (2005), 393–412.

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

[30] Wakif, A. *Numerical inspection of two-dimensional MHD mixed bioconvective flows of radiating Maxwell nanofluids nearby a convectively heated vertical surface*, Waves Random Complex Media, (2023), 1–22.

[31] Wakif, A., Abderrahmane, A., Guedri, K., Bouallegue, B., Kaewthongrach, R., Kaewmesri, P. and Jirawattanapanit, A., *Importance of exponentially falling variability in heat generation on chemically reactive*

von kármán nanofluid flows subjected to a radial magnetic field and controlled locally by zero mass flux and convective heating conditions: a differential quadrature analysis, Front. Phys. 10 (2022), 988275.

[32] Wu, J. *Theory and applications of partial functional-differential equations*, Applied Mathematical Sciences, 119. Springer-Verlag, New York, 1996.

[33] Zhang, K., Shah, N.A., Alshehri, M., Alkarni, S., Wakif, A. and Eldin, S.M. *Water thermal enhancement in a porous medium via a suspension of hybrid nanoparticles: MHD mixed convective Falkner's-Skan flow case study*, Case Stud. Therm. Eng. 47 (2023), 103062.

Iranian Journal of Numerical Analysis and Optimization
Vol. 14, No. 1, 2024, pp 107–135
https://doi.org/10.22067/ijnao.2023.83112.1288
https://ijnao.um.ac.ir/

CC BY NC

Research Article

Approximate proper solutions in vector optimization with variable ordering structure

S. Shahbeyk*,

Abstract

In this paper, we study approximate proper efficient (nondominated and minimal) solutions of vector optimization problems with variable ordering structures (VOSs). In vector optimization with VOS, the partial ordering cone depends on the elements of the image set. Approximate proper efficient/nondominated/ minimal solutions are defined in different senses (Henig, Benson, and Borwein) for problems with VOSs from new standpoints. The relationships among the introduced notions are studied, and some scalarization approaches are developed to characterize these solutions. These scalarization results based on new functionals defined by elements from the dual cones are given. Moreover, some existing results are addressed.

AMS subject classifications (2020): Primary 90C29; Secondary 90C30; 90C59.

*Corresponding author

Received 1 July 2023; revised 5 October 2023; accepted 9 October 2023

Shokouh Shahbeyk
Department of Mathematics, Faculty of Statistics, Mathematics, and Computer Science, Allameh Tabataba'i University , Tehran, Iran. e-mail: sh_shahbeyk@atu.ac.ir

How to cite this article
Shahbeyk, S., Approximate proper solutions in vector optimization with variable ordering structure. *Iran. J. Numer. Anal. Optim.*, 2024; 14(1): 107-135. https://doi.org/10.22067/ijnao.2023.83112.1288

Keywords: Approximate proper solutions; Variable ordering structure; Scalarization; Vector optimization.

1 Introduction

In vector optimization, elements in the objective space are compared by ordering (nontrivial, convex, closed, and pointed) cone. In recent years, the variable ordering structure (VOS) concept has been introduced [5, 10, 15, 44]. In vector optimization with VOS, the partial ordering cone depends on the elements of the image set. A candidate element is said to be nondominated optimal if and only if it is not dominated by other reference elements with respect to their corresponding ordering. Another notion of optimal is called minimality. For that notion, only the ordering of the candidate element itself is considered. Vector optimization with a VOS has recently gained more interest due to several applications in medical image registration, dynamical models, economic theory, behavioral sciences, and so on [2, 14, 18, 31, 42].

One of the most exciting and important notions in vector optimization is proper optimality. Properly optimal elements (solutions) are optimal with additional properties. This notion has been defined and investigated by various scholars; see, for example, Kuhn and Tucker [29], Geoffrion [20], Benson [4], Borwein [6], and Henig [25], among others. See also [22, 19, 35, 36, 45]. The introduction of a properly optimal solution in vector optimization with a VOS was first done by Eichfelder and Kasimbeyli [17]. The relationships between properly optimal solutions with a VOS in different senses, including Borwein, Benson, and Henig, were studied in [12, 13, 17, 16]. After that, Hartley properly optimal solutions, super optimal solutions, and robust solutions in vector optimization with a VOS were investigated [37].

On the other hand, in recent decades, there has been a lot of attention to approximate solutions to optimization problems due to two facts. Firstly, numerical algorithms may generate only approximate solutions, and secondly, the set of exact solutions might be empty in some practical problems, where approximate solutions exist. For Problems with a fixed order structure, the definition of approximate solutions has been extended by various scholars [3,

21, 23, 28, 30, 33, 40, 41, 42, 43]. Based on different ideas of proper efficiency, some researchers proposed concepts of approximate proper efficiency and investigated their characterizations and applications in vector optimization. For example, Maghri and Pareto-Fenchel [32] proposed approximate proper efficiency in the sense of Henig based on the ideas of Henig proper efficiency. Rong proposed the concept of Benson approximate proper efficiency based on the ideas of Benson proper efficiency [34]. Recently, approximate proper efficiency in an infinite-dimensional space was identified [26]. Soleimani and Tammer incorporated VOSs in approximate solutions of vector optimization problems [38, 39].

In the present paper, we introduce approximate properly optimal solutions in different senses (Henig, Benson, and Borwein) for problems with a VOS. Each of these proper solutions has its own advantages. Henig's notion provides easy-to-check criteria for proper optimality, and optimality has a close connection with stability. Benson and Borwein optimality notions deal with efficient solutions, which are efficient, respectively, for a closed cone generated by the transferred image space and the tangent cone of the transferred image space. Besides the well-featured approximate solutions, mentioned in the literature review, these concepts do not need to consider the compactness conditions to show the existence. The relationships between the introduced notions are studied. Furthermore, we give necessary and sufficient conditions for various approximate proper optimal notions.

The rest of the paper is organized as follows. In Section 2, we give some preliminary definitions and results. Section 3 is devoted to introducing proper approximate solutions with a VOS and investigating their relations. Section 4 continues the paper with necessary and sufficient conditions.

2 Preliminaries

Let X and Y be two real normed spaces. A set $C \subseteq Y$ is said to be cone if $\lambda C \subseteq C$ for each $\lambda \in [0, +\infty)$. A cone C is said to be convex if $C + C \subseteq C$, and it is pointed if $C \cap (-C) \subseteq \{0\}$. Furthermore, it is called nontrivial if $C \neq \{0\}$, and $C \neq Y$. A cone C is called an ordering cone if it is nontrivial, convex, closed, and pointed. A nonempty convex subset B of a cone C is

called a base of C if each $y \in C$ has a unique representation of the form $y = \lambda d$ for some $\lambda > 0$ and some $d \in B$.

For a set $A \subseteq X$, int$A$, cl$A$, and bd$A$ stand for the interior, the closure, and the boundary of A, respectively. Furthermore, $cone(A) := \bigcup_{\lambda \geq 0} \lambda A$ is the cone generated by A. The set A is called starshaped at $\bar{y} \in A$, if $\lambda y + (1-\lambda)\bar{y} \in A$ for every $y \in A$ and every $\lambda \in [0, 1]$.

Let $A \subseteq X$ and let $\bar{x} \in A$. The contingent cone or Bouligand tangent cone to A at $\bar{x}$, denoted by $T(A, \bar{x})$, is defined as the collection of all $v \in X$ such that there are sequences $\{x_j\} \subseteq A$ and $\{t_j\} \subseteq (0, +\infty)$ satisfying

$$x_j \to \bar{x} \quad \text{and} \quad t_j(x_j - \bar{x}) \to v \quad as \quad j \to \infty.$$

This cone is closed, while it is not convex necessarily. Consider a vector optimization problem,

$$\min\{f(x) : x \in \Omega\}, \tag{1}$$

where $f : X \to Y$ and $\emptyset \neq \Omega \subset X$.

In the following, we present one of the approximate solution concepts, which was first introduced by Kutateladze [30]. It is the most popular notion of approximate efficiency; see [24, 43] for more details. Let C be an ordering cone.

Definition 1. Let $\bar{x} \in \Omega$, let $\varepsilon > 0$, and let $k^0 \in C \setminus \{0\}$.
a) $\bar{x}$ is called an εk^0-efficient solution to (1) with respect to C if

$$\Big(f(\bar{x}) - \varepsilon k^0 - C \setminus \{0\}\Big) \cap f(\Omega) = \emptyset.$$

b) If int$C \neq \emptyset$, then $\bar{x}$ is called a weakly εk^0-efficient solution to (1) with respect to C if

$$\Big(f(\bar{x}) - \varepsilon k^0 - \text{int}C\Big) \cap f(\Omega) = \emptyset.$$

We denote the set of εk^0-efficient and weakly εk^0-efficient solutions of (1) with respect to C by $\varepsilon k^0 - E(\Omega, f, C)$, and $\varepsilon k^0 - wE(\Omega, f, C)$, respectively.

Now, we will study approximate solutions of vector optimization problems with a VOS.

We assume that $C : Y \rightrightarrows 2^Y$ is a given set-valued mapping that satisfies $0 \in bdC(y)$ and that $C(y)$ is an ordering cone for each $y \in Y$. Furthermore, assume that $k^0 \in Y$ is a nonzero vector satisfying $C(y) + [0, \infty)k^0 \subseteq C(y)$ for all $y \in Y$. Moreover, let $\varepsilon > 0$ be given.

The following definitions of minimality and nondomination for vector optimization problems with VOS, defined by a cone-valued mapping $C(\cdot)$, are known in the literature; see, for example, [7, 8, 9, 13, 15, 17, 44].

In order to introduce these notions, we consider the following two domination relations: For $y^1, y^2 \in Y$, we have

$$y^1 \leq_1 y^2 \quad if \quad y^2 \in y^1 + \big(C(y^1) \setminus \{0\}\big),$$
$$y^1 \leq_2 y^2 \quad if \quad y^2 \in y^1 + \big(C(y^2) \setminus \{0\}\big).$$

Definition 2. [38] Let $\bar{x} \in \Omega$.
a) $\bar{x}$ is called an εk^0-minimal solution to (1) with respect to the mapping $C(\cdot)$ if there is no feasible solution $x \in \Omega$ such that $f(x) + \varepsilon k^0 \leq_2 f(\bar{x})$, that is,

$$\Big(f(\bar{x}) - \varepsilon k^0 - \big(C(f(\bar{x})) \setminus \{0\}\big)\Big) \cap f(\Omega) = \emptyset.$$

b) If $\mathrm{int}C(f(\bar{x})) \neq \emptyset$, then $\bar{x}$ is called a weakly εk^0-minimal solution to (1) with respect to the mapping $C(\cdot)$ if

$$\Big(f(\bar{x}) - \varepsilon k^0 - \mathrm{int}C(f(\bar{x}))\Big) \cap f(\Omega) = \emptyset.$$

We denote the set of εk^0-minimal and weakly εk^0-minimal solutions of (1) with respect to the mapping $C(\cdot)$ by $\varepsilon k^0 - M(\Omega, f, C)$ and $\varepsilon k^0 - wM(\Omega, f, C)$, respectively.

Definition 3. [38] Let $\bar{x} \in \Omega$.
a) $\bar{x}$ is called an εk^0-nondominated solution to (1) with respect to the mapping $C(\cdot)$ if

$$\Big(f(\bar{x}) - \varepsilon k^0 - \big(C(f(x)) \setminus \{0\}\big)\Big) \cap \{f(x)\} = \emptyset, \quad \text{for all } \ x \in \Omega.$$

b) Assuming $\mathrm{int}C(f(x)) \neq \emptyset$ for all $x \in \Omega$, the vector $\bar{x}$ is called a weakly εk^0-nondominated solution to (1) with respect to the mapping $C(\cdot)$ if

$$\Big(f(\bar{x}) - \varepsilon k^0 - \text{int}C(f(x))\Big) \cap \{f(x)\} = \emptyset, \quad \text{for all } \ x \in \Omega.$$

We denote the set of εk^0-nondominated and weakly εk^0-nondominated solutions of (1) with respect to the mapping $C(\cdot)$ by $\varepsilon k^0 - N(\Omega, f, C)$ and $\varepsilon k^0 - wN(\Omega, f, C)$, respectively.

3 Proper approximate solutions

A fundamental solution concept in vector optimization, which plays a vital role from both theoretical and practical points of view, is the proper solution. This concept has been studied in many publications to eliminate the situations in which the trade-off between the criteria is unbounded. In a recent paper, Soleimani [38] investigated the concept of approximate solutions in vector optimization problems with VOSs. To our knowledge, no studies have been done on approximate proper solutions with VOSs in infinite-dimensional spaces.

We start the section by reviewing the definitions of proper approximate solutions with a fixed ordering structure.

Definition 4. [32, 34] Let $\bar{x} \in \Omega$.
a) $\bar{x}$ is called a properly εk^0-efficient solution to (1) with respect to C in the sense of Henig if there is a convex cone C' such that $C \setminus \{0\} \subseteq \text{int}C'$ and $\bar{x} \in \varepsilon k^0 - E(\Omega, f, C')$.
b) $\bar{x}$ is called a properly εk^0-efficient solution to (1) with respect to C in the sense of Benson if $f(\bar{x})$ is an εk^0-efficient element of the set

$$\{f(\bar{x}) - \varepsilon k^0\} + clcone\Big(f(\Omega) + C - \{f(\bar{x}) - \varepsilon k^0\}\Big)$$

with respect to C.
c) $\bar{x}$ is called a properly εk^0-efficient solution to (1) with respect to C in the sense of Borwein if $f(\bar{x})$ is an εk^0-efficient element of the set

$$\{f(\bar{x}) - \varepsilon k^0\} + T\Big(f(\Omega) + C, f(\bar{x}) - \varepsilon k^0\Big)$$

with respect to C.

We denote the set of properly εk^0-efficient solutions of (1) with respect to C in the sense of Henig, Benson, and Borwein by $\varepsilon k^0 - He(\Omega, f, C)$, $\varepsilon k^0 - Be(\Omega, f, C)$, and $\varepsilon k^0 - Bo(\Omega, f, C)$, respectively.

In the following, we introduce and study proper approximate solutions of vector optimization problems with a VOS in the sense of Henig, Benson, and Borwein. The notions introduced in Definitions 5 and 6 are extensions to the approximate case of the concepts of exact proper efficiency with VOSs given by Eichfelder and Kasimbeyli in [17, Definition 6].

Definition 5. Let $\bar{x} \in \Omega$.

a) $\bar{x}$ is called a properly εk^0-minimal solution to (1) with respect to the mapping $C(\cdot)$ in the sense of Henig if there is a set-valued mapping $C' : Y \rightrightarrows 2^Y$ with $C'(f(x))$ a convex cone and $C(f(x)) \setminus \{0\} \subseteq \text{int}C'(f(x))$ for all $x \in \Omega$ such that $\bar{x} \in \varepsilon k^0 - M(\Omega, f, C')$.

b) $\bar{x}$ is called a properly εk^0-minimal solution to (1) with respect to the mapping $C(\cdot)$ in the sense of Benson if $f(\bar{x})$ is an εk^0-minimal element of the set

$$A_1 := \{f(\bar{x}) - \varepsilon k^0\} + clcone\Big(f(\Omega) + C(f(\bar{x})) - \{f(\bar{x}) - \varepsilon k^0\}\Big) \tag{2}$$

with respect to the mapping $C(\cdot)$.

c) $\bar{x}$ is called a properly εk^0-minimal solution to (1) with respect to the mapping $C(\cdot)$ in the sense of Borwein if $f(\bar{x})$ is an εk^0-minimal element of the set

$$A_2 := \{f(\bar{x}) - \varepsilon k^0\} + T\Big(f(\Omega) + C(f(\bar{x})), f(\bar{x}) - \varepsilon k^0\Big) \tag{3}$$

with respect to the mapping $C(\cdot)$.

We denote the set of properly εk^0-minimal solutions of (1) with respect to the mapping $C(\cdot)$ in the sense of Henig, Benson, and Borwein by $\varepsilon k^0 - MHe(\Omega, f, C)$, $\varepsilon k^0 - MBe(\Omega, f, C)$, and $\varepsilon k^0 - MBo(\Omega, f, C)$, respectively.

Definition 6. Let $\bar{x} \in \Omega$.

a) $\bar{x}$ is called a properly εk^0-nondominated solution to (1) with respect to the mapping $C(\cdot)$ in the sense of Henig if there is a set-valued mapping $C' : Y \rightrightarrows 2^Y$ with $C'(f(x))$ a convex cone and $C(f(x)) \setminus \{0\} \subseteq \text{int}C'(f(x))$ for all $x \in \Omega$ such that $\bar{x} \in \varepsilon k^0 - N(\Omega, f, C')$.

b) $\bar{x}$ is called a properly εk^0-nondominated solution to (1) with respect to the mapping $C(\cdot)$ in the sense of Benson if $f(\bar{x})$ is an εk^0-nondominated element of the set

$$A_3 := \{f(\bar{x}) - \varepsilon k^0\} + clcone\Big(\bigcup_{\omega\in\Omega} \big(\{f(\omega)\} + C(f(\omega))\big) - \{f(\bar{x}) - \varepsilon k^0\}\Big) \tag{4}$$

with respect to the mapping $C(\cdot)$.

c) The element $\bar{x}$ is a properly εk^0-nondominated solution to (1) with respect to the mapping $C(\cdot)$ in the sense of Borwein if $f(\bar{x})$ is an εk^0-nondominated element of the set

$$A_4 := \{f(\bar{x}) - \varepsilon k^0\} + T\Big(\bigcup_{\omega\in\Omega} \big(\{f(\omega)\} + C(f(\omega))\big), f(\bar{x}) - \varepsilon k^0\Big) \tag{5}$$

with respect to the mapping $C(\cdot)$.

We denote the set of properly εk^0-nondominated solutions of (1) with respect to the mapping $C(\cdot)$ in the sense of Henig, Benson, and Borwein by $\varepsilon k^0 - NHe(\Omega, f, C)$, $\varepsilon k^0 - NBe(\Omega, f, C)$, and $\varepsilon k^0 - NBo(\Omega, f, C)$, respectively.

Lemma 1. It holds that $\bar{x}$ is a properly εk^0-minimal solution to (1) with respect to mapping $C(\cdot)$ in the sense of Henig if and only if there is a convex cone C with $C(f(\bar{x})) \setminus \{0\} \subseteq \text{int}C$ such that

$$\big(f(\bar{x}) - \varepsilon k^0 - C \setminus \{0\}\big) \cap f(\Omega) = \emptyset.$$

Proof. The proof of this lemma is similar to that of [17, Lemma 4] and is hence omitted. □

Lemma 2. [11] Let $P \subseteq Y$ be a weakly closed cone, and let C be a cone with a weakly compact base such that $P \cap C = \{0\}$. Then there exists an ordering cone C' such that $C \setminus \{0\} \subseteq \text{int}C'$ and $P \cap C' = \emptyset$.

Corollary 1. By the definitions and Lemma 1, $\bar{x}$ is a properly εk^0-minimal solution to (1) with respect to the mapping $C(\cdot)$ in the sense of Henig (resp., Benson/Borwein) if and only if it is a properly εk^0-efficient solution to (1) with respect to the ordering cone $C := C(f(\bar{x}))$ in the sense of Henig (resp., Benson/Borwein).

Corollary 2. It is not difficult to see that if $\bar{x}$ is a properly εk^0-efficient solution to (1) with respect to C in the sense of Henig (Benson/Borwein), then it is an εk^0-efficient solution of (1) with respect to C.

The following theorem shows that every properly εk^0-minimal (resp., εk^0-nondominated) solution is an εk^0-minimal (resp., εk^0-nondominated) solution with respect to the ordering mapping $C(\cdot)$.

Theorem 1. Let $\bar{x} \in \Omega$.

a) If $\bar{x} \in \varepsilon k^0 - MHe(\Omega, f, C)$ or $\bar{x} \in \varepsilon k^0 - MBe(\Omega, f, C)$ or $\bar{x} \in \varepsilon k^0 - MBo(\Omega, f, C)$ then $\bar{x} \in \varepsilon k^0 - M(\Omega, f, C)$

b) If $\bar{x} \in \varepsilon k^0 - NHe(\Omega, f, C)$ or $\bar{x} \in \varepsilon k^0 - NBe(\Omega, f, C)$ or $\bar{x} \in \varepsilon k^0 - NBo(\Omega, f, C)$, then $\bar{x} \in \varepsilon k^0 - N(\Omega, f, C)$

Proof. a) Let $C'(\cdot)$ be the mapping in definition of proper εk^0-minimality in the sense of Henig. We have $C(f(x)) \subseteq C'(f(x))$ for all $x \in \Omega$, $f(\Omega) \subseteq A_1$, where A_1 is as defined in (2) and $f(\Omega) \subseteq A_3$, where A_3 is as defined in (4). Hence, $\bar{x} \in \varepsilon k^0 - MHe(\Omega, f, C)$ or $\bar{x} \in \varepsilon k^0 - MBe(\Omega, f, C)$ implies $\bar{x} \in \varepsilon k^0 - M(\Omega, f, C)$.
To complete the proof of part (a), we have

$$\begin{aligned} & \bar{x} \in \varepsilon k^0 - MBo(\Omega, f, C) \\ \overset{Corollary\ 1}{\Longrightarrow} \quad & \bar{x} \in \varepsilon k^0 - Bo(\Omega, f, C(f(\bar{x}))), \\ \overset{Corollary\ 2}{\Longrightarrow} \quad & \bar{x} \in \varepsilon k^0 - E(\Omega, f, C(f(\bar{x}))), \\ \Longrightarrow \quad & \bar{x} \in \varepsilon k^0 - M(\Omega, f, C). \end{aligned}$$

b) The proof for Henig and Benson cases is similar to part (a). Now, let $\bar{x} \in \varepsilon k^0 - NBo(\Omega, f, C)$ and let $\bar{x} \notin \varepsilon k^0 - N(\Omega, f, C)$. Then there exist $\hat{y} \in f(\Omega) \subseteq \bigcup_{\omega\in\Omega} \Big(\{f(\omega)\} + C(f(\omega)) \Big)$ and $d \in C(\hat{y}) \setminus \{0\}$ such that $f(\bar{x}) - \varepsilon k^0 = \hat{y} + d$. Now define $d^n := (1 - \frac{1}{n})d \in C(\hat{y})$, $t_n := n$ and $y^n := \hat{y} + d^n \in \{\hat{y}\} + C(\hat{y}) \subseteq \bigcup_{\omega\in\Omega} \Big(\{f(\omega)\} + C(f(\omega)) \Big)$. Hence, we have

$$y^n = \hat{y} + d^n = \big(f(\bar{x}) - \varepsilon k^0 - \frac{1}{n} d\big) \to f(\bar{x}) - \varepsilon k^0,$$
$$t_n \Big(y^n - (f(\bar{x}) - \varepsilon k^0) \Big) = n\big(\hat{y} + d^n - f(\bar{x}) + \varepsilon k^0\big) = n(d^n - d) \to -d.$$

Therefore, we conclude

$$-d \in T\Big(\bigcup_{\omega\in\Omega} \big(\{f(\omega)\} + C(f(\omega))\big), f(\bar{x}) - \varepsilon k^0 \Big)$$

and

$$\hat{y} = f(\bar{x}) - \varepsilon k^0 - d \in \{f(\bar{x}) - \varepsilon k^0\} + T\Big(\bigcup_{\omega\in\Omega} \big(\{f(\omega)\} + C(f(\omega))\big), f(\bar{x}) - \varepsilon k^0 \Big).$$

This implies that $\bar{x}$ is not properly εk^0-nondominated solution to (1) with respect to C in the sense of Borwein, and the proof is completed.

□

It is shown in the following propositions that the set of approximate proper solutions does not grow as tolerance (ε) gets smaller.

Proposition 1. Let $0 \leq \varepsilon_1 \leq \varepsilon_2$. Then
a) $\varepsilon_1 k^0 - MBe(\Omega, f, C) \subseteq \varepsilon_2 k^0 - MBe(\Omega, f, C)$.
b) $\varepsilon_1 k^0 - MBo(\Omega, f, C) \subseteq \varepsilon_2 k^0 - MBo(\Omega, f, C)$.
c) $\varepsilon_1 k^0 - MHe(\Omega, f, C) \subseteq \varepsilon_2 k^0 - MHe(\Omega, f, C)$.

Proof. a)Let $\bar{x} \in \varepsilon_1 k^0 - MBe(\Omega, f, C)$. Then $\bar{x} \in \varepsilon_1 k^0 - M(A_1, C)$ in which A_1 is defined in (2). By [39, Theorem 2], $\bar{x} \in \varepsilon_2 k^0 - M(A_1, C)$, and therefore, $\bar{x} \in \varepsilon_2 k^0 - MBe(\Omega, f, C)$. Parts (b) and (c) can be proved similarly. □

The following proposition can be proved similar to Proposition 1.

Proposition 2. Let $0 \leq \varepsilon_1 \leq \varepsilon_2$. Then
a) $\varepsilon_1 k^0 - NBe(\Omega, f, C) \subseteq \varepsilon_2 k^0 - NBe(\Omega, f, C)$.

b) $\varepsilon_1 k^0 - NBo(\Omega, f, C) \subseteq \varepsilon_2 k^0 - NBo(\Omega, f, C)$.
c) $\varepsilon_1 k^0 - NHe(\Omega, f, C) \subseteq \varepsilon_2 k^0 - NHe(\Omega, f, C)$.

Proposition 3. a) If $C(f(x)) \subseteq C(f(\bar{x}))$ for all $x \in \Omega$, then every properly εk^0-minimal solution to (1) with respect to the mapping $C(\cdot)$ in the sense of Henig (resp., Benson/Borwein) is a properly εk^0-nondominated solution to (1) with respect to the mapping $C(\cdot)$ in the sense of Henig (resp., Benson/Borwein).
b) If $C(f(\bar{x})) \subseteq C(f(x))$ for all $x \in \Omega$, then every properly εk^0-nondominated solution to (1) with respect to the mapping $C(\cdot)$ in the sense of Henig (resp., Benson/Borwein) is a properly εk^0-minimal solution to (1) with respect to the mapping $C(\cdot)$ in the sense of Henig (resp., Benson/Borwein).

Proof. According to definitions of properly εk^0-minimal and nondominated solution and the inequality $\varepsilon k^0 - M(\Omega, f, C) \subseteq \varepsilon k^0 - N(\Omega, f, C)$ holds [39, Theorem 6], the proof is straightforward □

In the following, the relationships between the properly approximate solutions in different senses have been studied.

Theorem 2. Let $\bar{x} \in \Omega$. Then the following properties hold:
a)

$$\varepsilon k^0 - MBe(\Omega, f, C) \subseteq \varepsilon k^0 - MBo(\Omega, f, C)$$

$$\varepsilon k^0 - NBe(\Omega, f, C) \subseteq \varepsilon k^0 - NBo(\Omega, f, C).$$

b) If $f(\Omega) + C(f(\bar{x}))$ is starshaped with respect to $f(\bar{x})$, then

$$\varepsilon k^0 - MBo(\Omega, f, C) \subseteq \varepsilon k^0 - MBe(\Omega, f, C).$$

c) If $\bigcup_{\omega \in \Omega}(\{f(\omega)\} + C(f(\omega)))$ is starshaped with respect to $f(\bar{x})$, then

$$\varepsilon k^0 - NBo(\Omega, f, C) \subseteq \varepsilon k^0 - NBe(\Omega, f, C).$$

Proof. For a given set Λ and element $\bar{y} \in \Lambda$, we have $T(\Lambda, \bar{y}) \subseteq cl(cone(\Lambda - \{\bar{y}\})$, and the inclusion holds as equality when Λ is starshaped at $\bar{y}$. Therefore, parts (a), (b), and (c) are proved. □

Theorem 3. Let $\bar{x} \in \Omega$. Then the following properties hold:
a)

$$\varepsilon k^0 - MHe(\Omega, f, C) \subseteq \varepsilon k^0 - MBe(\Omega, f, C).$$

b) If $C(f(\bar{x}))$ has a weakly compact base, then

$$\varepsilon k^0 - MBe(\Omega, f, C) \subseteq \varepsilon k^0 - MHe(\Omega, f, C).$$

c) If $C(f(x))$ has a weakly compact base for all $x \in X$, then

$$\varepsilon k^0 - NBe(\Omega, f, C) \subseteq \varepsilon k^0 - NHe(\Omega, f, C).$$

Proof. a) Let $\bar{x} \in \varepsilon k^0 - MHe(\Omega, f, C)$. We have $\bar{x} \in \varepsilon k^0 - M(\Omega, f, C')$, where C' is a set-valued map $C' : Y \rightrightarrows 2^Y$ with $C'(f(x))$ an convex cone and $C(f(x)) \setminus \{0\} \subseteq \text{int}C'(f(x))$ for all $x \in X$. Therefore

$$\Big(f(\Omega) - \{f(\bar{x}) - \varepsilon k^0\}\Big) \cap \Big(- C'(f(\bar{x})) \setminus \{0\}\Big) = \emptyset.$$

Hence, by [25, Lemma 3.7],

$$\Big(f(\Omega) + C(f(\bar{x})) - \{f(\bar{x}) - \varepsilon k^0\}\Big) \cap \Big(- C'(f(\bar{x})) \setminus \{0\}\Big) = \emptyset.$$

Regarding to Definition of $cone(A)$ and since $C'(f(x))$ is an convex cone, we have

$$cone\Big(f(\Omega) + C(f(\bar{x})) - \{f(\bar{x}) - \varepsilon k^0\}\Big) \cap \Big(- C'(f(\bar{x})) \setminus \{0\}\Big) = \emptyset. \quad (6)$$

If $\bar{x} \notin \varepsilon k^0 - MBe(\Omega, f, C)$, then there exists $\hat{y} \in A_1$ (as defined in (2)) such that $\hat{y} = f(\bar{x}) - \varepsilon k^0 + d$ with $d \in -C(f(\bar{x})) \setminus \{0\}$. Therefore

$$d \in clcone\Big(f(\Omega) + C(f(\bar{x})) - \{f(\bar{x}) - \varepsilon k^0\}\Big) \cap -C(f(\bar{x})) \setminus \{0\}.$$

Thus

$$d \in clcone\Big(f(\Omega) + C(f(\bar{x})) - \{f(\bar{x}) - \varepsilon k^0\}\Big) \cap -\text{int}C'(f(\bar{x})).$$

Therefore, there exists a nonzero sequence $\{d_n\}$ such that

$$d_n \in cone\Big(f(\Omega) + C(f(\bar{x})) - \{f(\bar{x}) - \varepsilon k^0\}\Big) \cap -C'(f(\bar{x})),$$

with

$$\lim_{n\to+\infty} d_n = d.$$

This contradicts (6).

b) Let $\bar{x} \in \varepsilon k^0 - MBe(\Omega, f, C)$. By Corollary 1, $\bar{x} \in \varepsilon k^0 - Be(\Omega, f, C(f(\bar{x}))$. Thus

$$clcone\Big(f(\Omega) + C - \{f(\bar{x}) - \varepsilon k^0\}\Big) \cap -C(f(\bar{x})) = \{0\}.$$

By Lemma 2, there exists a closed convex pointed cone C' with $C(f(\bar{x})) \setminus \{0\} \subseteq \text{int}C'$ and $clcone\Big(f(\Omega) - (f(\bar{x}) - \varepsilon k^0)\Big) \cap -C' = \{0\}$. Hence,

$$\Big(f(\bar{x}) - \varepsilon k^0 - C' \setminus \{0\}\Big) \cap f(\Omega) = \emptyset.$$

This implies $\bar{x} \in \varepsilon k^0 - He(\Omega, f, C(f(\bar{x})))$ and by Corollary 1, $\bar{x} \in \varepsilon k^0 - MHe(\Omega, f, C)$.

c) Let $\bar{x} \in \varepsilon k^0 - NBe(\Omega, f, C)$ and let $\hat{x} \in \Omega$. Since $f(\Omega) \subseteq A_3$, by Definition 6, we have

$$f(\hat{x}) - (f(\bar{x}) - \varepsilon k^0) \notin -C(f(\hat{x})) \setminus \{0\}.$$

Thus

$$clcone\Big(f(\hat{x}) - (f(\bar{x}) - \varepsilon k^0)\Big) \cap -C(f(\hat{x})) = \{0\}.$$

By Lemma 2, there exists a closed convex pointed cone $C'(f(\hat{x}))$ with $C(f(\hat{x})) \setminus \{0\} \subseteq \text{int}C'(f(\hat{x}))$ and $clcone\big(f(\hat{x}) - (f(\bar{x}) - \varepsilon k^0)\big) \cap -C'(f(\hat{x})) = \{0\}$. Hence,

$$f(\hat{x}) - (f(\bar{x}) - \varepsilon k^0) \notin -C'(f(\hat{x}))$$

and

$$\big(f(\bar{x}) - \varepsilon k^0 - \big(C'(f(\hat{x})) \setminus \{0\}\big)\big) \cap \{f(\hat{x})\} = \emptyset.$$

Therefore, $\bar{x} \in \varepsilon k^0 - NHe(\Omega, f, C)$.

□

The following example shows that the converse of part (c) of Theorem 3 may not hold.

Example 1. Let $\Lambda = f(\Omega) = \{(y_1, y_2) \in \mathbb{R}^2 \mid 0 \leq y_1 \leq 1,\ 0 \leq y_2 \leq 1\}$. Let $\bar{y} = (\varepsilon, \varepsilon)$, $\varepsilon \in (0, 1)$, let $k^0 = (1, 1)$ and let

$$C(y) := \begin{cases} \mathbb{R}^2_+, & y \in \Lambda \setminus \{(1,0)\}, \\ cone(conv(\{(1,1),(1,-1)\})), & y = (1,0), \\ cone(conv(\{(1,1),(-1,1)\})), & y \in Y \setminus \Lambda. \end{cases}$$

If we define

$$C'(y) := \begin{cases} cone(conv(\{(1,-\delta),(-\delta,1)\})), & y \in \Lambda \setminus \{(1,0)\}, \\ cone(conv(\{(1,1+\delta),(1,-1-\delta)\})), & y = (1,0), \\ cone(conv(\{(1+\delta,1),(-1-\delta,1)\})), & y \in Y \setminus \Lambda, \end{cases}$$

for some small $\delta > 0$, then $C(y) \setminus \{0\} \subseteq \text{int}C'(y)$ for all $y \in Y$ and $\bar{y}$ is an εk^0-nondominated solution to (1) with respect to the mapping $C'(\cdot)$. Therefore $\bar{y} \in \varepsilon k^0 - NHe(\Omega, f, C)$.
Furthermore,

$$\hat{y} = (1, -1) \in A_3 = \{(y_1, y_2) \in \mathbb{R}^2 : y_1 + y_2 \geq 0,\ y_1 \geq 0\},$$

and

$$\big(\bar{y} - \varepsilon k^0 - \big(C(\hat{y}) \setminus \{0\}\big)\big) \cap \{\hat{y}\} \neq \emptyset.$$

Hence $\bar{y} \notin \varepsilon k^0 - NBe(\Omega, f, C)$.

4 Scalarization

In this section, some scalarization results are provided to characterize the properly εk^0 minimal, nondominated solution of (1). Let Y^* denote the topological dual space of Y, and it is equipped with the norm $\|\cdot\|_*$.
The dual cone of a set $C \subseteq Y$ is defined as

$$C^* := \{\zeta^* \in Y^* : \langle \zeta^*, c \rangle \geq 0, \quad \text{for all } c \in C\},$$

and the positive dual cone of C is defined as

$$C^{*^0} := \{\zeta^* \in Y^* : \langle \zeta^*, c\rangle > 0, \quad \text{for all } c \in C \setminus \{0\}\}.$$

Let a map $y^* : Y \to Y^*$ and an element $\bar{x} \in \Omega$ be given. We consider two functionals $\varphi^{\varepsilon}_{\bar{x}}, \psi^{\varepsilon}_{\bar{x}} : Y \to \mathbb{R}$ with

$$\begin{aligned} \varphi^{\varepsilon}_{\bar{x}}(y) &:= \langle y^*(f(\bar{x})), y - (f(\bar{x}) - \varepsilon k^0)\rangle \qquad for\ all\ \ y \in Y, \\ \psi^{\varepsilon}_{\bar{x}}(y) &:= \langle y^*(y), y - (f(\bar{x}) - \varepsilon k^0)\rangle \qquad for\ all\ \ y \in Y. \end{aligned}$$

Therefore,

$$\begin{aligned} \varphi^{\varepsilon}_{\bar{x}}(f(\bar{x}) - \varepsilon k^0) &= \langle y^*(f(\bar{x})), (f(\bar{x}) - \varepsilon k^0) - (f(\bar{x}) - \varepsilon k^0)\rangle = 0, \\ \psi^{\varepsilon}_{\bar{x}}(f(\bar{x}) - \varepsilon k^0) &= \langle y^*(f(\bar{x}) - \varepsilon k^0), (f(\bar{x}) - \varepsilon k^0) - (f(\bar{x}) - \varepsilon k^0)\rangle = 0. \end{aligned}$$

We use the following lemmas for some forthcoming proofs.

Lemma 3.

a) [27, Theorem 3.18] Let S be a nonempty closed convex subset of the normed space Y. Then $y \in Y \setminus S$ if and only if there are a continuous linear functional $l \in Y^* \setminus \{0\}$ and a real number α with

$$l(y) < \alpha \leq l(s) \quad for\ all\ s \in S.$$

b) [27, Theorem 3.22] Let the topology give Y as the topological dual space of Y^*. Moreover, let S and T be closed convex cones in Y with $\text{int}S^* \neq \emptyset$. Then $(-S) \cap (T) = \{0\}$ if and only if there is a continuous linear functional $l \in Y^* \setminus \{0\}$ with

$$l(x) \leq 0 \leq l(y) \quad for\ all\ x \in -S\ \ and\ \ y \in T$$

and

$$l(x) < 0 \quad for\ all\ x \in -S \setminus \{0\}.$$

Definition 7. Let Λ be a nonempty subset of a subset Y of a partially ordered linear space with an ordering cone C.
a) A functional $y^* : Y \to \mathbb{R}$ is called monotonically increasing on Λ, if for

every $\bar{y} \in \Lambda$

$$y \in (\bar{y} - C) \cap \Lambda \quad \Rightarrow \quad y^*(y) \leq y^*(\bar{y}).$$

b) A functional $y^* : Y \to \mathbb{R}$ is called strongly monotonically increasing on Λ, if for every $\bar{y} \in \Lambda$

$$y \in (\bar{y} - C) \cap \Lambda,\ y \neq \bar{y} \quad \Rightarrow \quad y^*(y) < y^*(\bar{y}).$$

Remark 1. Let Λ be any subset of a partially ordered linear space Y with the ordering cone C. Every linear functional $y^* \in C^*$ is monotonically increasing on Λ. Furthermore, every linear functional $y^* \in C^{*^0}$ is strongly monotonically increasing on Λ.

Lemma 4. Let Λ be a nonempty subset of a partially ordered linear space Y with a pointed cone C.
a) If there is a linear functional $y^* \in C^*$ and an element $\bar{y} \in \Lambda$ with

$$y^*(\bar{y} - \varepsilon k^0) < y^*(y) \quad for\ all\ \ y \in \Lambda \setminus \{\bar{y} - \varepsilon k^0\}, \tag{7}$$

then $\bar{y} \in \varepsilon k^0 - E(\Lambda, C)$.
b) If there are a linear functional $y^* \in C^{*^0}$ and an element $\bar{y} \in \Lambda$ with

$$y^*(\bar{y} - \varepsilon k^0) \leq y^*(y) \quad for\ all\ \ y \in \Lambda, \tag{8}$$

then $\bar{y} \in \varepsilon k^0 - E(\Lambda, C)$.

Proof. For the proof of both parts, we let $\bar{y} \notin \varepsilon k^0 - E(\Lambda, C)$. Then there exists $\hat{y} \neq \bar{y} - \varepsilon k^0$ such that

$$\hat{y} \in (\bar{y} - \varepsilon k^0 - C) \cap \Lambda.$$

In part (a), since $y^* \in C^*$, y^* is monotonically increasing, and hence $y^*(\hat{y}) \leq y^*(\bar{y} - \varepsilon k^0)$, which contradicts (7).
In the part (b), since $y^* \in C^{*^0}$, y^* is strongly monotonically increasing, and hence $y^*(\hat{y}) < y^*(\bar{y} - \varepsilon k^0)$, which contradicts (8). □

Example 2. Let $\Lambda = f(\Omega) = \{(y_1, y_2) \in \mathbb{R}^2 \mid 0 \leq y_1 \leq 1,\ 0 \leq y_2 \leq 1\}$. Let $\bar{y} = (\varepsilon, \varepsilon)$, let $\varepsilon \in (0, 1)$, let $k^0 = (1, 1)$, and let $C = \mathbb{R}^2_+$. It is easy to see that $\bar{y} \in \varepsilon k^0 - E(\Lambda, C)$. We define a map $y^* : \mathbb{R}^2 \to \mathbb{R}$ by $y^*(y) := (1, 1)^T y$.

Then for any $y \in \Lambda \setminus \{\bar{y} - \varepsilon k^0\} = \Lambda \setminus \{(0,0)\}$, we obtain $y^*(\bar{y} - \varepsilon k^0) = 0 < y^*(y) = y_1 + y_2$.

4.1 Characterizing approximate minimal elements

In this subsection, we provide some characterizations of approximate minimal elements. Theorem 4, given below, provides two characterizations of εk^0-minimal elements.

Theorem 4. Let $\bar{x} \in \Omega$.
a) Let $y^* : Y \to Y^*$ be a map such that $y^*(f(\bar{x})) \in C(f(\bar{x}))^*$. If

$$\varphi^{\varepsilon}_{\bar{x}}(y) > \varphi^{\varepsilon}_{\bar{x}}(f(\bar{x}) - \varepsilon k^0) = 0 \qquad for\ all\ \ y \in f(\Omega) \setminus \{f(\bar{x})\}, \tag{9}$$

then $\bar{x} \in \varepsilon k^0 - M(\Omega, f, C)$.
b)If $\bar{x} \in \varepsilon k^0 - M(\Omega, f, C)$, then there is a map $y^* : Y \to Y^*$ with $y^*(f(\bar{x})) \in C(f(\bar{x}))^*$ such that

$$\psi^{\varepsilon}_{\bar{x}}(y) > \psi^{\varepsilon}_{\bar{x}}(f(\bar{x}) - \varepsilon k^0) = 0 \qquad for\ all\ \ y \in f(\Omega) \setminus \{f(\bar{x})\}.$$

c) If $\mathrm{int}C(f(\bar{x}))^* \neq \emptyset$, then $\bar{x} \in \varepsilon k^0 - M(\Omega, f, C)$ if and only if there is a map $y^* : Y \to Y^*$ with $y^*(f(\bar{x})) \in C(f(\bar{x}))^{*^0} \setminus \{0_{Y^*}\}$ such that

$$\varphi^{\varepsilon}_{\bar{x}}(y) \geq \varphi^{\varepsilon}_{\bar{x}}(f(\bar{x}) - \varepsilon k^0) = 0 \qquad for\ all\ \ y \in f(\Omega) \setminus \{f(\bar{x})\}. \tag{10}$$

Proof. a) Let $y^* : Y \to Y^*$ be a map such that $y^*(f(\bar{x})) \in C(f(\bar{x}))^*$ and $\varphi^{\varepsilon}_{\bar{x}}(y) > 0$ for all $y \in f(\Omega) \setminus \{f(\bar{x})\}$. If $\bar{x} \notin \varepsilon k^0 - M(\Omega, f, C)$, then

$$\big(f(\bar{x}) - \varepsilon k^0 - \big(C(f(\bar{x})) \setminus \{0\}\big)\big) \cap f(\Omega) \neq \emptyset.$$

Therefore, there exists $\hat{y} \in f(\Omega)$ such that $f(\bar{x}) - \varepsilon k^0 - \hat{y} \in C(f(\bar{x})) \setminus \{0\}$. Since $y^*(f(\bar{x})) \in C(f(\bar{x}))^*$, then $\langle y^*(f(\bar{x})), f(\bar{x}) - \varepsilon k^0 - \hat{y}\rangle \geq 0$, and hence

$$\varphi^{\varepsilon}_{\bar{x}}(\hat{y}) = \langle y^*(f(\bar{x})), \hat{y} - (f(\bar{x}) - \varepsilon k^0)\rangle \leq 0,$$

and this is in contradiction with (9).
b) To prove (b), let $\bar{x} \in \varepsilon k^0 - M(\Omega, f, C)$ and $y \in f(\Omega) \setminus \{f(\bar{x})\}$ be arbitrarily

chosen. Then $f(\bar{x}) - \varepsilon k^0 - y \notin C(f(\bar{x}))$. Therefore by Lemma 3(a), there exist $y_1^*(y) \in Y^* \setminus \{0\}$ and $\alpha \in \mathbb{R}$ such that

$$\langle y_1^*(y), \big((f(\bar{x}) - \varepsilon k^0) - y\big)\rangle < \alpha \leq \langle y_1^*(y), d\rangle \qquad \text{for all } d \in C(f(\bar{x})).$$

We can show $\alpha = 0$. Hence by setting $y^*(y) = y_1^*(y)$ for all $y \in f(\Omega)\backslash\{f(\bar{x})\}\}$, we obtain a map $y^* : Y \to Y^*$ with $y^*(f(\bar{x})) \in C(f(\bar{x}))^*$ and $\psi_{\bar{x}}^{\varepsilon}(y) > \psi_{\bar{x}}^{\varepsilon}(f(\bar{x}) - \varepsilon k^0) = 0$ for all $y \in f(\Omega) \setminus \{f(\bar{x})\}$.
c) The argument of this part is similar to the first part.

□

Remark 2. By the definitions, $\bar{x} \in \varepsilon k^0 - M(\Omega, f, C)$ if and only if $\bar{x} \in \varepsilon k^0 - E(\Omega, f, C(f(\bar{x})))$.

Theorem 5 presents necessary and sufficient conditions for a given feasible solution to be Benson and Borwein properly εk^0 minimal solutions.

Theorem 5. Let $\bar{x} \in \Omega$.

a) If there exists a map $y^* : Y \to Y^*$ such that $y^*(f(\bar{x})) \in C^{*^0} = C(f(\bar{x}))^{*^0}$ and

$$\varphi_{\bar{x}}^{\varepsilon}(y) \geq 0 \qquad for\ all\ \ y \in f(\Omega) \setminus \{f(\bar{x})\}, \tag{11}$$

then $\bar{x} \in \varepsilon k^0 - MBe(\Omega, f, C)$ and $\bar{x} \in \varepsilon k^0 - MBo(\Omega, f, C)$.

b) If the topology gives Y as the topological dual space of Y^*, $\text{int}C(f(\bar{x}))^* \neq \emptyset$ and $f(\Omega) + C(f(\bar{x}))$ is convex, then $\bar{x} \in \varepsilon k^0 - MBe(\Omega, f, C) = \varepsilon k^0 - MBo(\Omega, f, C)$ if and only if there exists a map $y^* : Y \to Y^*$ with $y^*(f(\bar{x})) \in C(f(\bar{x}))^{*^0}$ such that (11) holds.

c) If the topology gives Y as the topological dual space of Y^*, $\text{int}C(f(\bar{x}))^* \neq \emptyset$ and $\bar{x} \in \varepsilon k^0 - MBe(\Omega, f, C)$, then there exists a map $y^* : Y \to Y^*$ with $y^*(y) \in C(f(\bar{x}))^{*^0}$ for all $y \in f(\Omega) \setminus \{f(\bar{x})\}$ such that

$$\psi_{\bar{x}}^{\varepsilon}(y) \geq 0 \qquad for\ all\ \ y \in f(\Omega). \tag{12}$$

Proof. a) From remark 2, it is sufficient to prove that $\bar{x} \in \varepsilon k^0 - Be(\Omega, f, C)$ and $\bar{x} \in \varepsilon k^0 - Bo(\Omega, f, C)$.
Let $y \in A_1$. Then $y = f(\bar{x}) - \varepsilon k^0 + d$ such that $d \in clcone\Big(f(\Omega) + C(f(\bar{x})) -$

$\{f(\bar{x}) - \varepsilon k^0\}\Big)$. Therefore there are sequences $\{y_n\} \subseteq f(\Omega) + C(f(\bar{x}))$ and $\{t_n\} \subseteq [0, +\infty)$ such that

$$d = \lim_{n\to\infty} t_n(y_n - f(\bar{x}) - \varepsilon k^0).$$

Then

$$\langle y^*(f(\bar{x})), d\rangle = \langle y^*(f(\bar{x})), \lim_{n\to\infty} t_n(y_n - f(\bar{x}) - \varepsilon k^0)\rangle.$$

Since the linear functional y^* is continuous and strongly monotonically increasing on $f(\Omega)$, it is strongly monotonically increasing on $f(\Omega) + C(f(\bar{x}))$, and therefore,

$$\langle y^*(f(\bar{x})), d\rangle) = \lim_{n\to\infty} t_n\big(\langle y^*(f(\bar{x})), y_n\rangle - \langle y^*(f(\bar{x})), (f(\bar{x}) - \varepsilon k^0)\rangle\big) \geq 0$$

and

$$\begin{aligned}\langle y^*(f(\bar{x})), y\rangle &= \langle y^*(f(\bar{x})), d\rangle + \langle y^*(f(\bar{x})), (f(\bar{x}) - \varepsilon k^0)\rangle \\ &\geq \langle y^*(f(\bar{x})), (f(\bar{x}) - \varepsilon k^0)\rangle.\end{aligned}$$

Hence, we obtain $\langle y^*(f(\bar{x})), y\rangle \geq \langle y^*(f(\bar{x})), (f(\bar{x}) - \varepsilon k^0)\rangle$ for all $y \in A_1$. Consequently, by Lemma 4, (b) $f(\bar{x}) \in \varepsilon k^0 - E(A_1, C)$. This completes the proof.

b) Sufficient condition follows immediately from (a) and Theorem 2. To prove the necessary condition, we assume that $\bar{x} \in \varepsilon k^0 - MBe(\Omega, f, C)$. Then $\bar{x}$ is εk^0-minimal element of set A_1 with respect to $C(\cdot)$. Therefore

$$\big(A_1 - \{f(\bar{x}) - \varepsilon k^0\}\big) \cap \big(-C(f(\bar{x}))\big) = \{0\}. \tag{13}$$

Since the set $f(\Omega) + C(f(\bar{x}))$ is convex, A_1 is convex and closed. Then, by Lemma 3(b), the set (13) is equivalent to a continuous linear functional $y_1^* \in Y^* \setminus \{0\}$ with

$$\langle y_1^*, -d\rangle \leq 0 \quad \text{for all } d \in C(f(\bar{x})),$$

$$\langle y_1^*, \hat{y}\rangle \geq 0 \quad \text{for all } \hat{y} \in A_1 - \{f(\bar{x}) - \varepsilon k^0\},$$

and

$$\langle y_1^*, d\rangle > 0 \quad \text{for all } d \in C(f(\bar{x})) \setminus \{0\}.$$

With these inequalities, we conclude

$$y_1^* \in C(f(\bar{x}))^{*^0},$$
$$\langle y_1^*, y - (f(\bar{x}) - \varepsilon k^0)\rangle \geq 0 \quad \text{for all } y \in A_1.$$

By setting $y^*(f(\bar{x})) := y_1^*$, we are done.

c) Let $\bar{x} \in \varepsilon k^0 - MBe(\Omega, f, C)$ and let $y = f(\bar{x}) - \varepsilon k^0 + d \in A_1$. Then

$$\big(f(\bar{x}) - \varepsilon k^0 - \big(C(f(\bar{x})) \setminus \{0\}\big)\big) \cap \{y\} = \emptyset,$$

and therefore,

$$cone\big(y - (f(\bar{x} - \varepsilon k^0)\big) \cap \big(- C(f(\bar{x}))\big) = \{0\}.$$

By Lemma 3(b), there exists $y_1^* \in Y^* \setminus \{0\}$ such that

$$\langle y_1^*, -d\rangle \leq 0 \quad \text{for all } d \in C(f(\bar{x})),$$

$$\langle y_1^*, \hat{y}\rangle \geq 0 \quad \text{for all } \hat{y} \in cone\big(y - (f(\bar{x} - \varepsilon k^0)\big)$$

and

$$\langle y_1^*, d\rangle > 0 \quad \text{for all } d \in C(f(\bar{x})) \setminus \{0\},$$

and then

$$y_1^* \in C(f(\bar{x}))^{*^0},$$
$$\langle y_1^*, y - (f(\bar{x}) - \varepsilon k^0)\rangle \geq 0.$$

By setting $y^*(y) := y_1^*$ for all $0 \neq y \in A_1$, we are done. □

Theorem 6 presents necessary and sufficient conditions for a given feasible solution to be Henig properly εk^0 minimal solutions with different assumptions.

Theorem 6. Let $\bar{x} \in \Omega$.

a) If $C(f(\bar{x}))$ has a weakly compact base, $y^* : Y \to Y^*$ is a map such that $y^*(f(\bar{x})) \in C(f(\bar{x}))^{*^0}$, and (11) holds, then $\bar{x} \in \varepsilon k^0 - MHe(\Omega, f, C)$.

b) Suppose that Y is a locally convex Hausdorff topological vector space and that there is a locally convex topology compatible with the dual pairing in Y^*. If there exists a map such that $y^*(f(\bar{x})) \in C(f(\bar{x}))^{*^0}$ and (11) holds, then $\bar{x} \in \varepsilon k^0 - MHe(\Omega, f, C)$.

c) If the topology gives Y as the topological dual space of Y^*, $\text{int}C(f(\bar{x}))^* \neq \emptyset$, $f(\Omega) + C(f(\bar{x}))$ is convex and $\bar{x} \in \varepsilon k^0 - MHe(\Omega, f, C)$, then there exists a map $y^* : Y \to Y^*$ with $y^*(f(\bar{x})) \in C(f(\bar{x}))^{*^0}$ such that (11) holds.

d) If $\bar{x} \in \varepsilon k^0 - MHe(\Omega, f, C)$, then there exists a map $y^* : Y \to Y^*$ with $y^*(y) \in C(f(\bar{x}))^{*^0}$ for all $y \in f(\Omega) \setminus \{f(\bar{x})\}$ such that

$$\psi^{\varepsilon}_{\bar{x}}(y) > 0 \qquad for\ all\ \ y \in f(\Omega) \setminus \{f(\bar{x})\}. \tag{14}$$

Proof. a) This part follows immediately from Theorems 5(a) and 3(b).

b) We consider the set-valued mapping C' from Y to 2^Y defined as follows: $C'(y) = Y$, if $y \neq f(\bar{x})$, and

$$C'(f(\bar{x})) := \{y \in Y \ \mid \ \langle y^*(f\bar{x})), y\rangle > 0\} \cup \{0\}.$$

Since $y^*(f(\bar{x})) \in C(f(\bar{x}))^{*^0}$, it is easy to see that $C'(f(x))$ is solid and convex and $C(f(x)) \setminus \{0\} \subset \text{int}C'(f(x))$ for all $x \in \Omega$. It follows that $\bar{x} \in \varepsilon k^0 - M(f, \Omega, C')$. Otherwise, there would exist $\hat{x} \in \Omega$ such that $f(\hat{x}) - f(\bar{x}) + \varepsilon k^0 \in -C'(f\bar{x})) \setminus \{0\}$ (note that from the definition of k^0 and C' it is easy to see that $f(\bar{x}) \neq f(\hat{x})$), which by the definition of C' means that

$$\varphi^{\varepsilon}_{\bar{x}}(y) = \langle y^*(f(\bar{x})), f(\hat{x}) - f(\bar{x}) + \varepsilon k^0\rangle < 0,$$

which contradicts (11). Then $\bar{x} \in \varepsilon k^0 - M(f, \Omega, C')$, and so $\bar{x} \in \varepsilon k^0 - MHe(f, \Omega, C)$.

c) Let $\bar{x} \in \varepsilon k^0 - MHe(\Omega, f, C)$. By Theorem 3(a), $\bar{x} \in \varepsilon k^0 - MBe(\Omega, f, C)$, and then by Theorem 5(b), there exists a map $y^* : Y \to Y^*$ with $y^*(f(\bar{x})) \in C(f(\bar{x}))^{*^0}$ such that (11) holds.

d) If $\bar{x} \in \varepsilon k^0 - MHe(\Omega, f, C)$, then by Lemma 1, there is a convex cone C' with $C(f(\bar{x})) \setminus \{0\} \subseteq \text{int}C'$ such that

$$\big(f(\bar{x}) - \varepsilon k^0 - C' \setminus \{0\}\big) \cap f(\Omega) = \emptyset.$$

We can assume the coneC' to be closed. Let $\hat{y} \in f(\Omega) \setminus \{f(\bar{x})\}$. Then

$$(f(\bar{x}) - \varepsilon k^0) - \hat{y} \notin C'.$$

Therefore by Lemma 3(a), there are a continuous linear functional $y_1^* \in Y^* \setminus \{0\}$ and a real number α with

$$\langle y_1^*, \big((f(\bar{x}) - \varepsilon k^0) - \hat{y}\big)\rangle < \alpha \leq \langle y_1^*, d\rangle \qquad for\ all\ d \in C'.$$

We can show $\alpha = 0$ and hence $y_1^* \in C'^* \setminus \{0\}$ and $\langle y_1^*, \hat{y} - (f(\bar{x}) - \varepsilon k^0)\rangle > 0$. By [27, Lemma 3.21], it holds

$$C(f(\bar{x})) \setminus \{0\} \subseteq \text{int}C' = \{y \in Y : y^*(y) > 0 \ \text{ for all } y^* \in C'^* \setminus \{0\}\},$$

and then $y_1^* \in C(f(\bar{x}))^{*^0}$. By setting $y^*(y) := y_1^*$ for all $y \in f(\Omega) \setminus \{f(\bar{x})\}$, we are done. □

4.2 Characterizing approximate nondominated elements

The characterization results of approximate nondominated elements are presented in the current subsection work. Theorem 7, given below, provides two characterizations of εk^0-nondominated elements.

Theorem 7. Let $\bar{x} \in \Omega$.
a) $\bar{x} \in \varepsilon k^0 - N(\Omega, f, C)$ if and only if there is a map $y^* : Y \to Y^*$ with $y^*(y) \in C(y)^* \setminus \{0\}$ for all $y \in f(\Omega) \setminus \{f(\bar{x}\})$ such that

$$\psi_{\bar{x}}^{\varepsilon}(y) > \psi_{\bar{x}}^{\varepsilon}(f(\bar{x}) - \varepsilon k^0) = 0 \qquad for\ all\ \ y \in f(\Omega) \setminus \{f(\bar{x})\}.$$

b) Suppose that $\text{int}C(y)^* \neq \emptyset$ for all $y \in f(\Omega)$. Then $\bar{x} \in \varepsilon k^0 - N(\Omega, f, C)$ if and only if there is a map $y^* : Y \to Y^*$ with $y^*(y) \in C(y)^{*^0} \setminus \{0\}$ for all $y \in f(\Omega) \setminus \{f(\bar{x}\})$ such that

$$\psi^{\varepsilon}_{\bar{x}}(y) \geq \psi^{\varepsilon}_{\bar{x}}(f(\bar{x}) - \varepsilon k^0) = 0 \qquad for\ all\ \ y \in f(\Omega) \setminus \{f(\bar{x})\}.$$

We omit proofs as the using same ideas in the proof of Theorem 4. Theorems 8 and 9 provide necessary and sufficient conditions for a given feasible solution to be Benson, Borwein, and Henig properly εk^0 nondominated solutions.

Theorem 8. Let $\bar{x} \in \Omega$, and let

$$\bar{C} := \bigcup_{y \in A_4} C(y).$$

a) If there exists a map $y^* : Y \to Y^*$ such that $y^*(y) \in \bar{C}^{*^0}$ for all $y \in f(\Omega) \setminus \{f(\bar{x})\}$ and (11) holds, then $\bar{x} \in \varepsilon k^0 - NBe(\Omega, f, C)$ and $\bar{x} \in \varepsilon k^0 - NBo(\Omega, f, C)$.

b) If $\bar{x} \in \varepsilon k^0 - NBe(\Omega, f, C)$, then there exists a map $y^* : Y \to Y^*$ with $y^*(y) \in C(y)^*$ for all $y \in f(\Omega) \setminus \{f(\bar{x})\}$ such that (14) holds.

c) If the topology gives Y as the topological dual space of Y^*, $\text{int}C(y)^* \neq \emptyset$ for all $y \in f(\Omega)$ and $\bar{x} \in \varepsilon k^0 - NBe(\Omega, f, C)$, then there exists a map $y^* : Y \to Y^*$ with $y^*(y) \in C(y)^{*^0}$ for all $y \in f(\Omega) \setminus \{f(\bar{x})\}$ such that (12) holds.

Proof. a) Let $y^*(y) \in \bar{C}^{*^0}$ and $\langle y^*(y), y - (f(\bar{x}) - \varepsilon k^0))\rangle \geq 0$ for all $y \in f(\Omega)$. Then $y^*(y) \in C(y)^{*^0}$ for all $y \in A_4$. Now, let $y \in A_4 \setminus \{f(\bar{x}) - \varepsilon k^0\}$, therefore there exist sequences $\{y_n\} \subseteq f(\Omega)$, $\{d_n\} \subseteq C(y_n)$, and $\{t_n\} \subseteq [0, \infty)$ such that $y = f(\bar{x}) - \varepsilon k^0 + d$, where

$$d = \lim_{n \to \infty} t_n(y_n + d_n - (f(\bar{x}) - \varepsilon k^0)).$$

Since the linear functional $y^*$is continuous, we have

$$\begin{aligned} \langle y^*(y), y - (f(\bar{x}) - \varepsilon k^0)\rangle &= y^*(d) \\ &= \lim_{n \to \infty} t_n\big(\langle y^*(y_n - (f(\bar{x}) - \varepsilon k^0))\rangle + \langle y^*(y), d_n\rangle\big) \\ &\geq 0. \end{aligned}$$

Hence, by Theorem 7(b) $\bar{x} \in \varepsilon k^0 - N(A_4, C)$. Thus $\bar{x} \in \varepsilon k^0 - NBe(\Omega, f, C)$, and then by Theorem 2, $\bar{x} \in \varepsilon k^0 - NBo(\Omega, f, C)$.

b) Let $\bar{x} \in \varepsilon k^0 - NBe(\Omega, f, C)$ and let $y = f(\bar{x}) - \varepsilon k^0 + d \in A_4$. Then

$$\big(f(\bar{x}) - \varepsilon k^0 - \big(C(y)\big) \setminus \{0\}\big)\big) \cap \{y\} = \emptyset,$$

and therefore,

$$(f(\bar{x}) - \varepsilon k^0) - y \notin C(y).$$

By Lemma 3(a), there are a continuous linear functional $y_1^* \in Y^* \setminus \{0\}$ and a real number α with

$$\langle y_1^*, \big((f(\bar{x}) - \varepsilon k^0) - y\big)\rangle < \alpha \leq \langle y_1^*, d\rangle \qquad \textit{for all } d \in C(y).$$

We can show $\alpha = 0$, and hence $y_1^* \in C(y)^* \setminus \{0\}$ and $\langle y_1^*, y - (f(\bar{x}) - \varepsilon k^0)\rangle > 0$. By setting $y^*(y) = y_1^*$ for all $0 \neq y \in A_4$, we are done.

c) Let $\bar{x} \in \varepsilon k^0 - NBe(\Omega, f, C)$ and let $y = f(\bar{x}) - \varepsilon k^0 + d \in A_4$. Then

$$\big(f(\bar{x}) - \varepsilon k^0 - \big(C(y)\big) \setminus \{0\}\big)\big) \cap \{y\} = \emptyset,$$

and therefore,

$$cone\big(y - (f(\bar{x} - \varepsilon k^0)\big) \cap \big(-C(f(y))\big) = \{0\}.$$

By Lemma 3(b), there exists $y_1^* \in Y^* \setminus \{0\}$ such that

$$\langle y_1^*, -d\rangle \leq 0 \quad \text{for all } d \in C(y),$$

$$\langle y_1^*, \hat{y}\rangle \geq 0 \quad \text{for all } \hat{y} \in cone\big(y - (f(\bar{x} - \varepsilon k^0)\big),$$

and

$$\langle y_1^*, d\rangle > 0 \quad \text{for all } d \in C(y) \setminus \{0\},$$

and then

$$y_1^* \in C(y)^{*^0},$$
$$\langle y_1^*, y - (f(\bar{x}) - \varepsilon k^0)\rangle \geq 0.$$

By setting $y^*(y) := y_1^*$ for all $0 \neq y \in A_4$, we are done. □

Theorem 9 provides necessary and sufficient conditions for a given feasible solution to be Henig properly εk^0 nondominated solutions.

Theorem 9. Let $\bar{x} \in \Omega$.

a) If there exists a map $y^* : Y \to Y^*$ such that $y^*(y) \in C(y)^{*^0}$ for all $y \in f(\Omega) \setminus \{f(\bar{x})\}$ such that (14) holds, then $\bar{x} \in \varepsilon k^0 - NHe(\Omega, f, C)$.

b) Suppose that $C(f(\bar{x}))$ has a weakly compact base for all $y \in f(\Omega) \setminus \{f(\bar{x})\}$. Then $\bar{x} \in \varepsilon k^0 - NHe(\Omega, f, C)$ if and only if there exists a map $y^* : Y \to Y^*$ with $y^*(y) \in C(y)^{*^0}$ such that (14) holds.

Proof. The proof is similar in spirit to Theorem 6. □

5 Conclusion

In this paper, some new approximate properly efficient solutions such as Henig, Benson, and Borwein minimal/nondominated solutions, in vector optimization with a VOS, were introduced, and theorems establishing the relationship between these concepts were proved. Necessary and sufficient conditions for these solutions were presented. We will consider these solutions in finite- dimensional spaces and provide an algorithm to generate these concepts in future work. It is recommended to study approximate Hartley properly and super solutions with a VOS.

References

[1] Adan, M. and Novo, V. *Proper efficiency in vector optimization on real linear spaces*, J. Optim. Theory Appl. 121 (2004), 515–540.

[2] Bao, T.Q., Mordukhovich, B.S. and Soubeyran, A. *Variational analysis in psychological modeling*, J. Optim. Theory Appl. 164 (2015), 290–315.

[3] Bazgan, C., Ruzika, S., Thielen, C. and Vanderpooten, D. *The power of the weighted sum scalarization for approximating multiobjective optimization problems* , Theory Comput. Syst. 66 (2022), 395–415.

[4] Benson, H.P. *An improved definition of proper efficiency for vector maximization with respect to cone*, J. Math. Appl. 71 (1979), 232–241.

[5] Bergstresser, K., Charnes, A. and Yu, PL. *Generalization of domination structures and nondominated solutions in multicriteria decision making*, J. Optim. Theory Appl., 18 (1976), 3–13.

[6] Borwein, J.M. *Proper efficient points for maximizations with respect to cones*, SIAM J. Control Optim. 15 (1977), 57–63.

[7] Chen, G.Y. *Existence of solutions for a vector variational inequality: an extension of the Hartmann Stampacchia theorem*, J. Optim. Theory Appl. 74 (1992), 445–456.

[8] Chen, G.Y. and Craven, B.D. *Existence and continuity of solutions for vector optimization*, J. Optim. Theory Appl. 81(3) (1994), 459–468.

[9] Chen, G.Y., Huang, X. and Yang, X. *Vector Optimization, Set-Valued and Variational Analysis*, Springer, Berlin, 2005.

[10] Chen, G.Y. and Yang, X.Q. *Characterizations of variable domination structures via nonlinear scalarization*, J. Optim. Theory Appl. 112 (2002), 97–110.

[11] Dauer, J.P. and Saleh, O.A. *A characterization of proper minimal points as solutions of sublinear optimizations problems*, J. Math. Anal. Appl. 178 (1993), 227–246.

[12] Durea, M., Florea E.A. and Strugriu, R. *Henig proper efficiency in vector optimization with variable ordering structure*, J. Ind. Manag. Optim. 15(2) (2019), 791–815.

[13] Eichfelder, G. *Optimal elements in vector optimization with a variable ordering structure*, J. Optim. Theory Appl. 151 (2011), 217–240.

[14] Eichfelder, G. *Vector optimization in medical engineering*, in Pardalos PM, Rassias TM (eds.) Mathematics Without Boundaries Springer, New York, (2014), 181–215.

[15] Eichfelder, G. *Variable Ordering Structures in Vector Optimization*, Springer, Berlin, 2014.

[16] Eichfelder, G. and Gerlach, T. *Characterization of properly optimal elements with variable ordering structures*, Optimization, 65 (2016), 571–588.

[17] Eichfelder, G., Kasimbeyli, R. *Properly optimal elements in vector optimization with variable ordering structure*, J. Global Optim. 60 (2014), 689–712.

[18] Engau, A. *Variable preference modeling with ideal-symmetric convex cones*, J. Global Optim. 42 (2008), 295–311.

[19] Foroutannia, D. and Mahmodinejad, A. *The concept of B-efficient solution in fair multiobjective optimization problems*, Iran. J. Numer. Anal. Optim. 7(1) (2017), 47–63.

[20] Geoffrion, A. *Proper efficiency and the theory of vector maximization*, J. Math. Anal. Appl. 22 (1968), 618–630.

[21] Ghaznavi-Ghosoni, B.A., Khorram E. and Soleimani-Damaneh, M. *Scalarization for characterization of approximate strong/weak/proper efficiency in multi-objective optimization*, Optimization, 62(6) (2013), 703–720.

[22] Guerraggio, A., Molho, E. and Zaffaroni, A. *On the notion of proper efficiency in vector optimization*, J. Optim. Theory Appl. 82 (1994), 1–21.

[23] Helbig, S. *On a new concept for ε-efficiency*, talk at Optimization Days 1992, Montreal, (1992).

[24] Helbig, S. and Pateva, D. *On several concepts for ε-efficiency*, OR Spectrum, 16 (1994), 179–186.

[25] Henig, M.I. *Proper efficiency with respect to cones*, J. Optim. Theory Appl. 36 (1982), 387–407.

[26] Hoseinpour, N. and Ghaznavi, M. *Identifying approximate proper efficiency in an infinite dimensional space*, RARO Operations Research, 57 (2023) 697–714.

[27] Jahn, J. *Vector optimization*, Springer, Berlin, 2012.

[28] Kiyani, E. and Soleimani-damaneh, M. *Approximate proper efficiency on real linear vector spaces*, Pacific Journal of Optimization, 10 (2014), 715–734.

[29] Kuhn, H. and Tucker, A. *Nonlinear programming*, in Neyman, J. (ed.) Proceeding of the Second Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley, California, (1951) 481–492.

[30] Kutateladze, S.S. *convex ε-programming*, Soviet. Math. Dokl, 20 (1979), 391–393.

[31] Luc, D.T. and Soubeyran, A. *Variable preference relations: existence of maximal elements*, J. Math. Econ. 49 (2013), 251–262.

[32] Maghri, E.l. and Pareto-Fenchel, M. *ε-subdifferential sum rule and ε-efficiency*, Optim. Lett. 6 (2012), 763–781.

[33] Németh, A.B. *A nonconvex vector minimization problem*, Nonlinear Anal. 10 (1986), 669–678.

[34] Rong, W. *Proper ε-efficiency in vector optimization problems with cone-subconvexlikeness*, Acta Sci. Natur. Univ. NeiMongol, 28 (1997), 609–613.

[35] Shahbeyk, S. and Soleimani-damaneh, M. *Proper minimal points of nonconvex sets in Banach spaces in terms of the limiting normal cone*, Optimization, 66 (2017), 473–489.

[36] Shahbeyk, S. and Soleimani-damaneh, M. *Limiting proper minimal points of nonconvex sets in finite-dimensional spaces*, Carpathian J. Math. 35(3) (2019), 370–384.

[37] Shahbeyk, S., Soleimani-damaneh, M. and Kasimbeyli, R. *Hartley properly and super nondominated solutions in vector optimization with a variable ordering structure*, J. Global Optim. 71 (2018), 383–405.

[38] Soleimani, B. *Characterization of Approximate Solutions of Vector Optimization Problems with a Variable Order Structure*, J. Optim. Theory Appl. 162 (2014), 605–632.

[39] Soleimani, B. and Tammer, C. *Concepts for approximate solutions of vector optimization problems with variable order structure*, Vietnam J. Math. 42 (2014), 43–566.

[40] Tammer, C. *Stability results for approximately efficient solutions*, OR Spectrum, 16 (1994), 47–52.

[41] Tanaka, T. *A new approach to approximation of solutions in vector optimization problems*, in Proceeding of APORS, 1994, M. Fushimi and K. Tone, eds., World Scientific Publishing, Singapore, (1995) 497–504.

[42] Wacker, M. and Deinzer, F. *Automatic robust medical image registration using a new democratic vector optimization approach with multiple measures*, in Yang GZ, Hawkes D, Rueckert D, Noble A, Taylor C, editors. Vol. 5761, Lecture notes in computer science, 12th International Conference, MICCAI 2009, 2009 Sept 20–24. Heidelberg, Springer, London, UK, (2009) 590–597.

[43] White, D.J. *Epsilon efficiency*, J. Optim. Theory Appl. 49 (1986), 319–337.

[44] Yu, P.l. *Cone convexity, cone extreme points, and nondominated solutions in decision problems with multiobjectives*, J. Optim. Theory Appl. 14 (1974), 319–377.

[45] Zamani, M. and Soleimani-damaneh, M. *Proper efficiency, scalarization and transformation in multi-objective optimization: unified approaches*, Optimization, (2020), 1–22.

Iranian Journal of Numerical Analysis and Optimization
Vol. 14, No. 1, 2024, pp 136–171
https://doi.org/10.22067/ijnao.2023.82954.1281
https://ijnao.um.ac.ir/

CC BY NC

Research Article

A study of global dynamics and sensitivity analysis of a discrete-time model of the COVID-19 epidemic

M. E.-B. Keddar and O. Belhamiti*,

Abstract

This study presents a novel approach to understanding the global dynamics of COVID-19 transmission with vaccination based on a discrete-time model. We establish biologically meaningful constraints on the model parameters and prove the existence of a disease-free equilibrium and an endemic equilibrium under these constraints, along with their theoretical stabilities. Furthermore, we identify the most sensitive operational parameters that have a substantial impact on the transmission of the epidemic. Through numerical simulations, we demonstrate the local stability of the two equilibria, depending on the parameter values. Our findings reveal that the discrete-time model is not only dynamically robust but also more

*Corresponding author

Received 14 June 2023; revised 3 November 2023; accepted 6 November 2023

Meriem El-Batoul Keddar
Department of Mathematics and Computer Science, Abdelhamid Ibn Badis Mostaganem University, Algeria. e-mail: batoulkeddar@gmail.com

Omar Belhamiti
Department of Mathematics and Computer Science, Abdelhamid Ibn Badis Mostaganem University, Algeria. e-mail: omar.belhamiti@univ-mosta.dz

How to cite this article
Keddar, M. E.-B. and Belhamiti, O., A study of global dynamics and sensitivity analysis of a discrete-time model of the COVID-19 epidemic. *Iran. J. Numer. Anal. Optim.*, 2024; 14(1): 136-171. https://doi.org/10.22067/ijnao.2023.82954.1281

realistic than its continuous counterpart under biologically meaningful constraints. These results provide a foundation for future research in this area and contribute to our understanding of the global dynamics of COVID-19 transmission.

AMS subject classifications (2020): Primary 39A12; Secondary 92C60, 92D30.

Keywords: Discrete-time epidemic model; Euler Scheme, Local stability; Vaccination reproduction number.

1 Introduction and Motivation

Coronaviruses are a family of viruses that cause respiratory and intestinal infections in humans and animals [6]. They were not dangerous to humans until the appearance of severe acute respiratory syndrome coronavirus (SARS-CoV) in 2002 in China [16]. This virus had infected 8,096 people in 25 countries, causing 774 deaths [19]. A decade later, the outbreak of another highly pathogenic coronavirus, Middle East respiratory syndrome coronavirus (MERS-CoV), occurred in Jeddah, Saudi Arabia [32]. The virus spread, infecting more than 2,400 people in 27 countries and causing over 850 deaths [13]. In December 2019, a third of these dangerous coronaviruses appeared in the city of Wuhan, China. Forty-four cases were reported in this country on January 3, 2020 [24]. The virus migrated to neighboring regions and countries. In less than a month, 106 confirmed cases were detected in 19 different countries, and the number of infected individuals reached 9,826, with 213 deaths. The World Health Organization (WHO) declared the outbreak to be a public health emergency of international concern [24]. The Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) is the virus that caused this disease, which has spread throughout the world. It was named COVID-19 by the WHO [23], which declared that this malady can be characterized as a pandemic [25]. On April 4, 2020, the number of infected people exceeded 1 million, with 56,986 deaths worldwide [26]. Many countries have gone into total confinement and closed their borders to reduce the spread of the virus. The number of infected people increases more and more rapidly over time.

In December of the following year, more than 75 million confirmed cases and 1.6 million deaths have been reported globally [27]. As of November 2022, COVID-19 still exists, and its spread worldwide has resulted in 629 million confirmed cases and 6.5 million deaths [28]. The number is increasing even now.

This pandemic has left the world feeling helpless due to its rapid spread and the lack of specific medicines or treatments available to combat it. As a result, many countries have implemented non-pharmaceutical measures, such as confinement, mask-wearing, social distancing, isolation, and treatment of infected persons, in an attempt to contain the virus. However, the effectiveness of these measures has proven to be limited. This has prompted epidemiologists to develop vaccines as a solution. Many of the developed vaccines are safe and effective in providing protection against severe forms of COVID-19, offering a glimmer of hope for an end to this pandemic that has impacted humanity for over two years. The COVID-19 pandemic has affected all sectors of health and security, including the involvement of mathematicians to control and stop its spread. Numerous mathematical models have been proposed, allowing for the control of the pandemic, prediction of its evolution, and estimation of key transmission parameters, such as the basic reproduction number R_0, as well as analyzing the impact of vaccines on disease dynamics. Most of these mathematical models are nonlinear continuous-time systems of the SEIR type, which study the dynamic transmission of COVID-19, such as those presented in [21, 12, 15]. Some models also include vaccination, as seen in [1, 31, 30]. However, only a few models are discretized-time models.

It should be noted that discrete-time models present a great advantage compared to continuous-time models, especially for COVID-19, as the individual goes through several phases (Susceptible-Exposed-Infected ...) and the phases of the disease represent the step of discretization, which can vary between five days to a week, depending on the individual. Despite these considerations, only a few articles have been published in discrete time that take into account the disease phases. In [8], the authors proposed a SEIR model with several characteristics. Firstly, the exposed subpopulation is infectious. They also incorporated a feedback vaccination control law on the susceptible. Furthermore, the model is subject to delayed partial re-susceptibility,

meaning that a partial immunity loss in the recovered individuals occurs after some time. By the same authors in [9], a new discrete SEIR model has been developed. This model considers the injection of two doses with different delays and a possible average different effectiveness. It also considers that the exposed have a transmission rate exposed-susceptible, which may be eventually distinct from the infectious-susceptible transmission rate. In [2], the authors proposed a new SIR model with vaccination. Their model has two equilibrium points, namely, a disease-free equilibrium (DFE) and an endemic equilibrium. By considering both the forward difference system and the backward difference system, some stability analyses of the DFE point are performed.

This paper describes a new discrete SEIR model for COVID-19 transmission dynamics that include vaccination. The model assumes that only symptomatic and asymptomatic infected individuals can transmit the disease and that hospitalized patients are isolated. It also considers the possibility of reinfection with the virus, as the vaccine may not be fully protective. [4]. Some exposed and asymptomatic infected can be vaccinated; this is due to the inefficiency of the tests [29]. The vaccination reproduction number R_v and the equilibrium points are calculated, and their sensibility and stability analysis are studied. Numerical simulations are performed considering various scenarios. We present detailed results in the form of graphs and tables, including new numerical results obtained for model validation.

The paper is organized as follows. Section 2 introduces the mathematical model and provides a preliminary overview. In section 3, the mathematical analysis is carried out to investigate the stability of the proposed model. This section focuses on the study of equilibrium points and their stability properties. Section 4 presents a sensitivity analysis of the model to investigate how changes in the model parameters affect the system behavior. Finally, section 5 presents the results and discussions of the model simulations, which provide insights into the system's behavior and its implications.

2 Mathematical model

In this section, we use the explicit Euler scheme to obtain a discrete-time model for the transmission of COVID-19. Based on the epidemiological status and the timeline of COVID-19 infection in [18], the population is divided as follows:

- **S** is the susceptible class,
- **E** is the exposed class, for example, the people in the incubation period,
- **V** is the vaccinated class,
- $\mathbf{I}^a$ is the asymptomatic infected class,
- $\mathbf{I}^s$ is a class of individuals diagnosed with moderate COVID-19 symptoms, which is further divided into two subclasses I^{s_1} refers to individuals experiencing symptoms in their first phase (i.e., within a week of symptom onset), as described in [11, 17] while $\mathbf{I}^{s_2}$ refers to individuals experiencing symptoms in their second phase (i.e., roughly one week after the end of the first phase of symptoms), also documented in references [11, 17]. That is,

$$\mathbf{I}^s = \mathbf{I}^{s_1} + \mathbf{I}^{s_2}.$$

- $\mathbf{I}^h$ refers to the class of infected individuals who require hospitalization, including those who need intensive care. This class is further subdivided into three subclasses based on the interval since symptom onset: $\mathbf{I}^{h_1}$ represents individuals who are hospitalized within a week of symptom onset, as detailed in references [11, 17], while $\mathbf{I}^{h_2}$ denotes those who are hospitalized around one week after the end of the first phase of symptoms. Lastly, $\mathbf{I}^{h_3}$ pertains to individuals who are hospitalized approximately one week after the end of the second phase of symptoms, also documented in references [11, 17]. That is,

$$\mathbf{I}^h = \mathbf{I}^{h_1} + \mathbf{I}^{h_2} + \mathbf{I}^{h_3}.$$

- **R** represents the class of individuals who have recovered from COVID-19.

- **D** represents the class of individuals who have died from COVID-19.

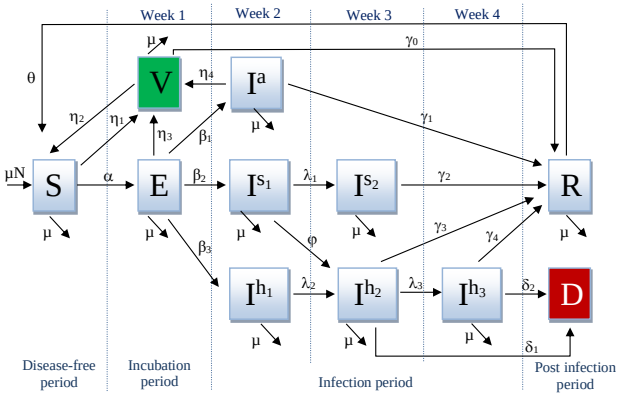


Figure 1: Timeline of infection and states transition scheme description.

By applying the explicit Euler scheme, we can obtain a discretized SEIR epidemic model that incorporates vaccination (see Figure 1). Considering the week-long interval required for individuals to transition between classes, the resulting model is as follows:

$$
\begin{cases}
\mathbf{S}(t+1) = \mathbf{S}(t) + \mu N - \eta_1 \mathbf{S}(t) + \eta_2 \mathbf{V}(t) + \theta \mathbf{R}(t) - \mu \mathbf{S}(t) \\
\qquad\qquad -\alpha \mathbf{S}(t) \dfrac{(\mathbf{I}^a(t) + \mathbf{I}^{s_1}(t) + \mathbf{I}^{s_2}(t))}{N}, \\
\mathbf{E}(t+1) = \mathbf{E}(t) + \alpha \mathbf{S}(t) \dfrac{(\mathbf{I}^a(t) + \mathbf{I}^{s_1}(t) + \mathbf{I}^{s_2}(t))}{N} \\
\qquad\qquad -(\beta_1 + \beta_2 + \beta_3 + \eta_3 + \mu)\mathbf{E}(t), \\
\mathbf{V}(t+1) = \mathbf{V}(t) + \eta_1 \mathbf{S}(t) + \eta_3 \mathbf{E}(t) + \eta_4 \mathbf{I}^a(t) - (\eta_2 + \gamma_0 + \mu)\mathbf{V}(t), \\
\mathbf{I}^a(t+1) = \mathbf{I}^a(t) + \beta_1 \mathbf{E}(t) - (\eta_4 + \gamma_1 + \mu)\mathbf{I}^a(t), \\
\mathbf{I}^{s_1}(t+1) = \mathbf{I}^{s_1}(t) + \beta_2 \mathbf{E}(t) - (\lambda_1 + \varphi + \mu)\mathbf{I}^{s_1}(t), \\
\mathbf{I}^{s_2}(t+1) = \mathbf{I}^{s_2}(t) + \lambda_1 \mathbf{I}^{s_1}(t) - (\gamma_2 + \mu)\mathbf{I}^{s_2}(t), \\
\mathbf{I}^{h_1}(t+1) = \mathbf{I}^{h_1}(t) + \beta_3 \mathbf{E}(t) - (\lambda_2 + \mu)\mathbf{I}^{h_1}(t), \\
\mathbf{I}^{h_2}(t+1) = \mathbf{I}^{h_2}(t) + \lambda_2 \mathbf{I}^{h_1}(t) + \varphi \mathbf{I}^{s_1}(t) - (\lambda_3 + \gamma_3 + \delta_1 + \mu)\mathbf{I}^{h_2}(t), \\
\mathbf{I}^{h_3}(t+1) = \mathbf{I}^{h_3}(t) + \lambda_3 \mathbf{I}^{h_2}(t) - (\gamma_4 + \delta_2 + \mu)\mathbf{I}^{h_3}(t), \\
\mathbf{R}(t+1) = \mathbf{R}(t) + \gamma_0 \mathbf{V}(t) + \gamma_1 \mathbf{I}^a(t) + \gamma_2 \mathbf{I}^{s_2}(t) + \gamma_3 \mathbf{I}^{h_2}(t) \\
\qquad\qquad +\gamma_4 \mathbf{I}^{h_3}(t) - (\theta + \mu)\mathbf{R}(t), \\
\mathbf{D}(t+1) = \mathbf{D}(t) + \delta_1 \mathbf{I}^{h_2}(t) + \delta_2 \mathbf{I}^{h_3}(t),
\end{cases}
\tag{1}
$$

We assume that the total population N remains constant,

$$N(t+1) = N(t).$$

Taking into account the previous assumption, we can normalize the variables in the model (1),

$$\begin{aligned}
&S(t) = \frac{\mathbf{S}(t)}{N}, \quad E(t) = \frac{\mathbf{E}(t)}{N}, \quad V(t) = \frac{\mathbf{V}(t)}{N}, \quad I^a(t) = \frac{\mathbf{I}^a(t)}{N},\\
&I^{s_1}(t) = \frac{\mathbf{I}^{s_1}(t)}{N}, \quad I^{s_2}(t) = \frac{\mathbf{I}^{s_2}(t)}{N}, \quad I^{h_1}(t) = \frac{\mathbf{I}^{h_1}(t)}{N}, \quad I^{h_2}(t) = \frac{\mathbf{I}^{h_2}(t)}{N},\\
&I^{h_3}(t) = \frac{\mathbf{I}^{h_3}(t)}{N}, \quad R(t) = \frac{\mathbf{R}(t)}{N}, \quad D(t) = \frac{\mathbf{D}(t)}{N}.
\end{aligned}$$

Because

$$\begin{aligned}
&S(t) + E(t) + V(t) + I^a(t) + I^{s_1}(t) + I^{s_2}(t)\\
&\quad + I^{h_1}(t) + I^{h_2}(t) + I^{h_3}(t) + R(t) + D(t) = 1,
\end{aligned}$$

then

$$\begin{aligned}
D(t) = 1 - (&S(t) + E(t) + V(t) + I^a(t) + I^{s_1}(t)\\
&+ I^{s_2}(t) + I^{h_1}(t) + I^{h_2}(t) + I^{h_3}(t) + R(t)\big).
\end{aligned}$$

Then, we have the following normalized reduced system:

$$\begin{cases}
S(t+1) = S(t) + \mu - \eta_1 S(t) + \eta_2 V(t) + \theta R(t) - \mu S(t)\\
\qquad -\alpha S(t)\left(I^a(t) + I^{s_1}(t) + I^{s_2}(t)\right),\\
E(t+1) = E(t) + \alpha S(t)\left(I^a(t) + I^{s_1}(t) + I^{s_2}(t)\right)\\
\qquad -(\beta_1 + \beta_2 + \beta_3 + \eta_3 + \mu) E(t),\\
V(t+1) = V(t) + \eta_1 S(t) + \eta_3 E(t) + \eta_4 I^a(t) - (\eta_2 + \gamma_0 + \mu) V(t),\\
I^a(t+1) = I^a(t) + \beta_1 E(t) - (\eta_4 + \gamma_1 + \mu) I^a(t),\\
I^{s_1}(t+1) = I^{s_1}(t) + \beta_2 E(t) - (\lambda_1 + \varphi + \mu) I^{s_1}(t),\\
I^{s_2}(t+1) = I^{s_2}(t) + \lambda_1 I^{s_1}(t) - (\gamma_2 + \mu) I^{s_2}(t),\\
I^{h_1}(t+1) = I^{h_1}(t) + \beta_3 E(t) - (\lambda_2 + \mu) I^{h_1}(t),\\
I^{h_2}(t+1) = I^{h_2}(t) + \lambda_2 I^{h_1}(t) + \varphi I^{s_1}(t) - (\lambda_3 + \gamma_3 + \delta_1 + \mu) I^{h_2}(t),\\
I^{h_3}(t+1) = I^{h_3}(t) + \lambda_3 I^{h_2}(t) - (\gamma_4 + \delta_2 + \mu) I^{h_3}(t),\\
R(t+1) = R(t) + \gamma_0 V(t) + \gamma_1 I^a(t) + \gamma_2 I^{s_2}(t) + \gamma_3 I^{h_2}(t)\\
\qquad +\gamma_4 I^{h_3}(t) - (\theta + \mu) R(t),
\end{cases} \tag{2}$$

where

μ represents the natural mortality,

α represents the transmission rate of COVID-19,

$(\eta_i)_{i=\overline{1,3,4}}$ represent the vaccination rates for each class,

$(\beta_i)_{i=\overline{1,3}}$ represent the rates at which exposed individuals become infected,

$(\gamma_i)_{i=\overline{0,4}}$ represent the recovery rates from each class,

$(\lambda_i)_{i=\overline{1,3}}$ represent the rates of transition from one class to another,

$(\delta_i)_{i=1,2}$ signify COVID-19 mortality rates per hospitalized group,

θ and η_2 denote COVID-19 reinfection rates, and

φ represents the hospitalization rate in the symptomatic phase I^{s_1}.

It is essential to assume that all model parameters are positive. Furthermore, some conditions on these parameters are necessary to ensure that the model has biological significance. In the following subsection, we present and illustrate these conditions.

2.1 Positive invariant set

The domain

$$\Omega = \Big\{\big(S, E, V, I^a, I^{s_1}, I^{s_2}, I^{h_1}, I^{h_2}, I^{h_3}, R\big) \in \mathbb{R}_+^{10} : \\ 0 \leqslant S + E + V + I^a + I^{s_1} + I^{s_2} + I^{h_1} + I^{h_2} + I^{h_3} + R \leqslant 1\Big\}$$

is a positive invariant set for system (2).

To render the model biologically meaningful, it is imperative to exhibit the belonging of solutions from the system (2) into Ω. This, in turn, hinges on ensuring that the combined outflow rates for each state do not surpass 1, as follows:

$$\begin{cases} \eta_1 + \alpha + \mu \leqslant 1, \\ \beta_1 + \beta_2 + \beta_3 + \eta_3 + \mu \leqslant 1, \\ \eta_2 + \gamma_0 + \mu \leqslant 1, \\ \eta_4 + \gamma_1 + \mu \leqslant 1, \\ \lambda_1 + \varphi + \mu \leqslant 1, \end{cases} \quad \text{and} \quad \begin{cases} \gamma_2 + \mu \leqslant 1, \\ \lambda_2 + \mu \leqslant 1, \\ \lambda_3 + \gamma_3 + \delta_1 + \mu \leqslant 1, \\ \gamma_4 + \delta_2 + \mu \leqslant 1, \\ \theta + \mu \leqslant 1. \end{cases}$$

Let $\big(S_n, E_n, V_n, I_n^a, I_n^{s_1}, I_n^{s_2}, I_n^{h_1}, I_n^{h_2}, I_n^{h_3}, R_n\big) \in \Omega$. Then, we have

$$\begin{cases} S_{n+1} = S_n + \mu - \eta_1 S_n + \eta_2 V_n + \theta R_n - \alpha S_n (I_n^a + I_n^{s_1} + I_n^{s_2}) - \mu S_n \\ \qquad \geqslant \mu + (1 - (\eta_1 + \alpha + \mu)) S_n \geqslant 0, \\ E_{n+1} = E_n + \alpha S_n (I_n^a + I_n^{s_1} + I_n^{s_2}) - (\beta_1 + \beta_2 + \beta_3 + \eta_3 + \mu) E_n \\ \qquad \geqslant (1 - (\beta_1 + \beta_2 + \beta_3 + \eta_3 + \mu)) E_n \geqslant 0, \\ V_{n+1} = V_n + \eta_1 S_n + \eta_3 E_n + \eta_4 I_n^a - (\eta_2 + \gamma_0 + \mu) V_n \\ \qquad \geqslant (1 - (\eta_2 + \gamma_0 + \mu)) V_n \geqslant 0, \\ I_{n+1}^a = I_n^a + \beta_1 E_n - (\eta_4 + \gamma_1 + \mu) I_n^a \\ \qquad \geqslant (1 - (\eta_4 + \gamma_1 + \mu)) I_n^a \geqslant 0, \\ I_{n+1}^{s_1} = I_n^{s_1} + \beta_2 E_n - (\lambda_1 + \varphi + \mu) I_n^{s_1} \\ \qquad \geqslant (1 - (\lambda_1 + \varphi + \mu)) I_n^{s_1} \geqslant 0, \\ I_{n+1}^{s_2} = I_n^{s_2} + \lambda_1 I_n^{s_1} - (\gamma_2 + \mu) I_n^{s_2} \\ \qquad \geqslant (1 - (\gamma_2 + \mu)) I_n^{s_2} \geqslant 0, \\ I_{n+1}^{h_1} = I_n^{h_1} + \beta_3 E_n - (\lambda_2 + \mu) I_n^{h_1} \\ \qquad \geqslant (1 - (\lambda_2 + \mu)) I_n^{h_1} \geqslant 0, \\ I_{n+1}^{h_2} = I_n^{h_2} + \lambda_2 I_n^{h_1} + \varphi I_n^{s_1} - (\lambda_3 + \gamma_3 + \delta_1 + \mu) I_n^{h_2} \\ \qquad \geqslant (1 - (\lambda_3 + \gamma_3 + \delta_1 + \mu)) I_n^{h_2} \geqslant 0, \\ I_{n+1}^{h_3} = I_n^{h_3} + \lambda_3 I_n^{h_2} - (\gamma_4 + \delta_2 + \mu) I_n^{h_3} \\ \qquad \geqslant (1 - (\gamma_4 + \delta_2 + \mu)) I_n^{h_3} \geqslant 0, \\ R_{n+1} = R_n + \gamma_0 V_n + \gamma_1 I_n^a + \gamma_2 I_n^{s_2} + \gamma_3 I_n^{h_2} + \gamma_4 I_n^{h_3} - (\theta + \mu) R_n \\ \qquad \geqslant (1 - (\theta + \mu)) R_n \geqslant 0. \end{cases}$$

Thus,

$$S_{n+1} + E_{n+1} + V_{n+1} + I_{n+1}^a + I_{n+1}^{s_1} + I_{n+1}^{s_2} + I_{n+1}^{h_1} + I_{n+1}^{h_2} + I_{n+1}^{h_3} + R_{n+1} \geqslant 0.$$

On the other hand, the sum of the system equations (2) gives

$$\begin{aligned} &S_{n+1} + E_{n+1} + V_{n+1} + I_{n+1}^a + I_{n+1}^{s_1} + I_{n+1}^{s_2} + I_{n+1}^{h_1} + I_{n+1}^{h_2} + I_{n+1}^{h_3} + R_{n+1} \\ &\quad = \mu + (1 - \mu) \left(S_n + E_n + V_n + I_n^a + I_n^{s_1} + I_n^{s_2} + I_n^{h_1} + I_n^{h_2} + I_n^{h_3} + R_n \right) - \left(\delta_1 I_n^{h_2} + \delta_2 I_n^{h_3} \right) \\ &\quad \leqslant \mu + (1 - \mu) \left(S_n + E_n + V_n + I_n^a + I_n^{s_1} + I_n^{s_2} + I_n^{h_1} + I_n^{h_2} + I_n^{h_3} + R_n \right) \\ &\quad \leqslant \mu + (1 - \mu) \leqslant 1. \end{aligned}$$

Finally, we can say that

$$\left(S_{n+1}, E_{n+1}, V_{n+1}, I_{n+1}^a, I_{n+1}^{s_1}, I_{n+1}^{s_2}, I_{n+1}^{h_1}, I_{n+1}^{h_2}, I_{n+1}^{h_3}, R_{n+1} \right) \in \Omega.$$

2.2 Vaccination reproduction number R_v

The basic reproduction number, R_v, is a fundamental epidemiological parameter that quantifies the contagiousness of an infectious disease. It represents the average number of secondary infections generated by one infected person in a susceptible population. If R_v is greater than 1, then the disease can sustain transmission; if it is less than 1, then it is likely to die out.

In this subsection, we describe the method used to estimate the vaccination reproduction number R_v (also called a control reproduction number) for our SEIR-V model. This method, proposed by Diekmann, Heesterbeek, and Metz [10], and further elaborated by Van Den Driessche and Watmough [22], involves using the next-generation matrix to determine R_v for an ordinary differential equation compartmental model.

Starting from the system (2), we can obtain the nontrivial DFE point as follows (is calculated in the next section):

$$E_0 = \left(S^*, 0, \frac{\eta_1 S^*}{(\eta_2+\gamma_0+\mu)}, 0, 0, 0, 0, 0, \frac{\gamma_0\eta_1 S^*}{(\theta+\mu)(\eta_2+\gamma_0+\mu)} \right)$$

with

$$S^* = \frac{(\theta+\mu)(\eta_2+\gamma_0+\mu)}{(\theta+\mu)(\eta_2+\gamma_0+\mu)+\eta_1(\theta+\mu)+\eta_1\gamma_0}.$$

Let $X = \left(E, V, I^a, I^{s_1}, I^{s_2}, I^{h_1}, I^{h_2}, I^{h_3}\right)^T$. Then the model can be written as

$$\begin{aligned} & X(t+1) - X(t) \\ & = \mathcal{F}(X(t)) - \mathcal{V}(X(t)) \\ & = \begin{pmatrix} \alpha S(t)\left(I^a(t) + I^{s_1}(t) + I^{s_2}(t)\right) - (\beta_1+\beta_2+\beta_3+\eta_3+\mu)E(t) \\ \eta_1 S(t) + \eta_3 E(t) + \eta_4 I^a(t) - (\eta_2+\gamma_0+\mu)V(t) \\ \beta_1 E(t) - (\eta_4+\gamma_1+\mu)I^a(t) \\ \beta_2 E(t) - (\lambda_1+\varphi+\mu)I^{s_1}(t) \\ \lambda_1 I^{s_1}(t) - (\gamma_2+\mu)I^{s_2}(t) \\ \beta_3 E(t) - (\lambda_2+\mu)I^{h_1}(t) \\ \lambda_2 I^{h_1}(t) + \varphi I^{s_1}(t) - (\lambda_3+\gamma_3+\delta_1+\mu)I^{h_2}(t) \\ \lambda_3 I^{h_2}(t) - (\gamma_4+\delta_2+\mu)I^{h_3}(t) \end{pmatrix}. \end{aligned}$$

We represent $\mathcal{F}(X)$ as the rate of appearance of new infections in each compartment, while $\mathcal{V}(X)$ signifies the disparity between the rate of individuals leaving each compartment and the rate of individuals entering the same compartment through all other means. It is assumed that each function is continuously differentiable at least twice in each variable. So, we have

$$\mathcal{F}(X) = \begin{pmatrix} \alpha S(I^a + I^{s_1} + I^{s_2}) \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

and

$$\mathcal{V}(X) = \begin{pmatrix} (\beta_1 + \beta_2 + \beta_3 + \eta_3 + \mu) E \\ (\eta_2 + \gamma_0 + \mu) V - \eta_1 S - \eta_3 E - \eta_4 I^a \\ (\eta_4 + \gamma_1 + \mu) I^a - \beta_1 E \\ (\lambda_1 + \varphi + \mu) I^{s_1} - \beta_2 E \\ (\gamma_2 + \mu) I^{s_2} - \lambda_1 I^{s_1} \\ (\lambda_2 + \mu) I^{h_1} - \beta_3 E \\ (\lambda_3 + \gamma_3 + \delta_1 + \mu) I^{h_2} - \lambda_2 I^{h_1} - \varphi I^{s_1} \\ (\gamma_4 + \delta_2 + \mu) I^{h_3} - \lambda_3 I^{h_2} \end{pmatrix}.$$

The vaccination reproduction number, R_v, is calculated by next generation technique. The $\widetilde{F}$ and $\widetilde{V}$ matrices at DFE is given as follows:

$$F = \left.\frac{\partial \mathcal{F}(X)}{\partial X}\right|_{E_0} = \alpha S^* \begin{pmatrix} 0\,0\,1\,1\,1\,0\,0\,0 \\ 0\,0\,0\,0\,0\,0\,0\,0 \\ 0\,0\,0\,0\,0\,0\,0\,0 \\ 0\,0\,0\,0\,0\,0\,0\,0 \\ 0\,0\,0\,0\,0\,0\,0\,0 \\ 0\,0\,0\,0\,0\,0\,0\,0 \\ 0\,0\,0\,0\,0\,0\,0\,0 \\ 0\,0\,0\,0\,0\,0\,0\,0 \end{pmatrix}$$

and

$$V = \left.\frac{\partial \mathcal{V}(X)}{\partial X}\right|_{X=E_0},$$

where

$$\begin{cases} V_{1,1} = (\beta_1 + \beta_2 + \beta_3 + \eta_3 + \mu), \\ V_{2,1} = -\eta_3, \\ V_{3,1} = -\beta_1, \\ V_{4,1} = -\beta_2, \\ V_{6,1} = -\beta_3, \\ V_{2,2} = (\eta_2 + \gamma_0 + \mu), \\ V_{2,3} = -\eta_4, \\ V_{3,3} = (\eta_4 + \gamma_1 + \mu), \\ V_{4,4} = (\lambda_1 + \varphi + \mu), \end{cases} \quad \begin{cases} V_{5,4} = -\lambda_1, \\ V_{5,5} = (\gamma_2 + \mu), \\ V_{6,6} = (\lambda_2 + \mu), \\ V_{7,4} = -\varphi, \\ V_{7,6} = -\lambda_2, \\ V_{7,7} = (\lambda_3 + \gamma_3 + \delta_1 + \mu), \\ V_{8,7} = -\lambda_3, \\ V_{8,8} = (\gamma_4 + \delta_2 + \mu), \\ 0 \qquad \text{otherwise.} \end{cases}$$

Therefore, the next generation matrix is

$$\begin{cases} \left(FV^{-1}\right)_{1,1} = \dfrac{\alpha}{(\beta_1 + \beta_2 + \beta_3 + \eta_3 + \mu)} \left(\dfrac{\beta_1}{(\eta_4 + \gamma_1 + \mu)} + \dfrac{\beta_2 (\gamma_2 + \mu) + \lambda_1 \beta_2}{(\lambda_1 + \varphi + \mu)(\gamma_2 + \mu)} \right) S^*, \\ \left(FV^{-1}\right)_{1,3} = \dfrac{\alpha}{(\eta_4 + \gamma_1 + \mu)} S^*, \\ \left(FV^{-1}\right)_{1,4} = \left(1 + \dfrac{\lambda_1}{(\gamma_2 + \mu)}\right) \dfrac{\alpha S^*}{(\lambda_1 + \varphi + \mu)}, \\ \left(FV^{-1}\right)_{1,5} = \dfrac{\alpha S^*}{(\gamma_2 + \mu)}, \\ \qquad 0 \quad \text{otherwise.} \end{cases}$$

According to Van Den Driessche and Watmough [22], the vaccination reproduction number, R_v, is the spectral radius of the next generation matrix, which is given by

$$R_v = \alpha S^* \frac{\beta_2 (\eta_4 + \gamma_1 + \mu)(\gamma_2 + \mu) + \beta_1 (\lambda_1 + \varphi + \mu)(\gamma_2 + \mu) + \lambda_1 \beta_2 (\eta_4 + \gamma_1 + \mu)}{(\beta_1 + \beta_2 + \beta_3 + \eta_3 + \mu)(\eta_4 + \gamma_1 + \mu)(\lambda_1 + \varphi + \mu)(\gamma_2 + \mu)}, \tag{3}$$

where

$$S^* = \frac{(\theta + \mu)(\eta_2 + \gamma_0 + \mu)}{(\theta + \mu)(\eta_2 + \gamma_0 + \mu + \eta_1) + \eta_1 \gamma_0},$$

2.3 Equilibrium points

The equilibrium points for the model (2) can be obtained by solving the following equations:

$$S(t+1) - S(t) = E(t+1) - E(t) = V(t+1) - V(t) = 0$$

$$I^a(t+1) - I^a(t) = I^{s_1}(t+1) - I^{s_1}(t) = I^{s_2}(t+1) - I^{s_2}(t) = 0$$
$$I^{h_1}(t+1) - I^{h_1}(t) = I^{h_2}(t+1) - I^{h_2}(t) = I^{h_3}(t+1) - I^{h_3}(t) = 0$$
$$R(t+1) - R(t) = 0.$$

The system (2) has two equilibrium points: the DFE and the endemic equilibrium. For strictly positive parameters, the DFE is given by

$$E_0 = \left(S^*, 0, \frac{\eta_1 S^*}{(\eta_2+\gamma_0+\mu)}, 0, 0, 0, 0, 0, \frac{\gamma_0\eta_1 S^*}{(\theta+\mu)(\eta_2+\gamma_0+\mu)}\right), \tag{4}$$

with

$$S^* = \frac{(\theta+\mu)(\eta_2+\gamma_0+\mu)}{(\theta+\mu)(\eta_2+\gamma_0+\mu+\eta_1)+\eta_1\gamma_0},$$

and the endemic equilibrium,

$$E_1 = \left(S^{**}, E^{**}, V^{**}, I^{a^{**}}, I^{s_1^{**}}, I^{s_2^{**}}, I^{h_1^{**}}, I^{h_2^{**}}, I^{h_3^{**}}, R^{**}\right) \tag{5}$$

exists if $R_v > 1$ with

$$S^{**} = \frac{1}{\alpha}\frac{K_1K_3K_4K_5}{\beta_2K_3(K_5+\lambda_1)+\beta_1K_4K_5},$$
$$E^{**} = \mu\frac{K_2K_3K_4K_5K_6K_7K_8K_9}{K}\left(1-\frac{1}{R_v}\right),$$
$$V^{**} = \frac{1}{\alpha}\frac{\eta_1}{K_2}\frac{K_1K_3K_4K_5}{\beta_2K_3(K_5+\lambda_1)+\beta_1K_4K_5},$$
$$+\mu\left(\eta_3+\frac{\eta_4\beta_1}{K_3}\right)\frac{K_3K_4K_5K_6K_7K_8K_9}{K}\left(1-\frac{1}{R_v}\right),$$
$$I^{a^{**}} = \mu\beta_1\frac{K_2K_4K_5K_6K_7K_8K_9}{K}\left(1-\frac{1}{R_v}\right),$$
$$I^{s_1^{**}} = \mu\beta_2\frac{K_2K_3K_5K_6K_7K_8K_9}{K}\left(1-\frac{1}{R_v}\right),$$
$$I^{s_2^{**}} = \mu\lambda_1\beta_2\frac{K_2K_3K_6K_7K_8K_9}{K}\left(1-\frac{1}{R_v}\right),$$
$$I^{h_1^{**}} = \mu\beta_3\frac{K_2K_3K_4K_5K_7K_8K_9}{K}\left(1-\frac{1}{R_v}\right),$$
$$I^{h_2^{**}} = \mu\left(\frac{\lambda_2\beta_3}{K_6}+\frac{\varphi\beta_2}{K_4}\right)\frac{K_2K_3K_4K_5K_6K_8K_9}{K}\left(1-\frac{1}{R_v}\right),$$
$$I^{h_3^{**}} = \mu\lambda_3\left(\frac{\lambda_2\beta_3}{K_6}+\frac{\varphi\beta_2}{K_4}\right)\frac{K_2K_3K_4K_5K_6K_9}{K}\left(1-\frac{1}{R_v}\right),$$

$$R^{**} = \frac{1}{\alpha}\frac{\gamma_0\eta_1}{K_9K_2}\frac{K_1K_3K_4K_5}{\beta_2K_3\left(K_5+\lambda_1\right)+\beta_1K_4K_5} + \mu\left(\frac{\gamma_0}{K_2}\left(\eta_3+\frac{\eta_4\beta_1}{K_3}\right)+\frac{\gamma_1\beta_1}{K_3}+\frac{\gamma_2\lambda_1\beta_2}{K_5K_4}\right.$$
$$\left.+\frac{\gamma_3}{K_7}\left(\frac{\lambda_2\beta_3}{K_6}+\frac{\varphi\beta_2}{K_4}\right)+\frac{\gamma_4\lambda_3}{K_8K_7}\left(\frac{\lambda_2\beta_3}{K_6}+\frac{\varphi\beta_2}{K_4}\right)\right)$$
$$\times\frac{K_2K_3K_4K_5K_6K_7K_8}{K}\left(1-\frac{1}{R_v}\right),$$

and

$$\begin{cases} K_0 = (\eta_1+\mu), \\ K_1 = (\beta_1+\beta_2+\beta_3+\eta_3+\mu), \\ K_2 = (\eta_2+\gamma_0+\mu), \\ K_3 = (\eta_4+\gamma_1+\mu), \\ K_4 = (\lambda_1+\varphi+\mu), \end{cases} \quad \begin{cases} K_5 = (\gamma_2+\mu), \\ K_6 = (\lambda_2+\mu), \\ K_7 = (\lambda_3+\gamma_3+\delta_1+\mu), \\ K_8 = (\gamma_4+\delta_2+\mu), \\ K_9 = (\theta+\mu), \end{cases} \tag{6}$$

$$\begin{aligned} K = {} & \mu K_2K_3K_4K_5K_6K_7K_8K_9 + \mu\eta_3K_3K_4K_5K_6K_7K_8K_9 \\ & +\mu\eta_3\gamma_0K_3K_4K_5K_6K_7K_8 + \mu\gamma_4\beta_2\lambda_3\varphi K_2K_3K_5K_6 \\ & +\mu\gamma_4\lambda_2\lambda_3\beta_3K_2K_3K_4K_5 + \mu\beta_3K_2K_3K_4K_5K_7K_8K_9 \\ & +\mu\gamma_3\lambda_2\beta_3K_2K_3K_4K_5K_8 + \mu\beta_2K_2K_3K_5K_6K_7K_8K_9 \\ & +\mu\lambda_1\beta_2K_2K_3K_6K_7K_8K_9 + \mu\beta_1\gamma_0K_4K_5K_6K_7K_8K_9 \\ & +\mu\gamma_2\lambda_1\beta_2K_2K_3K_6K_7K_8 + \mu\beta_1\eta_2K_4K_5K_6K_7K_8K_9 \\ & +\mu\gamma_1\beta_1\eta_2K_4K_5K_6K_7K_8 + \mu\beta_1\gamma_0\eta_4K_4K_5K_6K_7K_8 \\ & +\mu\gamma_3\beta_2\varphi K_2K_3K_5K_6K_8 + \mu\beta_1\left(\eta_4+\mu\right)K_4K_5K_6K_7K_8K_9 \\ & +\mu\beta_1\gamma_1\left(\gamma_0+\mu\right)K_4K_5K_6K_7K_8 + \lambda_2\beta_3\left(\delta_1+\mu\right)K_2K_3K_4K_5K_8K_9 \\ & +\lambda_2\lambda_3\beta_3\left(\delta_2+\mu\right)K_2K_3K_4K_5K_9 + \varphi\beta_2\left(\delta_1+\mu\right)K_2K_3K_5K_6K_8K_9 \\ & +\lambda_3\varphi\beta_2\left(\delta_2+\mu\right)K_2K_3K_5K_6K_9. \end{aligned} \tag{7}$$

3 Stability analysis

In this section, we study the asymptotic stability of the equilibrium points of our system (2) in which the Jacobian matrix is given by

$$J\left(S_n, E_n, V_n, I_n^a, I_n^{s_1}, I_n^{s_2}, I_n^{h_1}, I_n^{h_2}, I_n^{h_3}, R_n\right) = \left(M_{i,j}\right)_{i,j=\overline{1,10}}$$

such as

$$\begin{cases} M_{1,1} = -K_0 - \alpha\left(I_n^a + I_n^{s_1} + I_n^{s_2}\right), \\ M_{2,1} = \alpha(I_n^a + I_n^{s_1} + I_n^{s_2}), \\ M_{3,1} = \eta_1, \\ M_{2,2} = -K_1, \\ M_{3,2} = \eta_3, \\ M_{4,2} = \beta_1, \\ M_{5,2} = \beta_2, \\ M_{7,2} = \beta_3, \\ M_{1,3} = \eta_2, \\ M_{2,4} = M_{2,5} = M_{2,6} = \alpha S_n, \\ M_{1,4} = M_{1,5} = M_{1,6} = -\alpha S_n, \end{cases} \quad \begin{cases} M_{3,3} = -K_2, \\ M_{3,4} = \eta_4, \\ M_{4,4} = -K_3, \\ M_{10,3} = \gamma_0, \\ M_{5,5} = -K_4, \\ M_{6,5} = \lambda_1, \\ M_{10,4} = \gamma_1, \\ M_{6,6} = -K_5, \\ M_{8,5} = \varphi, \\ M_{7,7} = -K_6, \end{cases} \quad \begin{cases} M_{8,7} = \lambda_2, \\ M_{10,6} = \gamma_2, \\ M_{8,8} = -K_7, \\ M_{9,8} = \lambda_3, \\ M_{10,8} = \gamma_3, \\ M_{9,9} = -K_8, \\ M_{10,9} = \gamma_4, \\ M_{1,10} = \theta, \\ M_{10,10} = -K_9, \\ \\ 0 \qquad \text{otherwise}, \end{cases}$$

where $(K_i)_{i=0,\ldots,9}$ are positive constants given in (6).

3.1 Asymptotic stability of DFE E_0

Theorem 1. The DFE of the system (2) is locally-asymptotically stable in Ω whenever $R_v \leqslant 1$.

Proof. The system (2) has a unique DFE described in (4).

The Jacobian matrix at E_0 is given by

$$J\left(E_0\right) = \left(M_{i,j}\right)_{i,j=\overline{1,10}},$$

where

$$\begin{cases} M_{1,1} = -K_0, \\ M_{3,1} = \eta_1, \\ M_{2,2} = -K_1, \\ M_{3,2} = \eta_3, \\ M_{4,2} = \beta_1, \\ M_{5,2} = \beta_2, \\ M_{7,2} = \beta_3, \\ M_{1,3} = \eta_2, \\ M_{1,4} = M_{1,5} = M_{1,6} = \dfrac{-\alpha K_9 K_2}{K_9 K_2 + \eta_1 K_9 + \eta_1 \gamma_0}, \\ M_{2,4} = M_{2,5} = M_{2,6} = \dfrac{\alpha K_9 K_2}{K_9 K_2 + \eta_1 K_9 + \eta_1 \gamma_0}, \end{cases} \begin{cases} M_{3,3} = -K_2, \\ M_{3,4} = \eta_4, \\ M_{4,4} = -K_3, \\ M_{10,3} = \gamma_0, \\ M_{5,5} = -K_4, \\ M_{6,5} = \lambda_1, \\ M_{10,4} = \gamma_1, \\ M_{6,6} = -K_5, \\ M_{8,5} = \varphi, \\ M_{7,7} = -K_6, \end{cases} \begin{cases} M_{8,7} = \lambda_2, \\ M_{10,6} = \gamma_2, \\ M_{8,8} = -K_7, \\ M_{9,8} = \lambda_3, \\ M_{10,8} = \gamma_3, \\ M_{9,9} = -K_8, \\ M_{10,9} = \gamma_4, \\ M_{1,10} = \theta, \\ M_{10,10} = -K_9, \\ \\ 0 \quad \text{otherwise.} \end{cases}$$

The characteristic polynomial of $J(E_0)$ is given by,

$$P(X) = X^{10} + C_9 X^9 + C_8 X^8 + C_7 X^7 + C_6 X^6 + C_5 X^5 + C_4 X^4 + C_3 X^3 + C_2 X^2 + C_1 X + C_0.$$

If $R_v \leqslant 1$, then the coefficients $(C_i)_{i=\overline{0,9}}$ are positives (Details are provided in Appendix 6). According to the criteria of Descartes, all eigenvalues of $J(E_0)$ are negative reals or complexes with negative real parts. Hence, E_0 is locally asymptotically stable in Ω. □

3.2 Asymptotic stability of the endemic equilibrium E_1

Theorem 2. The endemic equilibrium of our system (2), E_1, is locally-asymptotically stable in Ω whenever $R_v > 1$.

Proof. The system (2) has a unique endemic equilibrium E_1 in Ω, given in 5.

The Jacobian matrix at E_1 is given by

$$J(E_1) = (M_{i,j})_{i,j=\overline{1,10}}$$

such as

$$\begin{cases} M_{1,1} = -\left(K_0 + \alpha\left(I^{a^{**}} + I^{s_1^{**}} + I^{s_2^{**}}\right)\right), \\ M_{2,1} = \alpha\left(I^{a^{**}} + I^{s_1^{**}} + I^{s_2^{**}}\right), \\ M_{3,1} = \eta_1, \\ M_{2,2} = -K_1, \\ M_{3,2} = \eta_3, \\ M_{4,2} = \beta_1, \\ M_{5,2} = \beta_2, \\ M_{7,2} = \beta_3, \\ M_{1,4} = M_{1,5} = M_{1,6} = -\alpha S^{**}, \\ M_{2,4} = M_{2,5} = M_{2,6} = \alpha S^{**}, \end{cases} \begin{cases} M_{1,3} = \eta_2, \\ M_{3,3} = -K_2, \\ M_{3,4} = \eta_4, \\ M_{4,4} = -K_3, \\ M_{10,3} = \gamma_0, \\ M_{5,5} = -K_4, \\ M_{6,5} = \lambda_1, \\ M_{10,4} = \gamma_1, \\ M_{6,6} = -K_5, \\ M_{8,5} = \varphi, \\ M_{7,7} = -K_6, \end{cases} \begin{cases} M_{8,7} = \lambda_2, \\ M_{10,6} = \gamma_2, \\ M_{8,8} = -K_7, \\ M_{9,8} = \lambda_3, \\ M_{10,8} = \gamma_3, \\ M_{9,9} = -K_8, \\ M_{10,9} = \gamma_4, \\ M_{1,10} = \theta, \\ M_{10,10} = -K_9, \\ \\ 0 \text{ otherwise}, \end{cases}$$

where

$$\begin{aligned} S^{**} &= \frac{1}{\alpha}\frac{K_1K_3K_4K_5}{\beta_2K_3\left(K_5+\lambda_1\right)+\beta_1K_4K_5}, \\ I^{a^{**}} &= \mu\beta_1\frac{K_2K_4K_5K_6K_7K_8K_9}{K}\left(1-\frac{1}{R_v}\right), \\ I^{s_1^{**}} &= \mu\beta_2\frac{K_2K_3K_5K_6K_7K_8K_9}{K}\left(1-\frac{1}{R_v}\right), \\ I^{s_2^{**}} &= \mu\lambda_1\beta_2\frac{K_2K_3K_6K_7K_8K_9}{K}\left(1-\frac{1}{R_v}\right). \end{aligned}$$

The characteristic polynomial of $J\left(E_1\right)$ is given by

$$P\left(X\right) = X^{10}+C_9X^9+C_8X^8+C_7X^7+C_6X^6+C_5X^5+C_4X^4+C_3X^3+C_2X^2+C_1X+C_0$$

If $R_v > 1$, then the coefficients $\left(C_i\right)_{i=\overline{0,9}}$ are positives (Details are provided in Appendix 6). According to the criteria of Descartes, all eigenvalues of $J\left(E_1\right)$ are negative reals or complexes with negative real parts. Hence, E_1 is locally asymptotically stable in Ω. □

4 Sensitivity analysis

To effectively reduce the transmission and prevalence of COVID-19, it is crucial to examine and understand the relative significance of various factors that contribute to their evolution. The vaccination reproduction number,

denoted as R_v, is a key parameter that directly affects disease transmission control. It measures the average number of cases an infected person generates during their infectious period in a population that is fully vaccinated. The long-term disease prevalence is represented by the endemic point E_1. Conducting sensitivity analyses on these indicators can help identify the operational parameters that have a substantial impact on them [5, 20]. This knowledge can then be used to develop effective intervention strategies.

The normalized forward sensitivity index of a variable, f, that is differentially dependent on a parameter, x, is defined as:

$$\Upsilon_x^f = \frac{\partial f}{\partial x}\frac{x}{f}. \tag{8}$$

This section focuses on conducting local sensitivity analyses. Initially, sensitivity indices of R_v are calculated, followed by sensitivity indices of E_1, using parameter values obtained from various sources listed in Table 3.

4.1 Sensitivity index of R_v

The formula for calculating R_v is explicitly given in (3). By using the analytical formulas provided in reference (8), sensitivity indices for R_v can be computed and presented in Table 1, using the parameter values from Table 3.

The sensitivity indices presented in Table 1 are sorted in descending order, indicating that the transmission rate α and the recovery rate of infected with moderate symptoms γ_2 are the most sensitive parameters. For instance, a 10% decrease (increase) in the transmission rate α results in a 10% decrease (increase) in R_v. Similarly, a 10% decrease (increase) in γ_2 leads to a 9.404% increase (decrease) in R_v. Therefore, it is crucial to accurately estimate these parameters, as even a small variation in them can cause significant quantitative changes. On the other hand, the lowest sensitive parameters, such as the recovery rate of vaccinated individuals γ_0 and the reinfection rate of vaccinated individuals η_2, have a negligible effect on R_v. A 100% increase (or decrease) in γ_0 or η_2 leads to only a 0.047043% or 0.041779% increase (or decrease) in R_v, respectively.

Table 1: Sensitivity indices of R_v. The parameters are ordered from most sensitive to least

Parameters	Sensitivity index
α	1.0000
γ_2	-0.9404
β_2	0.6249
β_1	-0.6125
γ_1	-0.0225
λ_1	-0.0202
η_3	-0.0104
η_4	-0.0040
β_3	-0.0018
φ	-0.0018
η_1	-0.0017
θ	0.00066625
γ_0	0.00047043
η_2	0.00041779

4.2 Sensitivity index of E_1

The endemic equilibrium for the COVID-19 system (2), is given by (5). Using (8), the same approach can be followed to determine the sensitivity indices of each of the endemic equilibrium state variables.

Table 2 reveals that the most sensitive parameters for all infected states are the transition rates $(\lambda_i)_{i=1,2,3}$, rates at which exposed individuals become infected $(\beta_i)_{i=1,2,3}$, and recovery rates $(\gamma_i)_{i=1,2,3,4}$, respectively. Therefore, any variations in these parameters will significantly impact the proportion of infectious individuals. The reinfection rate θ is another parameter that strongly and positively affects the prevalence of COVID-19, and it is the most sensitive parameter for the recovery state. Increasing this parameter of 10 percent results in a decrease of 9,466 percent in the proportion of recovered individuals, leading to an increased risk of COVID-19 spread. Lastly, the parameters η_4 (vaccination rate of asymptomatically infected individuals) and γ_2 (recovery rate of symptomatic infected) increase the proportion of

Table 2: Sensitivity indices of state variables at endemic equilibrium for baseline parameter values in Table 3

	S^{**}	E^{**}	V^{**}	$I^{a^{**}}$	$I^{s_1^{**}}$	$I^{s_2^{**}}$	$I^{H_1^{**}}$	$I^{H_2^{**}}$	$I^{H_3^{**}}$	R^{**}
α	-1	0.2800	0.2800	0.2800	0.2800	0.2800	0.2800	0.2800	0.2800	0.2734
β_1	0.6125	-0.2254	0.6763	**0.7746**	-0.2254	-0.2254	-0.2254	-0.2254	-0.2254	0.4102
β_2	-0.6249	-0.7206	-0.7206	**-0.7206**	0.2794	**-0.7206**	**-0.7206**	-0.4531	-0.4531	-0.3663
β_3	0.0018	-0.0069	-0.0069	-0.0069	-0.0069	**0.9931**	**0.9931**	**0.7256**	**0.7256**	-0.0051
η_1	0	0.0047	-0.000468	-0.000468	-0.000468	-0.000468	-0.000468	-0.000468	-0.000468	0.0047
η_2	0	-0.0198	-0.2479	0.0020	0.0020	0.0020	0.0020	0.0020	0.0020	-0.0198
η_3	0.0104	0.0045	0.0946	-0.0036	-0.0036	-0.0036	-0.0036	-0.0036	-0.0036	0.0045
η_4	0.0040	-0.0206	**0.7669**	-0.1495	0.000301	0.000301	0.000301	0.000301	0.000301	-0.0206
γ_0	0	0.0022	-0.7478	0.0022	0.0022	0.0022	0.0022	0.0022	0.0022	0.0239
γ_1	0.0225	0.0159	-0.7506	**-0.8341**	0.0159	0.0159	0.0159	0.0159	0.0159	0.0369
γ_2	0.9404	0.5741	0.5741	**0.5741**	-0.4145	-0.4145	**0.5741**	**0.5741**	**0.5741**	0.5800
γ_3	0	0.0036	0.0036	0.0036	0.0036	0.0036	0.0036	**-0.7376**	**-0.7376**	0.0034
γ_4	0	0.0025	0.0025	0.0025	0.0025	0.0025	0.0025	0.0025	**-0.9775**	0.0025
λ_1	0.0202	0.0146	0.0146	0.0146	**-0.9833**	0.0146	0.0146	-0.2523	-0.2523	0.0146
λ_2	0	0.000061	0.000061	0.000061	**1.0001**	**-0.9998**	**-0.9998**	0.000188	0.000188	0.000061
λ_3	0	0.0024	0.0024	0.0024	0.0024	0.0024	0.0024	-0.2445	**0.7555**	0.0025
δ_1	0	-0.0058	-0.0058	-0.0058	-0.0058	-0.0058	-0.0058	-0.0177	-0.0177	-0.0058
δ_2	0	-0.0025	-0.0025	-0.0025	-0.0025	-0.0025	-0.0025	-0.0025	-0.0223	-0.0025
θ	0	0.0535	0.0535	0.0535	0.0535	0.0535	0.0535	0.0535	0.0535	**-0.9466**
φ	0.0018	-0.0010	-0.0012	-0.0012	-0.0031	-0.0031	-0.0012	0.2658	0.2658	-0.0010

vaccinated individuals. Therefore, the population becomes immune to the disease, ultimately leading to the eradication of COVID-19.

5 Numerical simulation

This section aims to provide numerical simulations of the proposed model to demonstrate its practicality and effectiveness, and to illustrate our theoretical findings. The operational parameters used in the simulations are in Table 3.

Given that the model is discrete, the transitions between compartments are governed by the following rules:

- For the compartment of exposed individuals, those with a negative test result who will be vaccinated are represented by η_3, while the rest of the exposed individuals are divided into asymptomatic ($\beta_1 = 0.55 \times (1 - \eta_3 - \mu)$), symptomatic ($\beta_2 = 0.30 \times (1 - \eta_3 - \mu)$), and hospitalized ($\beta_3 = 0.15 \times (1 - \eta_3 - \mu)$) individuals, according to [18].

Table 3: The values of operational parameters per week

Parameters	Description	Range of values	Source
α	Transmission rate.	0.002-0.2	[7,14]
β_1	Asymptomatic infection rate among exposed.	0.4-0.55	[7,14]
β_2	Rate of infection with symptoms among exposed.	0.2-0.3	[7,14]
β_3	Rate of hospitalization among exposed individuals.	0.001-0.15	[7,14]
η_1	Vaccination rate of susceptible individuals.	0.00001-0.2	[7,14]
η_2	Co-infection rate of vaccinated individuals.	0.2-0.35	[7,14]
η_3	Vaccination rate of exposed individuals.	0.0001-0.2	[7,14]
η_4	Vaccination rate of asymptomatically infected.	0.1-0.55	[7,14]
γ_0	Recovery rate of vaccinated individuals.	0.65-0.75	[7,14]
γ_1	Recovery rate of asymptomatic infected.	0.45-0.85	[7,14]
γ_2	Recovery rate of symptomatic infected.	0.015-1.0	[7,14]
γ_3	Recovery rate of hospitalized infected in 2nd phase.	0.2-0.75	[7,14]
γ_4	Recovery rate of hospitalized infected in 3rd phase.	0.95-0.98	[7,14]
λ_1	Evolution rate symptomatic phase 1 to 2.	0.6-0.7	[7,14]
λ_2	Evolution rate hospitalized phase 1 to 2.	0.95-1.0	[7,14]
λ_3	Evolution rate hospitalized phase 2 to 3.	0.2-0.8	[7,14]
δ_1	Hospitalized COVID-19 mortality, phase 2.	0.01-0.02	[3,7,14]
δ_2	Hospitalized COVID-19 mortality, phase 3.	0.015-0.05	[3,7,14]
θ	Coinfection rate of recovery individuals	0.55-0.85	[7,14]
φ	The hospitalization rate of symptomatic in phase 1.	0.001-0.35	[3,7,14]
μ	Natural mortality rate	0.0001-0.002	[7,14]

- The efficacy of the vaccine is represented by γ_0 in the compartment of vaccinated individuals, with $\eta_2 = (1 - \gamma_0 - \mu)$ representing the proportion of vaccinated individuals who are not protected by the vaccine.

- In the compartment of asymptomatic infected individuals, the recovery rate is represented by γ_1, and $\eta_4 = (1-\gamma_1-\mu)$ represents the proportion of asymptomatic infected individuals who will be vaccinated.

- For the compartment of symptomatic infected individuals not requiring hospitalization, the proportion of individuals entering the second week of illness is represented by λ_1, while $\varphi = (1 - \lambda_1 - \mu)$ represents the proportion of individuals whose condition worsens and requires hospitalization.

- In the compartment of hospitalized infected individuals in the second phase, the recovery rate is represented by γ_3, λ_3 represents the proportion of individuals who remain hospitalized, and the mortality rate due to COVID-19 is represented by $\delta_1 = 1 - \gamma_3 - \lambda_3 - \mu$.
- Finally, in the compartment of hospitalized infected individuals in the final phase, the recovery rate is represented by γ_4, and the mortality rate due to COVID-19 is represented by $\delta_2 = 1 - \gamma_4 - \mu$.

5.1 Numerical stability

In this subsection, we verify the theoretical results obtained above through numerical simulations.

1. The following parameter values (per week) are used:

$$\begin{aligned} &\alpha = 0.0025, \quad \gamma_0 = 0.65, \quad \gamma_1 = 0.45, \quad \gamma_3 = 0.2, \\ &\gamma_4 = 0.95, \quad \eta_1 = 0.2, \quad \eta_3 = 0.2, \quad \theta = 0.55, \quad \lambda_3 = 0.78. \end{aligned} \tag{9}$$

   The existence and stability conditions in theorem 1 are satisfied,

$$R_v = 0.0015 < 1.$$

   Therefore, the system (2) has a DFE in Ω,

$$E_0 = (0.6962, 0, 0.1392, 0, 0, 0, 0, 0, 0, 0.1645)\,.$$

   Figure 2 indicates under some conditions that the long-term behavior of the solution curves of the system approaches the free equilibrium E_0, as revealed by numerical simulations. This suggests that the population under consideration gradually becomes free of the infection.

   Based on the numerical simulations depicted in Figure 2, it can be observed that the behavior of the solution curves of the system approaches the DFE E_0 over a long period of time. This simulation confirms the theoretical results outlined in section 3.

2. The following parameter values (per week) are used:

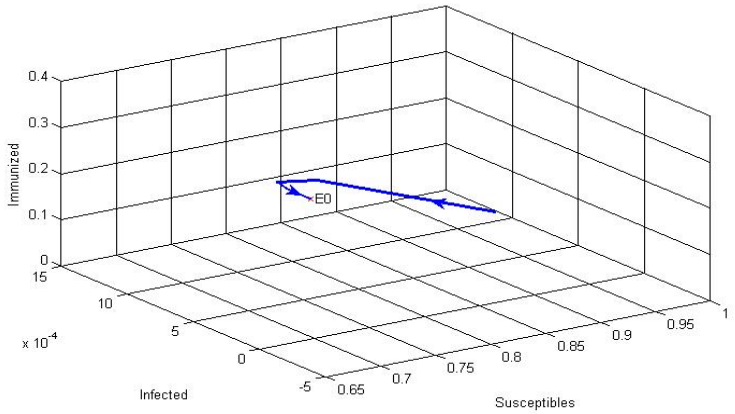


Figure 2: The proposed model has only one DFE E_0, which is asymptotically stable when $R_v \leqslant 1$.

$$\alpha = 0.19, \qquad \gamma_0 = 0.75, \qquad \gamma_1 = 0.85, \quad \gamma_3 = 0.75, \quad \gamma_4 = 0.98, \tag{10}$$
$$\eta_1 = 0.00089, \quad \eta_3 = 0.0089, \quad \theta = 0.85, \quad \lambda_3 = 0.24.$$

The existence and stability conditions in Theorem 2 are satisfied,

$$R_v = 4.5711 > 1.$$

There is only one endemic equilibrium point in Ω,

$$E_1 = \Big(0.21840, \quad 0.03385, \quad 0.00326, \quad 0.018451, \quad 0.015451,$$
$$0.661920, \quad 0.0000507, \quad 0.000068, \quad 0.000017, \quad 0.033088\Big).$$

Numerical simulations in Figure 3 demonstrate that the solution curves of the system approach the endemic equilibrium E_1 over an extended period.

Our proposed discrete model for coronavirus disease 2019 underwent numerical processing, which confirmed the existence of two equilibrium points. The first is the DFE E_0, which is locally asymptotically stable when $R_v \leqslant 1$.

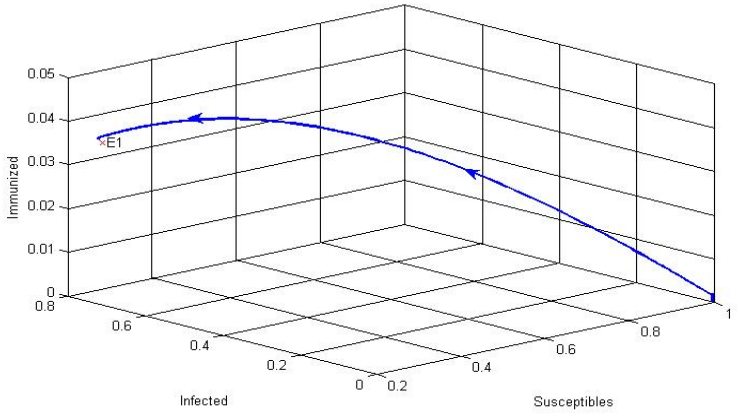


Figure 3: The proposed model has only one endemic equilibrium E_1, which is asymptotically stable when $R_v > 1$.

The second is the endemic equilibrium E_1, which is locally asymptotically stable when $R_v > 1$.

5.2 The effect of the most sensitive parameters of the model

It is important to note that the transmission rate α is mainly influenced by non-pharmaceutical measures such as physical distancing, the use of masks, quarantine and isolation of infected or exposed individuals, as well as limitations on nonessential gatherings and travel. Medical care is also an important factor that can impact the recovery rate γ_0, γ_1, γ_3, and γ_4 for COVID-19. This includes access to appropriate medical facilities, equipment, and trained health-care professionals. The availability of treatments and medications that can effectively manage COVID-19 symptoms can also impact the recovery rate by reducing the duration and severity of the illness.

The primary objective of the simulations was to observe how the model parameters most affected the transmission dynamics and spread of COVID-19. Table 3 values were used in conjunction with varying levels of α, γ_0, γ_1,

γ_3, and γ_4, which were increased in stages, ranging from mild to moderate and strict rates. The results of the simulations provide insight into the impact of control measures on the spread of the disease. By analyzing the effects of different levels, policymakers and public health officials can make informed decisions regarding the implementation of measures to curb the spread of COVID-19.

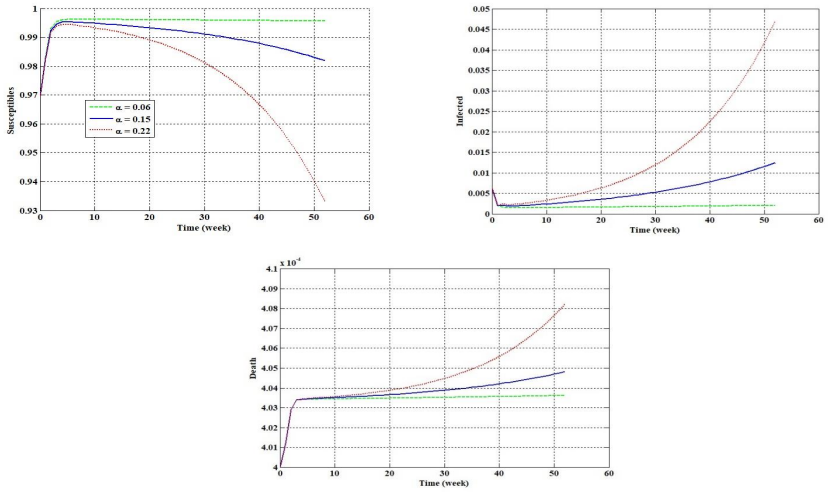


Figure 4: Effect of α transmission rate on different states of COVID-19.

Figures 4 and 5 present the results of numerical simulations conducted to investigate the effects of model parameters on the transmission dynamics and spread of COVID-19. The simulations were performed using different values of α, $\gamma_0, \gamma_1, \gamma_3$, and γ_4, and the results provide valuable insights into the impact of control measures on the spread of the disease.

Figure 4 shows that the rate of transmission is a critical factor in determining the number of people infected with COVID-19. The graph indicates that the higher the transmission rate α, the greater the number of infected individuals. When the transmission rate increases from the lowest 0.06 to the highest 0.22, the number of infections increases by more than 4 percent, as well as the COVID-19 mortality increases by $0.047e-2$ percent. This finding highlights the importance of implementing effective measures to control the

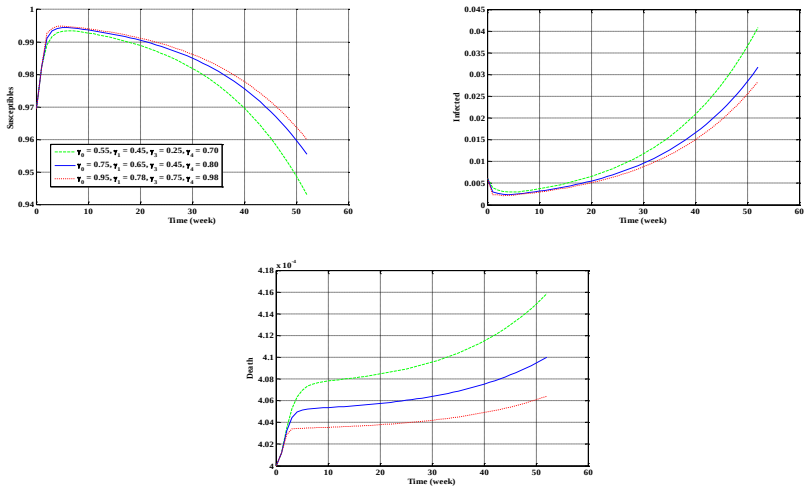


Figure 5: Effective recovery rates γ_0, γ_1, γ_3, and γ_4 on the number of COVID-19 infected cases.

spread of the virus, such as social distancing, wearing masks, and limiting large gatherings.

Figure 5 examines the impact of the recovery rates $\gamma_0, \gamma_1, \gamma_3$, and γ_4 on the number of infections. The graph shows that increasing the recovery rates results in a reduction in the number of infections. However, the effect of $\gamma_0, \gamma_1, \gamma_3$, and γ_4 is less pronounced than that of α. When the values of $\gamma_0, \gamma_1, \gamma_3$, and γ_4 are changed from the highest to the lowest, the difference in the number of infections is around 1.5 percent, as well as the mortality from COVID-19 decreases by $0.10e-2$ percent. This finding suggests that while recovery rates are essential in reducing the spread of the disease, they may not be sufficient on their own. A combination of measures, including vaccination, testing, and contact tracing, may be necessary to achieve effective control of the virus.

The results presented in Figures 4 and 5 are consistent with the sensitivity analysis results obtained earlier. The sensitivity analysis revealed that the transmission rate α had a significant impact on the number of infections, while the recovery rates $\gamma_0, \gamma_1, \gamma_3$, and γ_4 had a smaller effect. The simula-

tions conducted in Figures 4 and 5 confirm these findings and provide further evidence of the importance of controlling the transmission rate.

Overall, the numerical simulations provide important insights into the transmission dynamics and spread of COVID-19. By varying the values of α, $\gamma_0, \gamma_1, \gamma_3$, and γ_4, the simulations show how different control measures can affect the spread of the virus. The results indicate that while recovery rates are important in reducing the number of infections, the transmission rate is the primary determinant of the spread of the disease.

In conclusion, the numerical simulations presented in Figures 4 and 5 provide valuable insights into the transmission dynamics and spread of COVID-19. The simulations confirm the importance of controlling the transmission rate and suggest that recovery rates alone may not be sufficient in reducing the spread of the virus. The findings of the simulations can inform the development of effective control measures to curb the spread of COVID-19 and help mitigate the impact of the pandemic on public health and the economy.

6 Conclusion

In this research, we have conducted a comprehensive analysis of the factors affecting the transmission dynamics of COVID-19, employing both sensitivity analysis and numerical simulations. Our results shed light on the fundamental mechanisms governing the spread of the virus, offering critical insights for pandemic management and public health policy.

Sensitivity analysis revealed that the transmission rate α and the recovery rate of individuals with moderate symptoms (γ_2) are highly influential parameters. Small changes in these values can result in substantial alterations in the course of the pandemic. This underscores the imperative of precise parameter estimation and the significance of control measures that directly impact these variables.

Furthermore, numerical simulations provided a practical perspective, reinforcing the pivotal role of the transmission rate. The simulations demonstrated that variations in α have a profound effect on the number of infections and, subsequently, the COVID-19 mortality rate. This underscores the importance of strict public health interventions such as social distancing, mask-

wearing, and restrictions on gatherings to control viral transmission. While the recovery rates γ_0, γ_1, γ_3, and γ_4 also play a role in reducing infections, their impact is less pronounced compared to α. This suggests that a multi-faceted approach, encompassing vaccination, testing, and contact tracing, is essential to effectively manage the pandemic.

In sum, our study emphasizes the critical need for tailored public health strategies, underpinned by a thorough understanding of the varying impacts of different parameters. These findings are pertinent to real-world policy decisions, guiding efforts to mitigate the consequences of the COVID-19 pandemic on both public health and the economy.

Appendix

The following calculations were performed using the mathematical software (Maxima Version 5.42.1).

The characteristic polynomial of $J(E_0)$ is given by

$$P(X) = X^{10} + C_9X^9 + C_8X^8 + C_7X^7 + C_6X^6 + C_5X^5 + C_4X^4 + C_3X^3 + C_2X^2 + C_1X + C_0.$$

So we have

$$\eta_1\eta_2 < (\eta_1+\mu)(\eta_2+\gamma_0+\mu) < K_0K_2, \tag{11}$$

$$\begin{aligned} &\alpha S^*(\beta_1(K_4+K_5)+\beta_2(K_3+K_5)+\lambda_1\beta_2) \qquad (12)\\ &= \alpha S^*\left(\beta_1K_4K_5\left(\frac{1}{K_4}+\frac{1}{K_5}\right)+\beta_2K_3K_5\left(\frac{1}{K_3}+\frac{1}{K_5}\right)+\lambda_1\beta_2K_3\frac{1}{K_3}\right)\\ &< \alpha S^*(\beta_1K_4K_5+\beta_2K_3K_5+\lambda_1\beta_2K_3)\left(\frac{1}{K_3}+\frac{1}{K_4}+\frac{1}{K_5}\right)\\ &< K_1K_3K_4K_5R_v\left(\frac{1}{K_3}+\frac{1}{K_4}+\frac{1}{K_5}\right)\\ &< (K_1K_3K_4+K_1K_3K_5+K_1K_4K_5)R_v, \end{aligned}$$

and

$$\begin{aligned}\alpha S^*\left(\beta_1+\beta_2\right) &= \alpha S^*\left(\beta_1 K_4K_5\frac{1}{K_4K_5}+\beta_2K_3K_5\frac{1}{K_3K_5}\right) \\ &< \alpha S^*\left(\beta_1K_4K_5+\beta_2K_3K_5+\lambda_1\beta_2K_3\right)\left(\frac{1}{K_4K_5}+\frac{1}{K_3K_5}\right) \\ &< K_1K_3K_4K_5R_v\left(\frac{1}{K_4K_5}+\frac{1}{K_3K_5}\right) \\ &< \left(K_1K_3+K_1K_4\right)R_v.\end{aligned} \tag{13}$$

We pose

$$\begin{cases}\mathbf{P}_1 = K_0K_2-\eta_1\eta_2, \\ \mathbf{P}_2 = \left(K_1K_3K_4+K_1K_3K_5+K_1K_4K_5\right) \\ \qquad -\alpha S^*\left(\beta_1\left(K_4+K_5\right)+\beta_2\left(K_3+K_5+\lambda_1\right)\right), \\ \mathbf{P}_3 = \left(K_1K_3+K_1K_4\right)-\alpha S^*\left(\beta_1+\beta_2\right).\end{cases}$$

The coefficients of the characteristic polynomial are given by

$$C_0 = \frac{\mu}{S^*}K_1K_2K_3K_4K_5K_6K_7K_8K_9\left(1-R_v\right),$$

$$\begin{aligned}C_1 = &\frac{\mu}{S^*}K_2K_6K_7K_8K_9\left(K_3K_4K_5+\mathbf{P}_2\right) \\ &+K_1K_3K_4K_5K_6K_7K_8\left(K_0K_9+K_2K_9+\mathbf{P}_1\right)\left(1-R_v\right) \\ &+\frac{\mu}{S^*}K_1K_2K_3K_4K_5K_9\left(K_6K_7+K_6K_8+K_7K_8\right)\left(1-R_v\right),\end{aligned}$$

$$\begin{aligned}C_2 = &K_6K_7K_8K_9\left(K_1K_3K_4+K_1K_3K_5+K_1K_4K_5+K_3K_4K_5\right)\left(K_0+K_2\right) \\ &+K_1K_3K_4K_5K_8K_9\left(K_6K_7+K_2K_7+K_0K_7+K_0K_6+K_2K_6\right)\left(1-R_v\right) \\ &+K_1K_3K_4K_5K_6K_7\left(K_0K_9+K_2K_9+K_0K_8+K_2K_8\right)\left(1-R_v\right) \\ &+\frac{\mu}{S^*}K_1K_2K_3K_4K_5K_9\left(K_6+K_7+K_8\right)\left(1-R_v\right)+K_6K_7K_8\mathbf{P}_1\mathbf{P}_2 \\ &+\frac{\mu}{S^*}K_2K_9\left(K_1K_5K_6K_7K_8+K_3K_5K_6K_7K_8+K_4K_5K_6K_7K_8\right) \\ &+\frac{\mu}{S^*}K_2K_3K_4K_9\left(K_5K_6K_7+K_5K_6K_8+K_5K_7K_8+K_6K_7K_8\right) \\ &+\frac{\mu}{S^*}K_2K_9\left(K_6K_7+K_6K_8+K_7K_8\right)\mathbf{P}_2+K_6K_7K_8K_9\left(K_0+K_2\right)\mathbf{P}_2 \\ &+K_1K_3K_4K_5\left(K_6K_7+K_6K_8+K_8K_7\right)\mathbf{P}_1\left(1-R_v\right) \\ &+K_3K_4K_5K_6K_7K_8\mathbf{P}_1+\frac{\mu}{S^*}K_2K_6K_7K_8K_9\mathbf{P}_3,\end{aligned}$$

$$
\begin{aligned}
C_3 = {} & K_1K_3K_4K_5\left(K_6K_7\left(K_8+K_9\right)+\left(K_6+K_7\right)K_8K_9+\frac{\mu}{S^*}K_2K_9\right)\left(1-R_v\right)\\
& +K_5K_6K_7K_8K_9\left(\left(K_1+K_3+K_4\right)\left(K_0+K_2\right)+K_3K_4\right)+K_6K_7K_8\mathbf{P}_3\mathbf{P}_1\\
& +K_1K_3K_4K_5\left(K_8\left(K_7+K_6+K_9\right)+\left(K_7+K_6\right)K_9\right)\left(K_0+K_2\right)\left(1-R_v\right)\\
& +K_3K_4K_5K_6K_8K_9\left(K_0+K_2\right)+K_1K_3K_4K_5K_6K_7\left(K_0+K_2\right)\left(1-R_v\right)\\
& +K_3K_4\left(K_7K_8\left(K_6K_9+K_5K_9\right)+K_6K_7\left(K_5K_9+K_5K_8\right)\right)\left(K_0+K_2\right)\\
& +\frac{\mu}{S^*}K_2K_9\left(K_1K_5K_6K_8+K_3K_4K_5K_8+K_1K_5K_7K_8+K_3K_4K_6K_8\right)\\
& +\frac{\mu}{S^*}K_2K_9\left(K_7K_8+K_6K_8+K_7K_6\right)\mathbf{P}_3+K_6K_7K_8K_9\left(K_0+K_2\right)\mathbf{P}_3\\
& +\frac{\mu}{S^*}K_2K_9\left(K_7K_8\left(K_1K_6+K_3K_4+K_3K_5\right)+K_5K_6K_8\left(K_3+K_4\right)\right)\\
& +\frac{\mu}{S^*}K_2K_9\left(K_6K_7K_8\left(K_3+K_4+K_5\right)+K_4K_5K_7K_8+K_3K_4K_5K_6\right)\\
& +\frac{\mu}{S^*}K_2K_9\left(K_5K_6K_7\left(K_1+K_3+K_4\right)+K_3K_4K_5K_7+K_3K_4K_6K_7\right)\\
& +\frac{\mu}{S^*}K_2K_9\left(K_8+K_7+K_6\right)\mathbf{P}_2+K_1K_3K_4K_5\left(K_8+K_7+K_6\right)\mathbf{P}_1\left(1-R_v\right)\\
& +\left(K_0K_7\left(K_6K_8+K_6K_9+K_8K_9\right)+K_0K_6K_8K_9+K_2K_6K_7\left(K_8+K_9\right)\right)\mathbf{P}_2\\
& +\left(K_7K_8+K_6K_8+K_6K_7\right)\mathbf{P}_1\mathbf{P}_2+K_8K_9\left(K_2K_6+K_2K_7+K_6K_7\right)\mathbf{P}_2\\
& +K_5K_6K_7K_8\left(K_3+K_4+K_1\right)\mathbf{P}_1+K_7K_3K_4\left(K_5K_6+K_5K_8+K_6K_8\right)\mathbf{P},_1
\end{aligned}
$$

$$
\begin{aligned}
C_4 = {} & K_3K_9\left(K_7K_8\left(K_6+K_5+K_4\right)+K_8\left(K_5K_6+K_4K_6+K_5K_4\right)\right)\left(K_0+K_2\right)\\
& +K_4\left(K_8K_9\left(K_6K_7+K_5K_7+K_5K_6\right)+K_5K_6K_7\left(K_9+K_8\right)\right)\left(K_0+K_2\right)\\
& +K_3\left(K_6K_7\left(K_5K_8+K_4K_8+K_4K_5\right)+K_4K_5K_8\left(K_7+K_6\right)\right)\left(K_0+K_2\right)\\
& +K_1\left(K_8K_9\left(K_6K_7+K_5K_7+K_5K_6\right)+K_5K_6K_7\left(K_9+K_8\right)\right)\left(K_2+K_0\right)\\
& +\frac{\mu}{S^*}K_2K_9K_8K_7\left(K_6+K_5+K_4+K_3+K_1\right)+K_1K_3K_4K_5\left(1-R_v\right)\mathbf{P}_1\\
& +K_1K_3K_4K_5\left(\left(K_8+K_7+K_6\right)K_9+\left(K_7+K_6\right)K_8+K_6K_7\right)\left(1-R_v\right)\\
& +K_7K_8K_9\left(K_1K_5K_6+K_3K_5K_6+K_4K_5K_6+K_3K_4K_5+K_3K_4K_6\right)\\
& +K_3K_4K_5K_6\left(K_7K_8+K_7K_9+K_8K_9\right)+K_5K_6K_7K_8K_9\left(K_0+K_2\right)\\
& +K_3K_9\left(K_5K_6K_7+K_4K_6K_7+K_5K_4K_7+K_6K_5K_4\right)\left(K_0+K_2\right)\\
& +\frac{\mu}{S^*}K_2K_9K_6\left(K_8+K_7\right)\left(K_5+K_4+K_3+K_1\right)+K_6K_7K_8K_9\mathbf{P}_3\\
& +K_9\left(\left(K_7K_8+K_6K_8+K_6K_7\right)+\left(K_8+K_7+K_6\right)\left(K_0+K_2\right)\right)\mathbf{P}_2\\
& +\left(K_8+K_7+K_6\right)\mathbf{P}_1\mathbf{P}_2+\frac{\mu}{S^*}K_2K_3K_4K_9\left(K_5+K_6+K_7+K_8\right)\\
& +\left(K_7K_8+K_6K_8+K_6K_7\right)\mathbf{P}_1\mathbf{P}_3+K_3K_4K_5\left(K_8+K_7+K_6\right)\mathbf{P}_1
\end{aligned}
$$

$$
\begin{aligned}
&+ \left(K_7K_8K_9 + K_6K_8K_9 + K_6K_7K_9 + K_8K_7K_6\right)\left(K_2 + K_0\right)\mathbf{P}_3 \\
&+ K_6K_8\left(K_1K_5 + K_3K_4 + K_4K_5 + K_3K_7 + K_1K_7 + K_3K_5\right)\mathbf{P}_1 \\
&+ K_7K_8\left(K_3K_4 + K_1K_5 + K_4K_5 + K_4K_6 + K_5K_6 + K_3K_5\right)\mathbf{P}_1 \\
&+ K_6K_7\left(K_1K_5 + K_4K_5 + K_3K_4 + K_3K_5\right)\mathbf{P}_1 + \frac{\mu}{S^*}K_2K_9\mathbf{P}_2 \\
&+ \frac{\mu}{S^*}K_2K_9\left(K_6 + K_7 + K_8\right)\left(K_5\left(K_1 + K_3 + K_4\right) + \mathbf{P}_3\right) \\
&+ K_1K_3K_4K_5\left(K_8 + K_7 + K_9 + K_6\right)\left(K_0 + K_2\right)\left(1 - R_v\right) \\
&+ \left(K_6K_7K_8 + \left(K_7K_8 + K_6K_8 + K_7K_6\right)\left(K_0 + K_2\right)\right)\mathbf{P}_2,
\end{aligned}
$$

$$
\begin{aligned}
C_5 = {} & K_3\left(K_0 + K_2\right)\left(K_7K_6K_8 + \left(K_7K_8 + K_6K_8 + K_7K_6\right)\left(K_5 + K_4\right)\right) \\
&+ K_5\left(K_0 + K_1 + K_2 + K_3 + K_4\right)\left(K_9\left(K_8\left(K_7 + K_6\right) + K_7K_6\right)\right) \\
&+ K_4K_9\left(K_0 + K_2 + K_3\right)\left(K_8\left(K_7 + K_6 + K_5\right) + K_7\left(K_6 + K_5\right)\right) \\
&+ K_4\left(K_0 + K_2 + K_3\right)\left(K_8\left(K_7\left(K_5 + K_6\right) + K_6K_5\right) + K_7K_6K_5\right) \\
&+ K_3K_4K_5\left(K_0 + K_2\right)\left(K_8 + K_7 + K_6\right) + \left(K_8 + K_7 + K_6\right)\mathbf{P}_1\mathbf{P}_3 \\
&+ \left(\left(K_7K_8 + K_6K_8 + K_6K_7\right)K_9 + K_7K_8K_6\right)\mathbf{P}_3 + \frac{\mu}{S^*}K_2K_9\mathbf{P}_3 \\
&+ \left(K_7K_8 + K_6K_8 + K_9\left(K_8 + K_7 + K_6\right) + K_7K_6\right)\left(K_0 + K_2\right)\mathbf{P}_3 \\
&+ K_1K_3K_4K_5\left(K_9 + K_8 + K_7 + K_6 + K_2 + K_0\right)\left(1 - R_v\right) + \mathbf{P}_2\mathbf{P}_1 \\
&+ \frac{\mu}{S^*}K_2K_9\left(K_6\left(K_1 + K_3 + K_4 + K_5\right) + K_5\left(K_1 + K_3 + K_4\right)\right) \\
&+ \left(\left(K_7 + K_6\right)K_8 + \left(K_6 + K_7 + K_8 + K_9\right)\left(K_0 + K_2\right) + K_7K_6\right)\mathbf{P}_2 \\
&+ K_3K_9\left(K_0 + K_2\right)\left(K_8\left(K_7 + K_6\right) + \left(K_8 + K_7 + K_6\right)\left(K_5 + K_4\right)\right) \\
&+ \left(\left(K_8 + K_7 + K_6\right)K_9\right)\mathbf{P}_2 + K_3K_9\left(K_0 + K_2\right)\left(K_7K_6 + K_5K_4\right) \\
&+ K_7\left(\frac{\mu}{S^*}K_2K_9 + K_8\mathbf{P}_1\right)\left(K_1 + K_3 + K_4 + K_5 + K_6\right) \\
&+ K_1\left(K_2 + K_0\right)\left(K_7K_8K_9 + K_6K_5\left(K_9 + K_8 + K_7\right)\right) \\
&+ K_1\left(K_2 + K_0\right)\left(K_8K_9 + K_7K_9 + K_7K_8\right)\left(K_6 + K_5\right) \\
&+ \frac{\mu}{S^*}K_2K_9\left(K_8\left(K_1 + K_3 + K_4 + K_5 + K_6 + K_7\right)\right) \\
&+ \frac{\mu}{S^*}K_2K_3K_4K_9 + K_4K_5K_6K_9\left(K_0 + K_2 + K_3\right) \\
&+ K_9K_8K_7K_6\left(K_0 + K_1 + K_2 + K_3 + K_4 + K_5\right) \\
&+ \left(K_6K_8 + K_6K_7\right)\left(K_1 + K_3 + K_4 + K_5\right)\mathbf{P}_1
\end{aligned}
$$

$$
\begin{aligned}
&+ K_5K_6K_7K_8\left(K_0+K_1+K_2+K_3+K_4\right)\\
&+ K_3K_4\left(K_5+K_6+K_7+K_8\right)\mathbf{P}_1\\
&+ \left(K_5\left(K_7+K_6+K_8\right)\left(K_1+K_3+K_4\right)\right)\mathbf{P}_1,
\end{aligned}
$$

$$
\begin{aligned}
C_6 = {} & \left(K_8\left(K_7+K_6+K_5+K_4+K_3+K_1\right)+K_7\left(K_6+K_5+K_4+K_3+K_1\right)\right)\mathbf{P}_1\\
&+ \left(K_5\left(K_0+K_1+K_2+K_3+K_4\right)+K_3\left(K_2+K_0\right)\right)\left(K_8\left(K_7+K_6\right)+K_7K_6\right)\\
&+ K_2K_1\left(K_9\left(K_8+K_7+K_6+K_5\right)+K_8\left(K_7+K_6+K_5\right)+K_7\left(K_6+K_5\right)\right)\\
&+ \frac{\mu}{S^*}K_2K_9\left(K_1+K_3+K_4+K_5+K_6+K_7+K_8\right)+K_1K_3K_4K_5\left(1-R_v\right)\\
&+ K_4\left(K_0+K_2+K_3\right)\left(K_9\left(K_8+K_7+K_6+K_5\right)+K_8\left(K_7+K_6+K_5\right)\right)\\
&+ K_9\left(K_5\left(K_0+K_1+K_2+K_3+K_4\right)+K_3\left(K_2+K_0\right)\right)\left(K_8+K_7+K_6\right)\\
&+ K_7\left(K_0K_1+K_4K_3\right)\left(K_6+K_5\right)+K_4K_6K_7\left(K_0+K_2\right)+K_4K_3K_6K_5\\
&+ \left(K_9\left(K_8+K_7+K_6\right)+K_6K_7+\left(K_9+K_8+K_7+K_6\right)\left(K_0+K_2\right)\right)\mathbf{P}_3\\
&+ \left(K_3K_4\left(K_5+K_6+K_7+K_8+K_9\right)+K_4K_5\left(K_6+K_7\right)\right)\left(K_0+K_2\right)\\
&+ K_0K_1\left(K_9\left(K_8+K_7+K_6+K_5\right)+K_8\left(K_7+K_6+K_5\right)+K_6K_5\right)\\
&+ K_7K_8K_9\left(K_0+K_1+K_2+K_3+K_4+K_5+K_6\right)+K_1K_2K_5K_6\\
&+ K_6\left(K_0+K_1+K_2+K_3+K_4+K_5\right)\left(K_9\left(K_8+K_7\right)+K_8K_7\right)\\
&+ \left(K_6\left(K_5+K_4+K_3+K_1\right)+K_5\left(K_4+K_3+K_1\right)+K_4K_3\right)\mathbf{P}_1\\
&+ K_3K_5\left(K_6+K_7+K_8+K_9\right)\left(K_0+K_2\right)+K_8\left(K_7+K_6\right)\mathbf{P}_3\\
&+ \left(K_9+K_8+K_7+K_6+K_2+K_0\right)\mathbf{P}_2+\mathbf{P}_3\mathbf{P}_1,
\end{aligned}
$$

$$
\begin{aligned}
C_7 = {} & K_7K_8K_9+K_6\left(K_9+K_8+K_7\right)K_5+\left(K_9+K_8+K_7+K_6+K_2+K_0\right)\mathbf{P}_3\\
&+ \left(K_7K_9+K_7K_8+K_9K_8\right)\left(K_0+K_1+K_2+K_3+K_4+K_5+K_6\right)\\
&+ \left(\left(K_6+K_5\right)\left(K_9+K_8+K_7\right)+K_5K_6\right)\left(K_0+K_1+K_2+K_3+K_4\right)\\
&+ K_3\left(K_0+K_2\right)\left(K_9+K_8+K_7+K_6+K_5+K_4\right)+\frac{\mu}{S^*}K_2K_9+\mathbf{P}_2\\
&+ \left(\left(K_1+K_4\right)\left(K_0+K_2\right)+K_3K_4\right)\left(K_9+K_8+K_7+K_6+K_5\right)\\
&+ \left(K_8+K_7+K_6+K_5+K_4+K_3+K_1\right)\mathbf{P}_1,
\end{aligned}
$$

$$
\begin{aligned}
C_8 = {} & \left(K_9+K_8+K_7+K_6\right)\left(K_0+K_1+K_2+K_3+K_4+K_5\right)+K_1\left(K_2+K_0\right)\\
&+ K_5\left(K_0+K_1+K_2+K_3+K_4\right)+K_9\left(K_6+K_7+K_8\right)+K_8\left(K_6+K_7\right)\\
&+ K_4\left(K_0+K_2+K_3\right)+K_3\left(K_0+K_2\right)+K_6K_7+\mathbf{P}_1+\mathbf{P}_3,
\end{aligned}
$$

and

$$C_9 = K_0 + K_1 + K_2 + K_3 + K_4 + K_5 + K_6 + K_7 + K_8 + K_9.$$

Consequently, if $R_v \leqslant 1$ and inequalities (11)–(13) are true, we have $C_{i\in\{0,1,\ldots,9\}} > 0$.

Using a technique similar to the one previously employed, the coefficients of the characteristic polynomial for the Jacobian matrix at the endemic equilibrium $J(E_1)$ can be shown to be positive if $R_v > 1$.

References

[1] Alimerina, H., Ablaoui, N.L., Maamar, M.H., Bouzid, L., Tabharit, L. and Belhamiti, O. *Mathematical modeling of the impact of vaccination on infected linked Covid-19 disease*, International Conference on Recent Advances in Mathematics and Informatics (ICRAMI). IEEE, 2021.

[2] Almatroud, A.O., Djenina, N., Ouannas, A., Grassi, G. and Al-Sawalha, M.M. *A novel discrete-time COVID-19 epidemic model including the compartment of vaccinated individuals*, Math. Biosci. Eng. 19 (2022), 12387–12404.

[3] Balabdaoui, F. and Mohr, D. *Age-stratified discrete compartment model of the COVID-19 epidemic with application to Switzerland*, Sci. Rep. 10(1) (2020), 1–12.

[4] Cavanaugh, A.M., Spicer, K.B., Thoroughman, D., Glick, C. and Winter, K. *Reduced risk of reinfection with SARS-CoV-2 after COVID-19 vaccination–Kentucky, May–June 2021*, Morb. Mortal. Wkly. Rep. 70(32) (2021), 1081.

[5] Chitnis, N., Hyman, J.M. and Cushing, J.M. *Determining important parameters in the spread of malaria through the sensitivity analysis of a mathematical model*, Bull. Math. Biol. 70 (2008), 1272–1296. https://doi.org/10.1007/s11538-008-9299-0

[6] Cui, J., Li, F. and Shi, Z.L. *Origin and evolution of pathogenic coronaviruses*, Nat. Rev. Microbiol. 17(3) (2019), 181–192.

[7] Dammann, O., Gray, P. and Gressens, P., Wolkenhauer, O. and Leviton, A. *Systems epidemiology: What's in a name?*, Online J. Public Health Inform. 6(3) (2014).

[8] De la Sen, M., Alonso-Quesada, S. and Ibeas, A. *On a discrete SEIR epidemic model with exposed infectivity, feedback vaccination and partial delayed re-susceptibility*, Mathematics, 9(5) (2021), 520.

[9] De la Sen, M., Alonso-Quesada, S., Ibeas, A. and Nistal, R. *On a discrete SEIR epidemic model with two-doses delayed feedback vaccination control on the susceptible*, Vaccines, 9 (2021), 398.

[10] Diekmann, O., Heesterbeek, J.A.P. and Metz, J.A. *On the definition and the computation of the basic reproduction ratio R_0 in models for infectious diseases in heterogeneous populations*, J. Math. Biol. 28 (1990), 365–382.

[11] Imperial College COVID-19 Response Team, *Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand*, 20 (10.25561) (2020), 77482.

[12] Ivorra, B., Ferrández, M.R., Vela-Pérez, M. and Ramos, A.M. *Mathematical modeling of the spread of the coronavirus disease 2019 (COVID-19) taking into account the undetected infections. The case of China*, Commun. Nonlinear Sci. Numer. Simul. 88 (2020), 105303.

[13] Killerby, M.E., Biggs, H.M., Midgley, C.M., Gerber, S.I. and Watson, J.T. *Middle East respiratory syndrome coronavirus transmission*, Emerg. Infect. Dis. 26(2) (2020), 191.

[14] Kranz, J. and Hau, B. *Systems analysis in epidemiology*, Annu. Rev. Phytopathol. 18(1) (1980), 67–83.

[15] Krishna, M.V. *Mathematical modelling on diffusion and control of COVID–19*, Infect. Dis. Model., 5 (2020), 588–597.

[16] Ksiazek, T.G., Erdman, D., Goldsmith, C.S., Zaki, S.R., Peret, T., Emery, S., Tong, S., Urbani, C., Comer, J.A., Lim, W. and Rollin, P.E.

A novel coronavirus associated with severe acute respiratory syndrome, N. Engl. J. Med., 348(20) (2003), 1953–1966.

[17] Liu, Y., Yan, L.M., Wan, L., Xiang, T.X., Le, A., Liu, J.M. and Zhang, W. *Viral dynamics in mild and severe cases of COVID-19*, Lancet Infect. Dis. 20(6) (2020), 656–657.

[18] Morgan, G., de Azambuja, E., Punie, K., Ades, F., Heinrich, K., Personeni, N., Rahme, R., Ferrara, R., Pels, K., Garassino, M. and von Bergwelt-Baildon, M. *OncoAlert round table discussions: The global COVID-19 experience*, JCO Global Oncology, 7 (2021), 455–463.

[19] Peiris, J.S., Guan, Y. and Yuen, K. *Severe acute respiratory syndrome*, Nat. Med. 10(12) (2004), S88–S97.

[20] Rabitz, H., Kramer, M. and Dacol, D. *Sensitivity analysis in chemical kinetics*, Annu. Rev. Phys. Chem. 34 (1983), 419–461.

[21] Ramosa, A.M., Ferrández, M.R., Vela-Pérez, M. and Ivorra, B. *A simple but complex enough θ-SIR type model to be used with COVID-19 real data. Application to the case of Italy*, Phys. D: Nonlinear Phenom. 421 (2021), 132839.

[22] 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. https://doi.org/10.1016/S0025-5564(02)00108-6.

[23] World Health Organization, *Naming the coronavirus disease (COVID-19) and the virus that causes it*, Braz. J. Implantol. Health Sci. 2(3) (2020).

[24] World Health Organization, *Novel Coronavirus (2019-nCoV): situation report 1*, https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200121-sitrep-1-2019-ncov.pdf (accessed on 20 January 2023).

[25] World Health Organization, *Novel Coronavirus (2019-nCoV):situation report 51*, https://www.who.int/docs/default-source/

coronaviruse/situation-reports/20200311-sitrep-51-covid-19.pdf (accessed on 20 January 2023).

[26] World Health Organization, *Novel Coronavirus (2019-nCoV):situation report 75*, https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200404-sitrep-75-covid-19.pdf (accessed on 20 January 2023).

[27] World Health Organization, *Weekly epidemiological update on COVID-19 - 21 December 2021*, https://www.who.int/publications/m/item/weekly-epidemiological-update-on-covid-19---21-december-2021 (accessed on 20 January 2023).

[28] World Health Organization, *Weekly epidemiological update on COVID-19 09 November 2022*, https://www.who.int/publications/m/item/weekly-epidemiological-update-on-covid-19---9-november-2022 (accessed on 20 January 2023)

[29] Xu, M., Wang, D., Wang, H., Zhang, X., Liang, T., Dai, J., Li, M., Zhang, J., Zhang, K., Xu, D. and Yu, X. *COVID-19 diagnostic testing: technology perspective*, Clin. Transl. Med. 10(4) (2020), e158.

[30] Yang, B., Yu, Z. and Cai, Y. *The impact of vaccination on the spread of COVID-19: Studying by a mathematical model*, Phys. A: Stat. Mech. 590 (2022), 126717.

[31] Yavuz, M.Ö., Coşar, F., Günay, F. and Özdemir, F.N. *A new mathematical modeling of the COVID-19 pandemic including the vaccination campaign*, Open Journal of Modelling and Simulation, 9(3) (2021), 299–321.

[32] Zaki, A.M., Van Boheemen, S. , Bestebroer, T.M., Osterhaus, A.D. and Fouchier, R.A. *Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia*, New Engl. J. Med. 367(19) (2012), 1814–1820.

Iranian Journal of Numerical Analysis and Optimization
Vol. 14, No. 1, 2024, pp 172–199
https://doi.org/10.22067/ijnao.2023.83389.1292
https://ijnao.um.ac.ir/

CC BY NC

Research Article

Modal spectral Tchebyshev Petrov–Galerkin stratagem for the time-fractional nonlinear Burgers' equation

Y.H. Youssri*, and A.G. Atta

Abstract

Herein, we construct an explicit modal numerical solver based on the spectral Petrov–Galerkin method via a specific combination of shifted Chebyshev polynomial basis for handling the nonlinear time-fractional Burger-type partial differential equation in the Caputo sense. The process reduces the problem to a nonlinear system of algebraic equations. Solving this algebraic equation system will yield the approximate solution's unknown coefficients. Many relevant properties of Chebyshev polynomials are reported, some connection and linearization formulas are reported and proved, and

*Corresponding author

Received 12 June 2023; revised 21 October 2023; accepted 30 October 2023

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

How to cite this article
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.*, 2024; 14(1): 172-199. https://doi.org/10.22067/ijnao.2023.83389.1292

all elements of the obtained matrices are evaluated neatly. Also, convergence and error analyses are established. Various illustrative examples demonstrate the applicability and accuracy of the proposed method and depict the absolute and estimated error figures. Besides, the current approach's high efficiency is proved by comparing it with other techniques in the literature.

AMS subject classifications (2020): 65M60, 11B39, 40A05, 34A08.

Keywords: Time-fractional Burgers' equation; Chebyshev polynomials; Petrov–Galerkin method; Convergence analysis.

1 Introduction

The orthogonal polynomials of the shifted first-kind Chebyshev polynomials (S1KCPs) are defined on the interval $[-1, 1]$. Due to their advantageous characteristics and value in spectral approaches, they play a vital part in the numerical solution of partial differential equations (PDEs).

Using an orthogonal polynomial-based series expansion, spectral methods [11, 25, 19, 20, 10] are numerical approaches that try to approximate the solution of a PDE. These techniques are particularly well adapted to the orthogonality, clustering, and exponential convergence of the S1KCPs.

The S1KCPs were denoted as $T_n(x) = \cos(n\,\theta)$, where n is the degree of the polynomial and $\theta = \cos^{-1}(x)$ is the angle between the x-axis and the point $(x, 0)$ on the unit circle.

The clustering property of the S1KCPs is one of their main benefits. The Chebyshev polynomials are unique among orthogonal polynomial families in grouping at the ends of the interval $[-1, 1]$. This clustering property is beneficial when approximating functions with boundary layers or steep gradients at the endpoints.

Another feature of spectral methods based on Chebyshev polynomials is their exponential convergence. By truncating the series expansion involving the Chebyshev polynomials, spectral methods approximate the solution of a PDE. As the degree of the truncated series increases, the approximation

approaches the real solution exponentially. Chebyshev spectral methods are particularly efficient and precise due to their rapid convergence feature.

Chebyshev spectral methods for numerically solving PDEs require modeling the solution as a series expansion with S1KCPs as basis functions. By projecting the PDE onto the basis functions and solving a sequence of algebraic equations, the series expansion coefficients are determined.

The number of basis functions (polynomial degree) chosen depends on the required accuracy and the issue features. Higher degrees provide more precise results but also require more computational resources. In practice, it is critical to balance precision and efficiency.

Chebyshev spectral methods, such as [7, 9, 10], have been effectively used for various fractional PDEs, such as elliptic, parabolic, and hyperbolic equations. They have proven especially useful in problems with smooth solutions, periodic boundary conditions, and unbounded domains.

Finally, the S1KCPs [27] are valuable tools for numerically solving PDEs. They are well-suited for spectral approaches due to their orthogonality, clustering, and exponential convergence qualities. These polynomials can obtain accurate and efficient approximations of PDE solutions, making them a powerful tool in computational mathematics and engineering. Chebyshev polynomials have recently been widely employed to solve several forms of differential problems; for example, see [41, 8, 1, 38, 42, 17, 2, 6, 16, 26].

We can guarantee that knowledge of the properties and applications of S1KCPs in the numerical solution of PDEs is generally established and found in various references and textbooks on numerical methods for PDEs. Among the significant works in this field are [14, 35, 13].

The nonlinear time-fractional equation Burger's equation is a PDE with a fractional derivative in time that incorporates nonlinear convection and diffusion factors. It is a variation of the traditional Burger's equation, a simple model for various physical phenomena, such as fluid flow and traffic movement. Incorporating fractional derivatives in time allows the system to include nonlocal and memory effects.

The general form of the nonlinear time-fractional Burger's equation is given as [18]:

$$\frac{\partial^\alpha u}{\partial t^\alpha} + \sigma\, u\, \frac{\partial u}{\partial x} - \kappa\, \frac{\partial^2 u}{\partial x^2} = s(x,t), \quad 0 < \alpha < 1, \tag{1}$$

$$\begin{aligned} u(x,0) &= g^1(x), \quad 0 < x \leq 1, \\ u(0,t) &= g^2(t), \quad u(1,t) = g^3(t), \quad 0 < t \leq 1, \end{aligned} \tag{2}$$

where $u(x,t)$ is the unknown function representing the dependent variable, t is time, x is the spatial variable, σ is a constant coefficient that controls the strength of the convection term, κ is a constant coefficient controlling the diffusion term, and $s(x,t)$ is the source term.

Due to the presence of both nonlinear and fractional variables, solving the nonlinear time-fractional Burger's equation analytically is difficult. Numerical approaches, on the other hand, can be used to approximate its solutions. Many techniques, such as finite difference methods, finite element methods, and spectral methods, can be used to deal with fractional derivatives and nonlinearity.

Numerical approaches for solving the nonlinear time-fractional Burger's equation frequently involve griding the spatial domain and using time-stepping methods to approximate the temporal derivatives. Fractional difference operators and fractional integral transformations can approximate fractional derivatives. For further methods that studied other types of fractional differential equations, see [5, 22, 29, 23].

Furthermore, the behavior of the nonlinear time-fractional Burger's equation can exhibit fascinating phenomena, such as the generation of shock waves, solitons, and other nonlinear waves. Memory effects are introduced by the fractional derivative in time, which might alter the transmission and evolution of these nonlinear structures.

Scientists are constantly researching the properties, analytical solutions, and numerical approaches of the nonlinear time-fractional Burger's equation. It is used in a variety of domains, such as fluid dynamics, heat transfer, and nonlinear wave phenomena, where the inclusion of fractional derivatives in time allows a more precise representation of the system dynamics. The interested reader can see the recent diverse numerical methods used to solve Burgers' problem in [15, 31, 44, 36, 24].

The Petrov–Galerkin method [34, 27] is a numerical technique for solving PDEs. It is a variation of the more general Galerkin approach, which aims to approximate a PDE solution by projecting it onto a finite-dimensional trial function subspace. To improve the accuracy and stability of the approximation, the Petrov–Galerkin approach incorporates an additional weighting function known as the test function or the Petrov function.

The Petrov–Galerkin approach [43] is especially effective for PDEs with specific properties, such as those with convection-dominated terms or singularities. It solves the usual Galerkin method's difficulties in capturing precise solutions in these challenging settings.

The Petrov–Galerkin approach [28] has been effectively used for various PDEs, including convection-diffusion, Navier-Stokes, and advection-dominated situations. It provides a versatile framework for dealing with difficult PDEs and is more accurate and stable than the Galerkin technique.

Overall, the Petrov–Galerkin approach is a strong numerical methodology that extends the Galerkin method to meet the difficulties given by specific types of PDEs. It improves the accuracy and stability of the approximation by introducing Petrov functions as extra weighting functions, making it an important tool in the field of numerical PDE solving.

The structure of this article is as follows: The theory of fractional calculus and the relevant properties of Chebyshev polynomials are briefly introduced in section 2. A numerical spectral Petrov–Galerkin technique for solving the time-fractional Burgers' type equation is constructed in section 3. Section 4 discusses the convergence and error analysis of the method. Some numerical examples are given in section 5 to illustrate the theoretical conclusions. Section 6 contains conclusions.

2 Preliminaries and essential relations

2.1 The fractional derivative in the Caputo sense

Definition 1. [32] The Caputo fractional derivative of order s is defined as

$$D_x^s u(x) = \frac{1}{\Gamma(m-s)} \int_0^x (x-y)^{m-s-1} u^{(m)}(y) dy, \quad s > 0, \quad x > 0, \tag{3}$$

where $m-1 \leqslant s < m, \quad m \in \mathbb{N}$.

The following properties are satisfied by the operator D_x^s for $m-1 \leqslant s < m, \quad m \in \mathbb{N}$,

$$D_x^s c = 0, \ (c \text{ is a constant}) \tag{4}$$

$$D_x^s \, x^m = \begin{cases} 0, & \text{if } \ m \in \mathbb{N}_0 \quad \text{and} \quad m < \lceil s \rceil, \\ \frac{\Gamma(m+1)}{\Gamma(m-s+1)} \, x^{m-s}, & \text{if } \ m \in \mathbb{N}_0 \quad \text{and} \quad m \geq \lceil s \rceil, \end{cases} \tag{5}$$

where $\mathbb{N} = \{1, 2, 3, \ldots\}$, $\mathbb{N}_0 = \{0\} \cup \mathbb{N}$ and the notation $\lceil \alpha \rceil$ denotes the ceiling function.

2.2 An account on the S1KCPs

Let $T_j^*(x)$ be the S1KCPs defined in the interval $[0, 1]$ by $T_j^*(x) = T_j(2\,x-1)$. These polynomials can be defined as [8, 38]

$$T_m^*(x) = m \sum_{k=0}^{m} \frac{(-1)^{m-k} \, 2^{2\,k} \, (m+k-1)!}{(m-k)! \, (2\,k)!} \, x^k, \quad m > 0, \tag{6}$$

satisfying the following orthogonality relation with respect to the weight function $\hat{w}(x) = \frac{1}{\sqrt{x\,(1-x)}}$ [8, 38]:

$$\int_0^1 \hat{w}(x) \, T_m^*(x) \, T_n^*(x) \, dx = h_m \, \delta_{m,n}, \tag{7}$$

where

$$h_m = \begin{cases} \pi, & \text{if } \ m = 0, \\ \frac{\pi}{2}, & \text{if } \ m > 0, \end{cases} \tag{8}$$

and

$$\delta_{m,n} = \begin{cases} 1, & \text{if } m = n, \\ 0, & \text{if } m \neq n. \end{cases} \tag{9}$$

The recurrence relation of $T_m^*(x)$ is

$$T^*_{m+1}(x) = 2\,(2\,x-1)\,T^*_m(x) - T^*_{m-1}(x), \tag{10}$$

where $T^*_0(x) = 1$ and $T^*_1(x) = 2\,x - 1$.

Moreover, the inversion formula is [8, 38]

$$x^j = 2^{1-2\,j}\,(2\,j)!\sum_{p=0}^{j}\frac{\epsilon_p}{(j-p)!(j+p)!}\,T^*_p(x), \quad j \geq 0, \tag{11}$$

where

$$\epsilon_m = \begin{cases} \frac{1}{2}, & \text{if } m = 0, \\ 1, & \text{otherwise.} \end{cases} \tag{12}$$

The following relation between $T^*_i(x)$ and $U^*_i(x)$ is correct

$$D\,T^*_i(x) = 2i\,U^*_{i-1}(x), \qquad \text{for all } i \geq 1. \tag{13}$$

The following linearization formula is valid

$$T^*_r(x)\,U^*_s(x) = \frac{1}{2}\left(U^*_{s+r}(x) + U^*_{s-r}(x)\right), \qquad \text{for all } r, s \geq 0. \tag{14}$$

Corollary 1. [1] For every positive integer q, the qth derivative of $T^*_j(x)$ can be expressed in terms of their original polynomials as

$$D^q\,T^*_j(x) = \sum_{\substack{p=0 \\ (j+p+q)\ \text{even}}}^{j-q} d_{j,p,q}\;T^*_p(x), \tag{15}$$

where

$$d_{j,p,q} = \frac{j\,2^{2\,q}\,\epsilon_p\,(q)_{\frac{1}{2}(j-p-q)}}{\left(\frac{1}{2}(j-p-q)\right)!\,\left(\frac{1}{2}(j+p+q)\right)_{1-q}}, \tag{16}$$

and ϵ_p is defined in (12).

Lemma 1. [4] For all nonnegative integers m and n, the following linearization formula holds for the S1KCPs:

$$T^*_m(x)\,T^*_n(x) = \frac{1}{2}\left(T^*_{m+n}(x) + T^*_{|m-n|}(x)\right). \tag{17}$$

Lemma 2. Let j and i be any two nonnegative integers. The moments' formula for the S1KCPs are given by

$$x^j\, T_i^*(x) = \sum_{s=i-j}^{i+j} \frac{\Gamma(2j+1)}{4^j \Gamma(i+j-s+1)\Gamma(-i+j+s+1)}\, T_s^*(x). \tag{18}$$

Proof. Multiplying both sides of (11) with $T_i^*(x)$ and direct use of Lemma 1, we get the desired result. □

Remark 1. The following relation is satisfied:

$$\int_0^1 \hat{w}(x)\, U_i^*(x)\, T_r^*(x)\, dx = \sigma_{i,r}, \tag{19}$$

where $U_i^*(x)$ is the shifted Chebyshev polynomials of the second kind and

$$\sigma_{i,r} = \begin{cases} \pi, & \text{if } (i-r) \text{ even}, \ i \geq r, \\ 0, & \text{otherwise.} \end{cases} \tag{20}$$

Remark 2. The following relation is satisfied

$$\int_0^1 x^2\, \hat{w}(x)\, T_i^*(x)\, T_s^*(x)\, dx = \kappa_{i,s}, \tag{21}$$

where

$$\kappa_{i,s} = \frac{\pi}{32\, \epsilon_i \epsilon_s} \begin{cases} 1, & \text{if } |s-i| = 2, \\ 4, & \text{if } |s-i| = 1, \\ 6, & \text{if } s-i = 0,\ i > 1, \\ 3, & \text{if } s = i = 0, \\ 7, & \text{if } s = i = 1, \\ 0, & \text{otherwise.} \end{cases} \tag{22}$$

3 Petrov–Galerkin approach for the time-fractional Burgers' equation

In this section, we consider the following time-fractional Burgers' equation [33]:

$$\frac{\partial^\alpha \chi(x,t)}{\partial t^\alpha} + \chi(x,t)\, \frac{\partial \chi(x,t)}{\partial x} - \Psi\, \frac{\partial^2 \chi(x,t)}{\partial x^2} = S(x,t), \quad 0 < \alpha < 1, \tag{23}$$

subject to the following initial and boundary conditions:

$$\chi(x,0) = g(x), \quad 0 < x \leq 1,$$
$$\chi(0,t) = \zeta_1(t), \quad \chi(1,t) = \zeta_2(t), \quad 0 < t \leq 1, \tag{24}$$

where Ψ is the kinematic viscosity and $S(x,t)$ is the source term.

Now, to proceed with our proposed Petrov–Galerkin approach, we will use the following transformation:

$$\chi(x,t) := u(x,t) + \Upsilon(x,t), \tag{25}$$

where

$$\Upsilon(x,t) = (1-x)\ (\chi(0,t) - \chi(0,0)) + x\ (\chi(1,t) - \chi(1,0)) + \chi(x,0), \tag{26}$$

to convert (23) governed by the conditions (24) into the following modified equation:

$$\frac{\partial^\alpha u(x,t)}{\partial t^\alpha} + u(x,t)\frac{\partial u(x,t)}{\partial x} + u(x,t)\frac{\partial \Upsilon(x,t)}{\partial x} + \Upsilon(x,t)\frac{\partial u(x,t)}{\partial x} - \Psi\frac{\partial^2 u(x,t)}{\partial x^2} = F(x,t), \quad 0 < \alpha < 1, \tag{27}$$

governed by the following homogeneous conditions:

$$u(x,0) = 0, \quad 0 < x < 1,$$
$$u(0,t) = u(1,t) = 0, \quad 0 < t \leq 1, \tag{28}$$

where

$$F(x,t) = S(x,t) - \frac{\partial^\alpha \Upsilon(x,t)}{\partial t^\alpha} - \Upsilon(x,t)\frac{\partial \Upsilon(x,t)}{\partial x} + \Psi\frac{\partial^2 \Upsilon(x,t)}{\partial x^2}. \tag{29}$$

Therefore, instead of solving (23) governed by (24), we can solve the modified equation (27) governed by the homogeneous conditions (28).

3.1 Trial functions

Consider the following basis functions:

$$\lambda_i^*(x) = T_{i+2}^*(x) - T_i^*(x),$$

$$\phi_j^*(t) = t\, T_j^*(t). \tag{30}$$

Remark 3. The polynomials $\lambda_i^*(x)$ can be written alternatively in the following form:

$$\lambda_i^*(x) = -8\, x\, (1-x)\, U_i^*(x). \tag{31}$$

The orthogonality relations of $\lambda_i^*(x)$ and $\phi_j^*(t)$ are given by

$$\int_0^1 \lambda_m^*(x)\, \lambda_n^*(x)\, \frac{1}{(x\,(1-x))^{3/2}} dx = 8\,\pi\,\delta_{n,m}, \tag{32}$$

and

$$\int_0^1 \phi_m^*(t)\, \phi_n^*(t)\, \frac{1}{t^2\,\sqrt{t\,(1-t)}} dx = h_m\,\delta_{m,n}, \tag{33}$$

where h_m is defined in (8).

Lemma 3. For all nonnegative integers m and n, the following linearization formula holds:

$$\begin{aligned}\lambda_m^*(x)\,\frac{d\,\lambda_i^*(x)}{d\,x} = &-i\,U_{i-m-3}^*(x) + (2\,i+2)\,U_{i-m-1}^*(x) - (i+2)\,U_{i-m+1}^*(x)\\ &+ i\,U_{i+m-1}^*(x) - (2\,i+2)\,U_{i+m+1}^*(x) + (i+2)\,U_{i+m+3}^*(x).\end{aligned} \tag{34}$$

Proof. We express $D\,\lambda_i^*(x)$ as a combination of $U_i^*(x)$ via (13). Then we linearize $\lambda_m^*(x)\,D\,\lambda_i^*(x)$ using (14), we get the desired result. □

Lemma 4. For all nonnegative integers m and n, the following linearization formula holds:

$$\phi_n^*(t)\,\phi_j^*(t) = \frac{t^2}{2}\ \left(T_{j+n}^*(t) + T_{j-n}^*(t)\right). \tag{35}$$

Proof. The proof of this lemma is a direct result of Lemma 1. □

Theorem 1. The first and second derivatives of $\psi_m^*(x)$ can be expressed explicitly as

$$\begin{aligned}\frac{d\,\lambda_i^*(x)}{d\,x} &= \sum_{j=0}^{i+2} \varsigma_{j,i}\, T_j^*(x),\\ \frac{d^2\,\lambda_i^*(x)}{d\,x^2} &= \sum_{j=0}^{i+1} \tau_{j,i}\, T_j^*(x),\end{aligned} \tag{36}$$

where

$$\varsigma_{j,i} = 4 \begin{cases} j+1, & \text{if } i+1=j, \\ 2, & \text{if } (i-j)\, odd\,, j>0, \\ 1, & \text{if } (i-j)\, odd\,, j=0, \\ 0, & \text{otherwise,} \end{cases}$$

$$\tau_{j,i} = \delta_j \begin{cases} 8\left(3\,i\,(i+2) - j^2 + 4\right), & \text{if } (i-j)\, even\,, i \geq j, \\ 0, & \text{otherwise,} \end{cases}$$

$$\delta_j = \begin{cases} 1, & \text{if } j=0, \\ \frac{1}{2}, & \text{otherwise.} \end{cases} \tag{37}$$

Proof. Relations (36) can be deduced after using Corollary 1 when $q = 1, 2$ along with $\lambda_i^*(x)$ defined in (30), then collecting like terms and rearranging the summations. □

3.2 Petrov–Galerkin solution for the time-fractional Burgers' equation

Now, one may set

$$\begin{aligned} \Theta^N(\Omega) &= \text{span}\{\lambda_i^*(x)\,\phi_j^*(t) : i,j = 0,1,\ldots,N\}, \\ \Lambda^N(\Omega) &= \{u \in \Theta^N(\Omega) : u(x,0) = u(0,t) = u(1,t) = 0\}, \end{aligned} \tag{38}$$

where $\Omega = [0,1]^2$. Then any function $u^N(x,t) \in \Lambda^N(\Omega)$ may be written as

$$u^N(x,t) = \sum_{i=0}^{N} \sum_{j=0}^{N} c_{ij}\, \lambda_i^*(x)\, \phi_j^*(t). \tag{39}$$

The application of Petrov–Galerkin technique [39] is used to find $u^N(x,t) \in \Lambda^N$ such that

$$\left(\frac{\partial^\alpha u^N(x,t)}{\partial t^\alpha}, T_r^*(x)\, T_s^*(t)\right)_{\omega(x,t)} + \left(u^N(x,t)\, \frac{\partial u^N(x,t)}{\partial x}, T_r^*(x)\, T_s^*(t)\right)_{\omega(x,t)}$$

$$
\begin{aligned}
&+\left(u^N(x,t)\,\frac{\partial\,\Upsilon^N(x,t)}{\partial\,x},T_r^*(x)\,T_s^*(t)\right)_{\omega(x,t)}\\
&+\left(\Upsilon^N(x,t)\,\frac{\partial\,u^N(x,t)}{\partial\,x},T_r^*(x)\,T_s^*(t)\right)_{\omega(x,t)}\\
&-\Psi\,\left(\frac{\partial^2\,u^N(x,t)}{\partial\,x^2},T_r^*(x)\,T_s^*(t)\right)_{\omega(x,t)}\\
&\quad=\left(F(x,t),T_r^*(x)\,T_s^*(t)\right)_{\omega(x,t)},\quad 0\le r,s\le N,
\end{aligned}
\tag{40}
$$

where $T_r^*(x)\,T_s^*(t)$ is the test function and $\omega(x,t)=\hat{w}(x)\,\hat{w}(t)$.

Therefore, (40) can be written after using the definition of $u^N(x,t)$ (39) as

$$
\begin{aligned}
&\sum_{i=0}^{N}\sum_{j=0}^{N}c_{ij}\,g_{i,r}\,b_{j,s}+\sum_{m=0}^{N}\sum_{n=0}^{N}\sum_{i=0}^{N}\sum_{j=0}^{N}c_{mn}\,c_{ij}\,d_{m,i,r}\,h_{n,j,s}\\
&\quad+\sum_{m=0}^{N}\sum_{n=0}^{N}\sum_{i=0}^{N}\sum_{j=0}^{N}c_{mn}\,a_{ij}\,d_{m,i,r}\,h_{n,j,s}+\sum_{m=0}^{N}\sum_{n=0}^{N}\sum_{i=0}^{N}\sum_{j=0}^{N}b_{mn}\,c_{ij}\,d_{m,i,r}\,h_{n,j,s}\\
&\quad-\Psi\sum_{i=0}^{N}\sum_{j=0}^{N}c_{ij}\,Z_{i,r}\,Q_{j,s}=F_{r,s},\quad 0\le r,s\le N,
\end{aligned}
\tag{41}
$$

where a_{ij} and b_{ij} are determined from the following relations:

$$
\begin{aligned}
a_{ij}&=\frac{1}{8\,\pi\,h_j}\int_0^1\int_0^1\frac{\partial\,\Upsilon^N(x,t)}{\partial\,x}\,\lambda_i^*(x)\,\phi_j^*(t)\,\frac{1}{(x\,(1-x))^{3/2}}\,\frac{1}{t^2\,\sqrt{t\,(1-t)}}dx\,dt,\\
b_{ij}&=\frac{1}{8\,\pi\,h_j}\int_0^1\int_0^1\Upsilon^N(x,t)\,\lambda_i^*(x)\,\phi_j^*(t)\,\frac{1}{(x\,(1-x))^{3/2}}\,\frac{1}{t^2\,\sqrt{t\,(1-t)}}dx\,dt.
\end{aligned}
\tag{42}
$$

Also,

$$
\begin{aligned}
g_{i,r}&=\left(\lambda_i^*(x),T_r^*(x)\right)_{\hat{w}(x)},&\qquad b_{j,s}&=\left(\frac{d^\alpha\,\phi_j^*(t)}{d\,t^\alpha},T_s^*(t)\right)_{\hat{w}(t)},\\
d_{m,i,r}&=\left(\lambda_m^*(x)\,\frac{d\,\lambda_i^*(x)}{d\,x},T_r^*(x)\right)_{\hat{w}(x)},&\qquad h_{n,j,s}&=\left(\phi_n^*(t)\,\phi_j^*(t),T_s^*(t)\right)_{\hat{w}(t)},\\
Z_{i,r}&=\left(\frac{d^2\,\lambda_i^*(x)}{d\,x^2},T_r^*(x)\right)_{\hat{w}(x)},&\qquad Q_{j,s}&=\left(\phi_j^*(t),T_s^*(t)\right)_{\hat{w}(t)},
\end{aligned}
$$

and $F_{r,s} = \left(f(x,t), T_r^*(x)\, T_s^*(t)\right)_{\omega(x,t)}$.

Theorem 2. The elements $g_{i,r}$, $b_{j,s}$, $d_{m,i,r}$, $h_{n,j,s}$, $Z_{i,r}$, and $Q_{j,s}$ are given by

$$g_{i,r} = \begin{cases} -\frac{\pi}{2\epsilon_i}, & \text{if } r-i=0, \\ \frac{\pi}{2}, & \text{if } r-i=2, , \\ 0, & \text{otherwise}, \end{cases}$$

$$b_{j,s} = \pi\,(-1)^j\,\Gamma\left(\frac{3}{2}-\alpha\right)\ {}_4F_3\left(\left.\begin{matrix} 2, -j, j, \frac{3}{2}-\alpha \\ \\ \frac{1}{2}, -s-\alpha+2, s-\alpha+2 \end{matrix}\right| 1\right),$$

$$\begin{aligned} d_{m,i,r} = &-i\,\sigma_{i-m-3,r} + (2\,i+2)\,\sigma_{i-m-1,r} - (i+2)\,\sigma_{i-m+1,r} + i\,\sigma_{i+m-1,r} \\ &- (2\,i+2)\,\sigma_{i+m+1,r} + (i+2)\,\sigma_{i+m+3,r}, \end{aligned}$$

$$h_{n,j,s} = \frac{1}{2}\left(\kappa_{j+n,s} + \kappa_{j-n,s}\right),$$

$$Z_{i,r} = \begin{cases} 4\,\pi\,\left(3\,i\,(i+2) - r^2 + 4\right), & \text{if } (i-r)\,even,\ i \geq r, \\ 0, & \text{otherwise}, \end{cases}$$

$$Q_{j,s} = \begin{cases} \frac{\pi}{8\epsilon_i}, & \text{if } s-j=1, \\ \frac{\pi}{8\epsilon_s}, & \text{if } j-s=1, \\ \frac{\pi}{4\epsilon_j}, & \text{if } s-j=0, \\ 0, & \text{otherwise}. \end{cases} \tag{43}$$

Proof. The elements $g_{i,r}$ can be obtained after using the orthogonality relation (7) and the definition of $\lambda_i^*(x)$ defined in (30), then collecting like terms.

To find the elements $b_{j,s}$, based on the definition of Caputo's fractional derivative (3) and the power form of $T_j^*(x)$ (6), one can write

$$\begin{aligned} b_{j,s} &= \left(\frac{d^\alpha\,\phi_j^*(t)}{d\,t^\alpha}, T_s^*(t)\right)_{\hat{w}(t)} \\ &= \sum_{k=0}^{j} \frac{j 2^{2k} (k+1)! (-1)^{j-k} (j+k-1)!}{(2k)!(j-k)!(-\alpha+k+1)!} \int_0^1 T_s^*(t) t^{-\alpha+k+1}\,\hat{w}(t)\,dt \end{aligned}$$

$$= \sum_{k=0}^{j} \frac{\sqrt{\pi} j 4^k (-1)^{j-k} \Gamma(k+2) \Gamma(j+k) \Gamma\left(k-\alpha+\frac{3}{2}\right)}{\Gamma(2k+1)\Gamma(j-k+1)\Gamma(k-s-\alpha+2)\Gamma(k+s-\alpha+2)}. \quad (44)$$

The last relation can be summed to give the following result:

$$b_{j,s} = \pi\,(-1)^j\,\Gamma\left(\frac{3}{2}-\alpha\right)\ {}_4F_3\left(\left.\begin{matrix} 2, -j, j, \frac{3}{2}-\alpha \\ \\ \frac{1}{2}, -s-\alpha+2, s-\alpha+2 \end{matrix}\right| 1\right). \quad (45)$$

The elements $d_{m,i,r}$ can be obtained after using Lemma 3 along with Remark 1.

Similarly, the elements $h_{n,j,s}$ can be obtained after using Lemma 4 along with Remark 2.

The elements $Z_{i,r}$ can be obtained after using Theorem 1 along with the orthogonality relation (7), collecting like terms, and rearranging the summations.

Finally, the elements $Q_{j,s}$ can be obtained after using Lemma 2 along with the orthogonality relation (7) and doing some computations. □

Remark 4. The inner product of $(u(x,t), v(x,t))_{\omega(x,t)}$, $(u(x), v(x))_{\hat{w}(x)}$, and $(u(t), v(t))_{\hat{w}(t)}$ are defined as

$$\begin{aligned} (u(x,t), v(x,t))_{\omega(x,t)} &= \int_0^1 \int_0^1 u(x,t)\, v(x,t)\, \omega(x,t)\, dx\, dt, \\ (u(x), v(x))_{\hat{w}(x)} &= \int_0^1 u(x)\, v(x)\, \hat{w}(x)\, dx, \\ (u(t), v(t))_{\hat{w}(t)} &= \int_0^1 u(t)\, v(t)\, \hat{w}(t)\, dt. \end{aligned} \quad (46)$$

Remark 5. The $(N+1) \times (N+1)$ nonlinear system of equations in (41) in the unknown expansion coefficient c_{ij} can be solved through a suitable numerical solver such as Newton's iterative technique.

4 Error bound

Herein, we give an upper bound of the absolute errors (AEs) using Lagrange interpolation polynomials.

Algorithm 1: Coding algorithm for the proposed technique

Input Ψ, $g(x)$, $\zeta_1(t)$, $\zeta_2(t)$, α, and $S(x,t)$.

Step 1. Using transformation (25) to convert the nonlinear TFCE (23)–(24), into modified equation (27)–(28)

Step 2. Assume an approximate solution $u^N(x,t)$ as in (39).

Step 3. Apply Petrov–Galerkin method to obtain the system in (41).

Step 4. Use Theorem 2 to get the elements of $g_{i,r}$, $b_{j,s}$, $d_{m,i,r}$, $h_{n,j,s}$, $Z_{i,r}$, and $Q_{j,s}$.

Step 5. Use *FindRoot* command with initial guess $\{c_{ij} = 10^{-i-j},\, i,j : 0,1,\ldots,N\}$, to solve the system (41) to get c_{ij}.

Output $u^N(x,t)$.

Let $u^N(x,t) \in \Lambda^N(\Omega)$ be the best approximation of $u(x,t)$; then, the definition of the best approximation enables us to write the following inequality:

$$||u(x,t) - u^N(x,t)||_\infty \leq ||u(x,t) - v^N(x,t)||_\infty, \qquad \text{for all } v^N(x,t) \in \Lambda^N(\Omega). \tag{47}$$

Moreover, the previous inequality is also true if $\hat{u}^N$ denotes the interpolating polynomial for $u(x,t)$ at points (x_i,t_j), where x_i are the roots of $\lambda_i^*(x)$, while t_j are the roots of $\phi_j^*(t)$.

Using similar steps as in [12, 40], one has

$$\begin{aligned} u(x,t) - v^N(x,t) = {} & \frac{\partial^{N+1} u(\eta,t)}{\partial x^{N+1}\,(N+1)!} \prod_{i=0}^{N}(x-x_i) + \frac{\partial^{N+1} u(x,\mu)}{\partial t^{N+1}\,(N+1)!} \prod_{j=0}^{N}(t-t_j) \\ & - \frac{\partial^{2N+2} u(\hat{\eta},\hat{\mu})}{\partial x^{N+1}\,\partial t^{N+1}\,((N+1)!)^2} \prod_{i=0}^{N}(x-x_i) \prod_{j=0}^{N}(t-t_j), \end{aligned} \tag{48}$$

where $\eta, \hat{\eta}, \mu, \hat{\mu} \in [0,1]$.

Now,

$$\begin{aligned} ||u(x,t) - v^N(x,t)||_\infty \leq {} & \max_{(x,t)\in\Omega} \left| \frac{\partial^{N+1} u(\eta,t)}{\partial x^{N+1}} \right| \frac{||\prod_{i=0}^{N}(x-x_i)||_\infty}{(N+1)!} \\ & + \max_{(x,t)\in\Omega} \left| \frac{\partial^{N+1} u(x,\mu)}{\partial t^{N+1}} \right| \frac{||\prod_{j=0}^{N}(t-t_j)||_\infty}{(N+1)!} \end{aligned}$$

$$- \max_{(x,t)\in\Omega} \left| \frac{\partial^{2N+2} u(\hat{\eta},\hat{\mu})}{\partial x^{N+1}\, \partial t^{N+1}} \right| \frac{||\prod_{i=0}^{N}(x-x_i)||_\infty \, ||\prod_{j=0}^{N}(t-t_j)||_\infty}{((N+1)!)^2}. \tag{49}$$

Since u is a smooth function on Ω, then there exist three constants ℓ_1, ℓ_2, and ℓ_3 such that

$$\max_{(x,t)\in\Omega} \left| \frac{\partial^{N+1} u(x,t)}{\partial x^{N+1}} \right| \le \ell_1, \quad \max_{(x,t)\in\Omega} \left| \frac{\partial^{N+1} u(x,\mu)}{\partial t^{N+1}} \right| \le \ell_2,$$
$$\max_{(x,t)\in\Omega} \left| \frac{\partial^{2N+2} u(\hat{\eta},\hat{\mu})}{\partial x^{N+1}\, \partial t^{N+1}} \right| \le \ell_3. \tag{50}$$

To minimize the factor $||\prod_{i=0}^{N}(x-x_i)||_\infty$, let us use the one-to-one mapping $x = \frac{1}{2}(z+1)$ between the intervals $[-1,1]$ and $[0,1]$ to deduce that

$$\begin{aligned} \min_{x_i\in[0,1]} \max_{x\in[0,1]} \left| \prod_{i=0}^{N}(x-x_i) \right| &= \min_{z_i\in[-1,1]} \max_{z\in[-1,1]} \left| \prod_{i=0}^{N} \frac{1}{2}(z-z_i) \right| \\ &= \left(\frac{1}{2}\right)^{N+1} \min_{z_i\in[-1,1]} \max_{z\in[-1,1]} \left| \prod_{i=0}^{N}(z-z_i) \right| \\ &= \left(\frac{1}{2}\right)^{N+1} \min_{z_i\in[-1,1]} \max_{z\in[-1,1]} \left| \frac{\lambda_{N-1}(z)}{\bar{\lambda_N}} \right|, \end{aligned} \tag{51}$$

where $\bar{\lambda_N} = 2^N$ is the leading coefficient of $\lambda_{N-1}(z) = T_{N+1}(z) - T_{N-1}(z)$ and z_i are the roots of $\lambda_{N-1}(z)$.

Also, the factor $||\prod_{j=0}^{N}(t-t_j)||_\infty$ may be minimized by using the one-to-one mapping $t = \frac{1}{2}(\bar{t}+1)$ between the intervals $[-1,1]$ and $[0,1]$ to deduce that

$$\min_{t_j\in[0,1]} \max_{t\in[0,1]} \left| \prod_{j=0}^{N}(t-t_j) \right| = \left(\frac{1}{2}\right)^{N+1} \min_{\bar{t}_j\in[-1,1]} \max_{\bar{t}\in[-1,1]} \left| \frac{\phi_N(\bar{t})}{\hat{\phi}_N} \right|, \tag{52}$$

where $\hat{\phi}_N = 2^{N-2}$ is the leading coefficient of $\phi_N(\bar{t}) = \left(\frac{\bar{t}+1}{2}\right) T_N(\bar{t})$ and $\bar{t}_j$ are the roots of $\phi_N(\bar{t})$.

Since

$$\max_{z\in[-1,1]} |\lambda_{N-1}(z)| = |T_{N+1}(1)| + |T_{N-1}(1)| = 2,$$

$$\max_{\bar{t}\in[-1,1]} |\phi_{N+1}(\bar{t})| = |\phi_{N+1}(1)| = 1, \tag{53}$$

then, inequalities (50), (51), (52), and (53) enable us to get the following desired result:

$$||u(x,t)-u^N(x,t)||_\infty \leq \ell_1 \frac{(\frac{1}{2})^N}{\bar{\lambda}_N\,(N+1)!}+\ell_2 \frac{(\frac{1}{2})^{N+1}}{\hat{\phi}_N\,(N+1)!}+\ell_3 \frac{(\frac{1}{2})^{2\,N+1}}{\bar{\lambda}_N\,\hat{\phi}_N\,((N+1)!)^2}, \tag{54}$$

which represents an upper bound of the AE.

5 Illustrative examples

Example 1. [33, 37] Consider the time-fractional Burgers' equation of the form

$$\frac{\partial^\alpha u(x,t)}{\partial t^\alpha} + u(x,t)\,\frac{\partial u(x,t)}{\partial x} - 2\,\frac{\partial^2 u(x,t)}{\partial x^2} = g(x,t), \tag{55}$$

subject to the following initial and boundary conditions:

$$\begin{aligned} &u(x,0) = 0, \quad 0 < x \leq 1,\\ &u(0,t) = u(1,t) = 0, \quad 0 < t \leq 1, \end{aligned} \tag{56}$$

where $g(x,t) = \frac{2}{\Gamma(3-\alpha)}\,t^{2-\alpha}\sin(\pi\,x) + \pi\,t^4\,\sin(\pi\,x)\,\cos(\pi\,x) + 2\,\pi^2\,t^2\,\sin(\pi\,x)$, and $u(x,t) = t^2\,\sin(\pi\,x)$ is the exact solution of this problem.

Table 1 gives a comparison of AE between our method and the method in [33] at $\alpha = 0.7$ and $\alpha = 0.8$. Figure 1 shows the maximum absolute error (MAE) at different values of N when $\alpha = 0.5$. Table 2 gives a comparison of L_∞-error between our method and methods in [37, 33] at different values of α. Figure 2 shows the AE (left) and approximate solution (right) at $\alpha = 0.9, N = 14$.

Example 2. [33] Consider the time-fractional Burgers' equation of the form

$$\frac{\partial^\alpha u(x,t)}{\partial t^\alpha} + u(x,t)\,\frac{\partial u(x,t)}{\partial x} - \frac{\partial^2 u(x,t)}{\partial x^2} = g(x,t), \tag{57}$$

subject to the following initial and boundary conditions:

$$\begin{aligned} &u(x,0) = 0, \quad 0 < x \leq 1,\\ &u(0,t) = t^2, \quad u(1,t) = -t^2, \quad 0 < t \leq 1, \end{aligned} \tag{58}$$

Table 1: Comparison of AE for Example 1

	$\alpha = 0.7$		$\alpha = 0.8$	
x	Method in [33]	Our method at $M = 14$	Method in [33]	Our method at $N = 14$
0.1	2.18701×10^{-5}	5.55112×10^{-17}	8.10773×10^{-7}	5.55112×10^{-17}
0.2	4.20505×10^{-5}	1.11022×10^{-16}	1.57153×10^{-6}	1.11022×10^{-16}
0.3	5.88630×10^{-5}	2.22045×10^{-16}	2.22713×10^{-6}	1.11022×10^{-16}
0.4	7.06958×10^{-5}	3.33067×10^{-16}	2.71560×10^{-6}	0
0.5	7.61346×10^{-5}	2.22045×10^{-16}	2.97247×10^{-6}	0
0.6	7.41772×10^{-5}	3.33067×10^{-16}	2.94204×10^{-6}	0
0.7	6.44989×10^{-5}	4.44089×10^{-16}	2.59370×10^{-6}	1.11022×10^{-16}
0.8	4.76900×10^{-5}	3.33067×10^{-16}	1.93833×10^{-6}	1.11022×10^{-16}
0.9	2.53573×10^{-5}	5.55112×10^{-17}	1.03759×10^{-6}	5.55112×10^{-17}

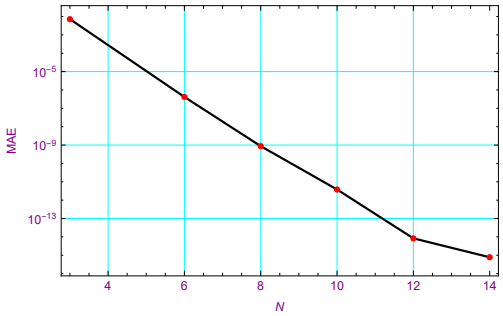


Figure 1: The MAE of Example 1 at $\alpha = 0.5$.

Table 2: Comparison of L_∞-error of Example 1

α	Method in [37] at $N = 2^7$, $M = 2^{12}$	Method in [33] at $N = 2^7$, $M = 2^{12}$	Our method at $N = 14$
0.2	5.47402×10^{-5}	4.76511×10^{-5}	3.33067×10^{-16}
0.3	5.46367×10^{-5}	4.75486×10^{-5}	3.33067×10^{-16}
0.4	5.44879×10^{-5}	4.74015×10^{-5}	2.22045×10^{-16}
0.5	5.42862×10^{-5}	4.72029×10^{-5}	7.77156×10^{-16}

where $g(x,t) = \frac{2}{\Gamma(3-\alpha)}\, t^{2-\alpha}\, \cos(\pi\, x) - \pi\, t^4\, \sin(\pi\, x)\, \cos(\pi\, x) + \pi^2\, t^2\, \cos(\pi\, x)$, and $u(x,t) = t^2\, \cos(\pi\, x)$ is the exact solution of this problem.

Table 3 reports the AE at $N = 12$ and $N = 14$ when $\alpha = 0.5$. Figure 3 shows the MAE at different values of N when $\alpha = 0.9$. Table 4 gives a comparison of L_∞-error between our method and method in [33] at different

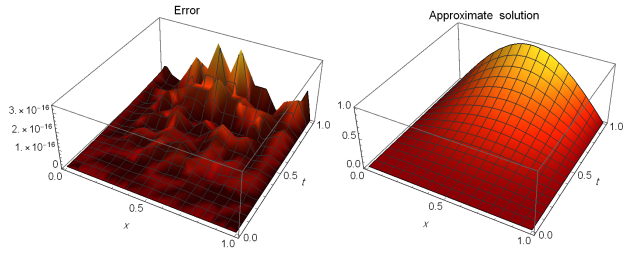


Figure 2: The AE (left) and approximate solution (right) for Example 1 at $\alpha = 0.9, N = 14$.

values of α. Figure 4 shows the exact and approximate solutions at $\alpha = 0.4, N = 14$.

Table 3: The AE of Example 2 at $\alpha = 0.5$

	$N = 12$			$N = 14$		
x	$t = 0.1$	$t = 0.5$	$t = 0.9$	$t = 0.1$	$t = 0.5$	$t = 0.9$
0.1	5.05238×10^{-16}	1.26565×10^{-14}	4.10366×10^{-14}	6.50521×10^{-19}	6.93889×10^{-18}	4.16334×10^{-17}
0.2	1.22949×10^{-15}	3.07462×10^{-14}	9.96425×10^{-14}	2.60209×10^{-18}	1.38778×10^{-17}	5.55112×10^{-17}
0.3	1.43115×10^{-15}	3.57214×10^{-14}	1.14991×10^{-13}	8.67362×10^{-19}	2.77556×10^{-17}	1.11022×10^{-16}
0.4	1.04387×10^{-15}	2.59792×10^{-14}	8.32667×10^{-14}	1.95156×10^{-18}	1.73472×10^{-17}	1.38778×10^{-17}
0.5	1.88161×10^{-19}	1.16453×10^{-17}	2.50372×10^{-17}	2.69252×10^{-19}	1.10588×10^{-17}	5.19085×10^{-17}
0.6	1.04431×10^{-15}	2.60071×10^{-14}	8.32112×10^{-14}	2.81893×10^{-18}	3.81639×10^{-17}	1.24901×10^{-16}
0.7	1.43071×10^{-15}	3.56937×10^{-14}	1.15047×10^{-13}	8.67362×10^{-19}	1.38778×10^{-17}	0
0.8	1.22949×10^{-15}	3.07601×10^{-14}	9.95871×10^{-14}	2.60209×10^{-18}	2.77556×10^{-17}	0
0.9	5.05238×10^{-16}	1.26427×10^{-14}	4.10644×10^{-14}	6.50521×10^{-19}	2.08167×10^{-17}	9.71445×10^{-17}

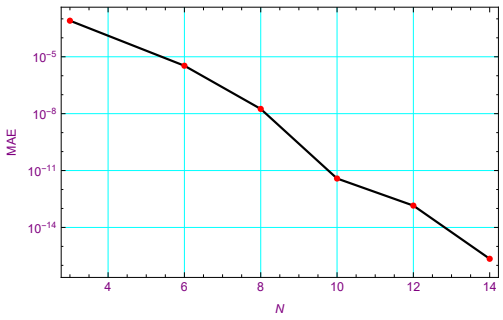


Figure 3: The MAE of Example 2 at $\alpha = 0.9$.

Table 4: Comparison of L_∞-error of Example 2

	Method in [33]		
α	$N = 2^7$, $M = 2^{12}$	$M = 2^7$, $N = 2^{11}$	Our method at $N = 14$
0.2	1.06427×10^{-5}	9.66513×10^{-6}	4.44089×10^{-16}
0.3	1.06304×10^{-5}	9.60898×10^{-6}	4.44089×10^{-16}
0.4	1.06126×10^{-5}	9.53096×10^{-6}	3.33067×10^{-16}
0.5	1.05886×10^{-5}	9.42186×10^{-6}	3.33067×10^{-16}

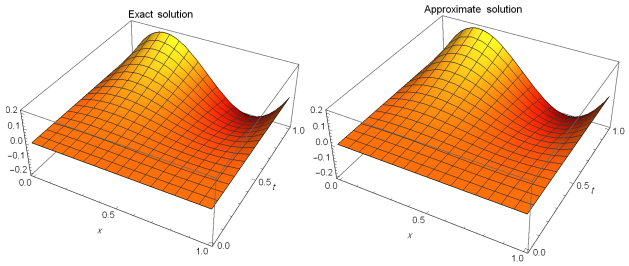


Figure 4: The exact and approximate solutions of Example 2 at $\alpha = 0.4, N = 14$.

Example 3. [33] Consider the time-fractional Burgers' equation of the form

$$\frac{\partial^\alpha u(x,t)}{\partial t^\alpha} + u(x,t)\,\frac{\partial u(x,t)}{\partial x} - \frac{\partial^2 u(x,t)}{\partial x^2} = g(x,t), \tag{59}$$

subject to the following initial and boundary conditions:

$$\begin{aligned} &u(x,0) = 0, \quad 0 < x \le 1, \\ &u(0,t) = t^2, \quad u(1,t) = e\,t^2, \quad 0 < t \le 1, \end{aligned} \tag{60}$$

where $g(x,t) = \frac{2}{\Gamma(3-\alpha)}\, e^x\, t^{2-\alpha} + t^4\, e^{2\,x} - t^2\, e^x$ and $u(x,t) = t^2\, e^x$ is the exact solution of this problem.

Table 5 gives a comparison of AE between our method and method in [33] at $\alpha = 0.9$. Table 6 gives a comparison of L_∞-error between our method and methods in [33] at different values of α. Figure 5 shows the AE (left) and approximate solution (right) at $\alpha = 0.3, N = 12$.

Table 5: Comparison of AE for Example 3 at $\alpha = 0.9$

	Method in [33]		
x	$N = 200, \Delta t = 0.0005$	$N = 80, \Delta t = 0.001$	Our method at $N = 12$
0.1	6.18496×10^{-9}	2.23568×10^{-7}	5.55112×10^{-17}
0.2	7.14787×10^{-9}	3.99993×10^{-7}	1.11022×10^{-16}
0.3	3.49621×10^{-9}	5.30327×10^{-7}	2.22045×10^{-16}
0.4	3.99695×10^{-9}	6.14693×10^{-7}	2.22045×10^{-16}
0.5	1.42437×10^{-8}	6.52199×10^{-7}	2.22045×10^{-16}
0.6	2.56063×10^{-8}	6.40791×10^{-7}	0
0.7	3.55155×10^{-8}	5.76962×10^{-7}	2.22045×10^{-16}
0.8	3.98360×10^{-8}	4.55240×10^{-7}	0
0.9	3.17808×10^{-8}	2.67253×10^{-7}	0

Table 6: Comparison of L_∞-error of Example 3

	Method in [33]		
α	$N = 2^7$, $M = 2^{11}$	$M = 2^7$, $N = 2^4$	Our method at $N = 12$
0.2	5.69465×10^{-7}	1.97043×10^{-5}	2.22045×10^{-16}
0.3	5.66088×10^{-7}	1.99857×10^{-5}	2.22045×10^{-16}
0.4	5.61387×10^{-7}	2.03716×10^{-5}	2.22045×10^{-16}
0.5	5.54967×10^{-7}	2.09231×10^{-5}	2.22045×10^{-16}

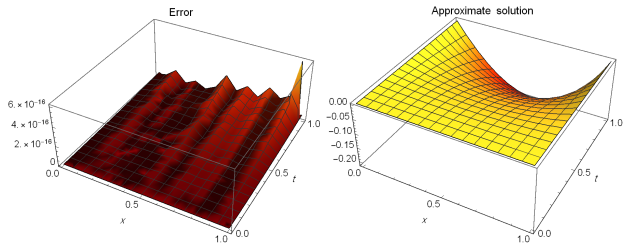


Figure 5: The AE (left) and approximate solution (right) for Example 3 at $\alpha = 0.3, N = 12$.

Example 4. Consider the following time-fractional Burgers' equation of two dimensional

$$\frac{\partial^{\alpha} u(x,y,t)}{\partial t^{\alpha}} + u(x,y,t)\left(\frac{\partial u(x,y,t)}{\partial x} + \frac{\partial u(x,y,t)}{\partial y}\right) - \frac{\partial^2 u(x,y,t)}{\partial x^2} - \frac{\partial^2 u(x,y,t)}{\partial y^2} = g(x,y,t), \tag{61}$$

subject to the following initial and boundary conditions:

$$\begin{aligned} &u(x,y,0) = 0, \quad 0 < x,y \le 1, \\ &u(0,y,t) = t^2 e^{-y}, \quad u(1,y,t) = t^2 e^{1-y}, \quad 0 < y,t \le 1, \\ &u(x,0,t) = t^2 e^{x}, \quad u(x,1,t) = t^2 e^{x-1}, \quad 0 < x,t \le 1, \end{aligned} \tag{62}$$

where $g(x,t) = \frac{2e^{x-y}}{\Gamma(3-\alpha)} t^{2-\alpha} - 2t^2 e^{x-y}$ and $u(x,t) = t^2 e^{x-y}$ is the exact solution of this problem.

Table 7 illustrates the AE at different values of t when $N = 6$ and $\alpha = 0.2$. Figure 6 shows the AE at different values of t at $N = 6$ and $\alpha = 0.2$.

Table 7: The AE of Example 4 at $\alpha = 0.2$

$x = y$	$t = 0.3$	$t = 0.6$	$t = 0.9$
0.2	5.50355×10^{-14}	2.19933×10^{-13}	4.94896×10^{-13}
0.4	4.41578×10^{-14}	1.7732×10^{-13}	4.01347×10^{-13}
0.6	4.40518×10^{-14}	1.75651×10^{-13}	3.92865×10^{-13}
0.8	5.50272×10^{-14}	2.19789×10^{-13}	4.94157×10^{-13}

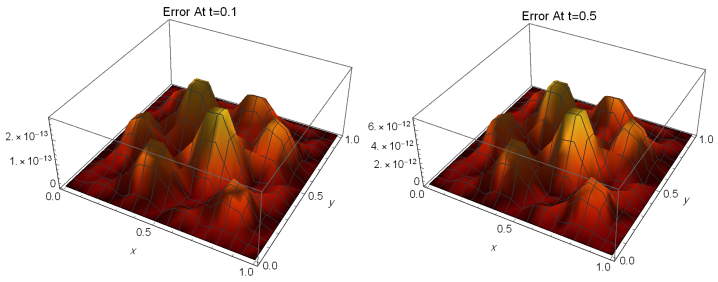


Figure 6: The AE for Example 4 at $\alpha = 0.2, N = 6$.

6 Concluding remarks

In this study, using a special combination of shifted Chebyshev polynomial bases, we built an explicit modal numerical solution based on the spectral Petrov–Galerkin technique to handle the nonlinear time-fractional Burger-type PDE in the Caputo sense. The procedure reduces the issue to a set of nonlinear algebraic equations. Numerous important Chebyshev polynomial characteristics were reported, along with various connection and linearization equations that were mentioned and verified. All components of the resultant matrices were also elegantly assessed. Additionally, studies of convergence and error were created. The applicability and accuracy of the suggested technique were illustrated through a number of illustrative instances, which also show the absolute and anticipated error rates. As an expected future work, we aim to employ the developed theoretical results in this paper along with suitable spectral methods to treat some other problems, for instance, [21, 3, 30]. All codes were written and debugged by *Mathematica* 11 on an HP Z420 Workstation, Processor: Intel (R) Xeon(R) CPU E5-1620 - 3.6 GHz, 16 GB RAM DDR3, and 512 GB storage.

Conflict of Interests

The authors have no conflicts of interest to declare.

Data Availability

No data associated with this research.

References

[1] Abd-Elhameed, W.M., Machado, J.A.T. and Youssri, Y.H. *Hypergeometric fractional derivatives formula of shifted Chebyshev polynomials: tau algorithm for a type of fractional delay differential equations*, Int. J. Nonlinear Sci. Numer. 23(7-8) (2022), 1253–1268.

[2] Abd-Elhameed, W.M., Doha, E.H., Youssri, Y.H. and Bassuony, M.A. *New TChebyshev-Galerkin operational matrix method for solving linear and nonlinear hyperbolic telegraph type equations*, Numer. Methods Partial Differ. Equ. 32(6) (2016), 1553–1571.

[3] Abd-Elhameed, W.M., Youssri, Y.H., Amin, A.K. and Atta, A.G. *Eighth-kind Chebyshev polynomials collocation algorithm for the nonlinear time-fractional generalized Kawahara equation*, Fractal Fract. 7(9) (2023), 652.

[4] Askey, R. *Orthogonal polynomials and special functions*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1975.

[5] Atangana, A. and Owolabi, K.M. *New numerical approach for fractional differential equations*, Math. Model. Nat. Phenom. 13(1) (2018), 3.

[6] Atta, A.G., Abd-Elhameed, W.M. and Youssri, Y.H. *Shifted fifth-kind Chebyshev polynomials Galerkin-based procedure for treating fractional diffusion-wave equation*, Int. J. Mod. Phys. C. 33(08) (2022), 2250102.

[7] Atta, A.G., Abd-Elhameed, W.M., Moatimid, G.M. and Youssri, Y.H. *Shifted fifth-kind Chebyshev Galerkin treatment for linear hyperbolic first-order partial differential equations*, Appl. Numer. Math. 167 (2021), 237–256.

[8] Atta, A.G., Abd-Elhameed, W.M., Moatimid, G.M. and Youssri, Y.H. *Advanced shifted sixth-kind Chebyshev tau approach for solving linear one-dimensional hyperbolic telegraph type problem*, Math. Sci. (2022), 1–15.

[9] Atta, A.G., Abd-Elhameed, W.M., Moatimid, G.M. and Youssri, Y.H. *A fast Galerkin approach for solving the fractional Rayleigh–Stokes problem via sixth-kind Chebyshev polynomials*, Mathematics 10(11) (2022), 1843.

[10] Atta, A.G., Abd-Elhameed, W.M., Moatimid, G.M. and Youssri, Y.H. *Novel spectral schemes to fractional problems with nonsmooth solutions*, Math. Methods Appl. Sci. 46(13) (2023), 14745–14764.

[11] Bernardi, C. and Maday, Y. *Spectral methods*, Handbook of numerical analysis, Vol. V, 209–485, Handb. Numer. Anal., V, North-Holland, Amsterdam, 1997.

[12] 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.

[13] Boyd, J.P. *Chebyshev and Fourier spectral methods*, Chebyshev and Fourier spectral methods. Second edition. Dover Publications, Inc., Mineola, NY, 2001.

[14] Canuto, C., Hussaini, M.Y., Quarteroni, A. and Zang, T.A. *Spectral methods: fundamentals in single domains*, Springer Science & Business Media, 2007.

[15] Chawla, R., Deswal, K.,Kumar, D. and Baleanu, D. *Numerical simulation for generalized time-fractional Burgers' equation with three distinct linearization schemes*, J. Comput. Nonlinear Dyn. 18(4) (2023), 041001.

[16] Costabile, F. and Napoli, A. *A class of collocation methods for numerical integration of initial value problems*, Computers & Mathematics with Applications 62(8) (2011), 3221–3235.

[17] Doha, E.H., Abd-Elhameed, W.M., Elkot, N.A. and Youssri, Y.H. *Integral spectral TChebyshev approach for solving space Riemann-Liouville and Riesz fractional advection-dispersion problems*, Advances in Difference Equations 2017(1) (2017), 1–23.

[18] Doley, S., Kumar, A.V., Singh, K.R. and Jino, L. *Study of Time Fractional Burgers' Equation using Caputo, Caputo-Fabrizio and Atangana-Baleanu Fractional Derivatives*, Engineering Letters 30(3) (2022).

[19] Gottlieb, D. and Orszag, S.A. *Numerical analysis of spectral methods: theory and applications*, CBMS-NSF Regional Conference Series in Applied Mathematics, No. 26. Society for Industrial and Applied Mathematics, Philadelphia, PA, 1977.

[20] Guo, B. *Spectral methods and their applications*, World Scientific Publishing Co., Inc., River Edge, NJ, 1998.

[21] Karaagac, B. and Owolabi, K.M. *Numerical analysis of polio model: A mathematical approach to epidemiological model using derivative with Mittag–Leffler Kernel*, Math. Methods Appl. Sci. 46(7) (2023), 8175–8192.

[22] Karaagac, B., Owolabi, K.M. and Pindza, E. *A computational technique for the Caputo fractal-fractional diabetes mellitus model without genetic factors*, International Journal of Dynamics and Control (2023), 1–18.

[23] Karaagac, B., Esen, A., Owolabi, K.M. and Pindza, E. *A collocation method for solving time fractional nonlinear Korteweg–de Vries–Burgers equation arising in shallow water waves*, Int. J. Mod. Phys. C. 34(07) (2023), 2350096.

[24] Korkmaz, E. and Yildirim, K. *A meshfree time-splitting approach for the time-fractional Burgers' equation*, Journal of Mathematics (2023), 2023.

[25] Magdy, E., Abd-Elhameed, W.M., Youssri, Y.H., Moatimid, G.M. and Atta, A.G. *A potent collocation approach based on shifted gegenbauer polynomials for nonlinear time fractional Burgers' equations*, Contemp. Math. 4(2) (2023), 647–665.

[26] Moghaddam, B.P., Dabiri, A., Lopes, A.M. and Machado, J.A.T. *Numerical solution of mixed-type fractional functional differential equations using modified Lucas polynomials*, Comput. Appl. Math. 38 (2019), 1–12.

[27] 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.

[28] Najafi, M., Arefmanesh, A. and Enjilela, V. *Meshless local Petrov–Galerkin method-higher Reynolds numbers fluid flow applications*, Eng. Anal. Bound. Elem. 36(11) (2012), 1671–1685.

[29] Owolabi, K.M. *Numerical approach to chaotic pattern formation in diffusive predator–prey system with Caputo fractional operator*, Numer. Methods Partial Differ. Equ. 37(1) (2021), 131–151.

[30] Owolabi, K.M. *Analysis and numerical simulation of cross reaction–diffusion systems with the Caputo–Fabrizio and Riesz operators*, Numer. Methods Partial Differ. Equ. 39(3) (2023), 1915–1937.

[31] Peng, X., Xu, D. and Qiu, W. *Pointwise error estimates of compact difference scheme for mixed-type time-fractional Burgers' equation*, Math. Comput. Simul. 208 (2023), 702–726.

[32] 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.

[33] Shafiq, M., Abbas, M., Abdullah, F.A., Majeed, A., Abdeljawad, T. and Alqudah, M.A. *Numerical solutions of time fractional Burgers' equation involving Atangana–Baleanu derivative via cubic B-spline functions*, Results Phys. 34 (2022), 105244.

[34] Shen, J. *A new dual-Petrov–Galerkin method for third and higher odd-order differential equations: application to the KdV equation*, SIAM J. Numer. Anal. 41(5) (2003), 1595–1619.

[35] Trefethen, L.N. *Spectral methods in MATLAB*, Software, Environments, and Tools, 10. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2000.

[36] Wang, H., Sun, Y., Qian, X. and Song, S. *A high-order compact difference scheme on graded mesh for time-fractional Burgers' equation*, Comput. Appl. Math. 42(1) (2023), 18.

[37] Yadav, S. and Pandey, R.K. *Numerical approximation of fractional Burgers equation with Atangana–Baleanu derivative in Caputo sense*, Chaos Solitons Fractals 133 (2020), 109630.

[38] Youssri, Y.H. and Atta, A.G. *Double TChebyshev spectral tau algorithm for solving KdV equation, with soliton application*, Solitons, New York, NY: Springer US, 2022, 451–467.

[39] Youssri , Y.H. and Atta, A.G. *Petrov-Galerkin Lucas polynomials procedure for the time-fractional diffusion equation*, *Contemp. Math.* (2023), 230–248.

[40] 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*, Fractal Fract. 7(2) (2023), 133.

[41] Youssri, Y.H. and Muttardi, M.M. *A mingled tau-finite difference method for stochastic first-order partial differential equations*, Int. J. Appl. Comput. Math. 9(2) (2023), 14.

[42] Youssri, Y.H., Abd-Elhameed, W.M. and Abdelhakem, M. *A robust spectral treatment of a class of initial value problems using modified Chebyshev polynomials*, Math. Methods Appl. Sci. 44(11) (2021), 9224–9236.

[43] Yuan, J., Shen, J. and Wu, J. *A dual-Petrov-Galerkin method for the Kawahara-type equations*, J. Sci. Comput. 34(1) (2008), 48–63.

[44] Zhang, Y. and Feng, M. *A local projection stabilization virtual element method for the time-fractional Burgers equation with high Reynolds numbers*, Appl. Math. Comput. 436 (2023), 127509.

Iranian Journal of Numerical Analysis and Optimization
Vol. 14, No. 1, 2024, pp 200–218
https://doi.org/10.22067/ijnao.2023.83361.1290
https://ijnao.um.ac.ir/

CC BY NC

Research Article

Minimization of sub-topical functions over a simplex

M.H. Daryaei*, and M.A. Yaghoobi

Abstract

This article investigates a particular version of the cutting angle method for finding the global minimizer of sub-topical (increasing and plus sub-homogeneous) functions over a simplex. The algorithm is based on the abstract convexity of sub-topical functions. Furthermore, we discuss the proof of convergence of the algorithm and provide results from numerical experiments.

AMS subject classifications (2020): 65K05, 90C30, 90C26.

Keywords: Abstract convexity; Global optimization; Sub-topical functions; Cutting angle method.

*Corresponding author

Received 11 July 2023; revised 21 October 2023; accepted 28 October 2023

Mohammad Hossein Daryaei
Department of Applied Mathematics, Faculty of Mathematics and Computer, Shahid Bahonar University of Kerman, Kerman, Iran. e-mail: daryaei@uk.ac.ir

Mohammad Ali Yaghoobi
Department of Applied Mathematics, Faculty of Mathematics and Computer, Shahid Bahonar University of Kerman, Kerman, Iran. e-mail: yaghoobi@uk.ac.ir

How to cite this article
Daryaei, M.H. and Yaghoobi, M.A., Minimization of sub-topical functions over a simplex. *Iran. J. Numer. Anal. Optim.*, 2024; 14(1): 200-218. https://doi.org/10.22067/ijnao.2023.83361.1290

1 Introduction

The cutting angle method is a powerful technique that can solve a wide range of global optimization problems based on abstract convexity. This method was proposed in 1999 as a method of global Lipschitz optimization and is a deterministic global optimization technique that involves constructing a sequence of lower approximations to an objective function [2]. In fact, this method was developed as a generalization of the cutting plane method of convex minimization. The original optimization problem is replaced by a sequence of relaxed auxiliary problems that are based on the lower approximations. Through this process, the sequence of global minimum solutions of the relaxed problems eventually converges to the global minimum of the original objective function. The articles [1, 3] discussed a method for globally minimizing an increasing positively homogeneous function over the unit simplex $S = \{x = (x_1, \ldots, x_n) \in \mathbb{R}^n_+ : \sum_{i=1}^n x_i = 1\}$. They provided a thorough examination of this method. In addition, Ferrer, Bagirov, and Beliakov [6] employed the cutting angle method to solve the problem of minimizing the difference of two convex functions.

A function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is classified as topical if it is increasing in the natural partial ordering of $\mathbb{R}^n$ and plus-homogeneous, meaning that $f(x + \lambda\mathbf{1}) = f(x) + \lambda$ for all $x \in \mathbb{R}^n$ and all $\lambda \in \mathbb{R}$, where $\mathbf{1}$ denotes the vector of the corresponding dimension with all components equal to one. Topical functions have been studied in [7, 8, 10] and have found numerous applications in different areas of applied mathematics, particularly in the modeling of discrete event systems (see [7]).

Functions that are both increasing and plus-sub-homogeneous are referred to as sub-topical functions, which are viewed as a natural extension of topical functions [4]. The class of sub-topical functions is one of the most important classes of abstract convex functions. In [8, 10], the authors showed that every sub-topical function defined on $\mathbb{R}^n$ can be characterized as an abstract convex function with respect to a set L that is called the set of elementary functions. For each $y \in \mathbb{R}^n$ and each $\alpha \in \mathbb{R}$, the function $\varphi_{(y,\alpha)} : \mathbb{R}^n \longrightarrow \mathbb{R}$ is defined by

$$\varphi_{(y,\alpha)}(x) = \sup\{\lambda \in \mathbb{R} : \lambda \leq \alpha,\ \lambda\mathbf{1} \leq x + y\} \quad \text{for all } x \in \mathbb{R}^n,$$

where

$$L = \{\varphi_{(y,\alpha)} : y \in \mathbb{R}^n,\ \alpha \in \mathbb{R}\}.$$

It is worth noting that L is a more general class of functions than the class of min-type functions [3, 9].

Here, we study the problem

$$\min_{x \in S_a} f(x),$$

where $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is a sub-topical function and S_a is a simplex in $\mathbb{R}^n$; that is, $S_a = \{x = (x_1, \ldots, x_n) \in \mathbb{R}^n_+ : \sum_{i=1}^n \frac{x_i}{a_i} = 1\}$, where $a = (a_1, \ldots, a_n) \in \mathbb{R}^n_{++}$. We developed a version of the cutting angle method to solve this problem.

The rest of this paper is organized as follows. Section 2 presents definitions and preliminary results on sub-topical functions. In Section 3, a cutting angle method is proposed for finding the global minimizers of sub-topical functions over a simplex. Then, we present an approach to the numerical solution of the subproblem of the algorithm. Finally, Section 4 presents the results of numerical experiments.

2 Preliminaries

Let $\mathbb{R}^n_+$ denote the cone of all n-vectors with nonnegative coordinates. We assume that the usual order relation is defined in this cone as follows. For $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ and $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$, $x \leq y$ if and only if $x_i \leq y_i$ for all i. A function $f : \mathbb{R}^n \longrightarrow \bar{\mathbb{R}} = [-\infty, +\infty]$ is said to be increasing if $f(x) \leq f(y)$ for each x, $y \in \mathbb{R}^n$ such that $x \leq y$. The function f is called plus sub-homogeneous if $f(x + \lambda \mathbf{1}) \leq f(x) + \lambda$ for all $x \in \mathbb{R}^n$ and all $\lambda \geq 0$, where $\mathbf{1} = (1, \ldots, 1) \in \mathbb{R}^n$. It is easy to see that f is plus sub-homogeneous if and only if $f(x + \lambda \mathbf{1}) \geq f(x) + \lambda$ for all $x \in \mathbb{R}^n$ and all $\lambda \leq 0$. The definitions and results that follow can be found in [9, 10].

Definition 1. A function $f : \mathbb{R}^n \longrightarrow \bar{\mathbb{R}}$ is called sub-topical if it is increasing and plus sub-homogeneous.

Remark 1. A function $f : \mathbb{R}^n \longrightarrow \bar{\mathbb{R}}$ is called topical if it is increasing and $f(x+\lambda\mathbf{1}) = f(x)+\lambda$ for all $x \in \mathbb{R}^n$ and all $\lambda \in \mathbb{R}$. It is clear that any topical function is sub-topical.

Lemma 1 (see [10])**.** Let $f : \mathbb{R}^n \longrightarrow \bar{\mathbb{R}}$ be a sub-topical function. The following statements hold:

(i) If $f(x) = +\infty$ for some $x \in \mathbb{R}^n$, then f is identically equal to $+\infty$.

(ii) If $f(x) = -\infty$ for some $x \in \mathbb{R}^n$, then f is identically equal to $-\infty$.

Lemma 1 implies that a sub-topical function is either finite (i.e., finite-valued at each $x \in \mathbb{R}^n$) or identically $+\infty$ or $-\infty$. Therefore, in the remainder of the paper, we only consider finite-valued sub-topical functions. Here are some simple examples of finite-valued sub-topical functions.

Example 1. Any sub-linear function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ such that $f(\mathbf{1}) \leq 1$ is plus sub-homogeneous. So, an increasing and sub-linear function is sub-topical.

Example 2. Any function of the form

$$f(x) = \frac{1}{\theta} \ln \big(\sum_{i=1}^{n} e^{\langle a_i, x \rangle} \big), \quad x \in \mathbb{R}^n,$$

where $a_i \in \mathbb{R}^n$, $a_i \geq 0$, $i = 1, 2, \ldots, n$, and $\theta \geq \max_{1 \leq i \leq n} \langle a_i, \mathbf{1} \rangle$, is sub-topical. Indeed,

$$\begin{aligned} f(x+\lambda\mathbf{1}) &= \frac{1}{\theta} \ln \big(\sum_{i=1}^{n} e^{\langle a_i, x+\lambda\mathbf{1} \rangle} \big) \\ &= \frac{1}{\theta} \ln \big(\sum_{i=1}^{n} e^{\langle a_i, x \rangle} e^{\lambda \langle a_i, \mathbf{1} \rangle} \big) \\ &\leq \frac{1}{\theta} \ln \big(e^{\lambda\theta} \sum_{i=1}^{n} e^{\langle a_i, x \rangle} \big) \\ &= \frac{1}{\theta} \Big(\ln \big(e^{\lambda\theta} \big) + \ln \big(\sum_{i=1}^{n} e^{\langle a_i, x \rangle} \big) \Big) \\ &= \lambda + \frac{1}{\theta} \ln \big(\sum_{i=1}^{n} e^{\langle a_i, x \rangle} \big) \\ &= \lambda + f(x). \end{aligned}$$

The set $\mathbf{\Gamma}$ of all sub-topical functions $f : \mathbb{R}^n \longrightarrow \bar{\mathbb{R}}$ has several notable properties, which we will mention below.

(1) We have $\mathbf{\Gamma} + \mathbb{R} = \mathbf{\Gamma}$, that is, if $f \in \mathbf{\Gamma}$ and $\gamma \in \mathbb{R}$, then $f + \gamma \in \mathbf{\Gamma}$.

(2) $\mathbf{\Gamma}$ is a convex set.

(3) Let $f_i \in \mathbf{\Gamma}$ for $1 \leq i \leq K$, and let

$$\hat{f}(x) = \min_{1 \leq i \leq K} f_i(x), \quad \tilde{f}(x) = \max_{1 \leq i \leq K} f_i(x), \quad (x \in \mathbb{R}^n).$$

Then the functions $\hat{f}$ and $\tilde{f}$ belong to $\mathbf{\Gamma}$.

(4) $\mathbf{\Gamma}$ is a complete lattice, that is, if $\{f_\beta\}_{\beta \in B}$ is an arbitrary family of elements of $\mathbf{\Gamma}$ and

$$f(x) = \sup_{\beta \in B} f_\beta(x), \quad (x \in \mathbb{R}^n),$$

then the function f belongs to $\mathbf{\Gamma}$.

Remark 2. If $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ is a sub-topical function, then f is continuous on $\mathbb{R}^n$. Indeed, consider an arbitrary point $x \in \mathbb{R}^n$ and a sequence $\{x_k\} \subseteq \mathbb{R}^n$ that $x_k \longrightarrow x$. Let $\epsilon > 0$ be arbitrary. Then, there exists $k_0 \geq 1$ such that $x - \epsilon \mathbf{1} \leq x_k \leq x + \epsilon \mathbf{1}$ for all $k \geq k_0$. Since f is increasing and plus sub-homogeneous, it follows that

$$f(x) - \epsilon \leq f(x - \epsilon \mathbf{1}) \leq f(x_k) \leq f(x + \epsilon \mathbf{1}) \leq f(x) + \epsilon \quad \text{for all } k \geq k_0.$$

This implies that $f(x_k) \longrightarrow f(x)$, which shows that f is continuous at x.

Now, we recall some definitions from abstract convexity (see for more details [9, 12]). Consider a set X and a set H of functions $h : X \longrightarrow \bar{\mathbb{R}}$. The function $f : X \longrightarrow \bar{\mathbb{R}}$ is called abstract convex with respect to H (or H-convex) if there exists a subset U of H such that

$$f(x) = \sup_{h \in U} h(x), \quad (x \in X).$$

The set H is called the set of elementary functions. Consider the function $\varphi : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R} \longrightarrow \mathbb{R}$ defined by

$$\varphi(x,y,\alpha)=\min_{1\le i\le n}\{\alpha,x_i+y_i\},\quad (x,\ y\in\mathbb{R}^n,\ \alpha\in\mathbb{R}).$$

Let $(y,\alpha)\in\mathbb{R}^n\times\mathbb{R}$ be arbitrary. Denote by $\varphi_{(y,\alpha)}$ the function defined on $\mathbb{R}^n$ by the formula $\varphi_{(y,\alpha)}(x)=\varphi(x,y,\alpha)$. Let $X_\varphi=\{\varphi_{(y,\alpha)}\ :\ y\in\mathbb{R}^n,\ \alpha\in\mathbb{R}\}$. Then it is known that any function f defined on $\mathbb{R}^n$ is sub-topical if and only if f is X_φ-convex (see for more details [5, Theorem 3.2]).

Theorem 1. (see [5, Theorem 3.1]) The function $f:\mathbb{R}^n\longrightarrow\bar{\mathbb{R}}$ is sub-topical if and only if it satisfies the following inequality:

$$f(x)\ge\varphi_{(y,\alpha)}(x)+f(-y+\alpha\mathbf{1})-\alpha\quad\text{for all }\ x,\ y\in\mathbb{R}^n\ \text{ for all }\ \alpha\in\mathbb{R}.\ (1)$$

Inequality (1) implies the following statement.

Proposition 1. (see [5, Theorem 3.2]) Let $f:\mathbb{R}^n\longrightarrow\mathbb{R}$ be a sub-topical function and let $\bar{x}\in\mathbb{R}^n$. Put $\bar{y}=-\bar{x}+f(\bar{x})\mathbf{1}$ and $\bar{\alpha}=f(\bar{x})$. Then $\varphi_{(\bar{y},\bar{\alpha})}(x)\le f(x)$ for all $x\in\mathbb{R}^n$ and $\varphi_{(\bar{y},\bar{\alpha})}(\bar{x})=f(\bar{x})$.

Remark 3. Let $\bar{x}\in S_a$, $\bar{y}=-\bar{x}+f(\bar{x})\mathbf{1}$ and let $\bar{\alpha}=f(\bar{x})$. Define the function $\varphi_{\bar{y}}$ as follows:

$$\varphi_{\bar{y}}(x)=\min_{1\le i\le n}\{x_i+\bar{y}_i\},\quad (x\in S_a).\qquad(2)$$

Then for each $x\in S_a$ we have $\varphi_{(\bar{y},\bar{\alpha})}(x)=\varphi_{\bar{y}}(x)$. Indeed, since $\bar{x},\ x\in S_a$, we get $\sum_{i=1}^n\frac{x_i-\bar{x}_i}{a_i}=0$. This implies that there exists $1\le i_0\le n$ such that $x_{i_0}-\bar{x}_{i_0}\le 0$. Therefore

$$x_{i_0}+\bar{y}_{i_0}=x_{i_0}-\bar{x}_{i_0}+f(\bar{x})\le f(\bar{x})=\bar{\alpha}.$$

Hence, $\min_{1\le i\le n}\{\bar{\alpha},x_i+\bar{y}_i\}=\min_{1\le i\le n}\{x_i+\bar{y}_i\}$.

3 Cutting angle method

We now present the cutting angle method for finding a global minimizer of the problem

$$\min_{x\in S_a}f(x),\qquad(3)$$

where $f:\mathbb{R}^n\longrightarrow\mathbb{R}$ is a sub-topical function and $S_a=\{x=(x_1,\dots,x_n)\in\mathbb{R}^n_+:\sum_{i=1}^n\frac{x_i}{a_i}=1\}$, where $a=(a_1,\dots,a_n)\in\mathbb{R}^n_{++}$. The ith unit vector is

denoted by $\mathbf{e}_i = (0, \ldots, 0, 1, 0, \ldots, 0)$. It is worth noting that, if $x^i = a_i\mathbf{e}_i$, $y^i = -x^i + f(x^i)\mathbf{1}$ and $\alpha_i = f(x^i)$, then $\varphi_{(y^i,\alpha_i)}(x) = \varphi_{y^i}(x) = x_i + y^i_i = x_i + f(x^i) - a_i$ for all $x = (x_1, \ldots, x_n) \in S_a$.

Algorithm 2:

Step **0:** (initialization)

a) Take points $x^i = a_i\mathbf{e}_i$ for $i = 1, \ldots, n$, and for each $1 \leq i \leq n$, construct the points $y^i = -x^i + f(x^i)\mathbf{1}$.

b) Define the function $h_n(x) = \max_{1\leq i\leq n} \varphi_{y^i}(x) = \max_{1\leq i\leq n}(x_i + y^i_i)$ for all $x = (x_1, \ldots, x_n) \in S_a$.

c) Set $k := n$.

Step **1:** Solve the following subproblem:

$$\min_{x\in S_a} h_k(x). \tag{4}$$

Step **2:** Let y^* be an optimal solution of Problem (4). Set $k := k + 1$ and $x^k = y^*$.

Step **3:** Compute $y^k = -x^k + f(x^k)\mathbf{1}$. Define the function

$$h_k(x) = \max(h_{k-1}(x),\ \varphi_{y^k}(x)) = \max_{1\leq j\leq k} \min_{1\leq i\leq n} (x_i + y^j_i)$$

for all $x = (x_1, \ldots, x_n) \in S_a$. Go to Step **1**.

In Algorithm 2, the function h_n is the first lower approximation of the objective function f over S_a (note that the objective function f is from $\mathbb{R}^n$ to $\mathbb{R}$). This algorithm produces the sequence $\{h_k\}_{k\geq n}$ of lower approximations of f such that

$$h_k(x) \leq h_{k+1}(x) \quad \text{for all}\, k \geq n \ \text{ for all } \ x \in S_a.$$

Note that Proposition 1 implies that $h_k(x) \leq f(x)$ for all $x \in S_a$ and all $k \geq n$.

Algorithm 2 is indeed a version of the cutting angle method, which is a powerful technique for global optimization of sub-topical functions over a

simplex. By a number of iterations of this algorithm, it produces x^k as an approximation of the global minimizer of f over S_a.

Let $\lambda_k = \min_{x\in S_a} h_k(x)$, and let $f(x^*) = \min_{x\in S_a} f(x)$. It is clear that $\lambda_k \le f(x^*)$ for all $k \ge n$.

Lemma 2. In Algorithm 2, let $k \ge n$. Then $y_i^i \le y_i^k$ for all $1 \le i \le n$.

Proof. By Proposition 1 we have $\varphi_{y^i}(x^k) \le f(x^k)$ for all $1 \le i \le n$. So

$$x_i^k + f(x^i) - a_i = x_i^k + y_i^i \le f(x^k), \quad i = 1, \dots, n.$$

This implies that $y_i^i \le y_i^k$ for all $1 \le i \le n$. □

Proposition 2. In Algorithm 2, let $k \ge n+1$. If $y_j^j = y_j^k$ for some $1 \le j \le n$, then x^k is a global minimizer of the Problem (3).

Proof. We have

$$\begin{aligned}
\lambda_{k-1} &= h_{k-1}(x^k) \\
&= \max_{1\le i\le k-1} \varphi_{y^i}(x^k) \\
&\ge \varphi_{y^j}(x^k) \\
&= y_j^j + x_j^k \\
&= y_j^k + x_j^k \\
&\ge \varphi_{y^k}(x^k) \\
&= f(x^k) \\
&\ge f(x^*) \\
&\ge \lambda_{k-1}.
\end{aligned}$$

Hence $\lambda_{k-1} = f(x^k) = f(x^*)$. This completes the proof. □

Proposition 3. In Algorithm 2, let $k \ge n+2$. If $y^k \le y^j$ for some $n+1 \le j \le k-1$, then x^k is a global minimizer of the Problem (3).

Proof. We have $\lambda_{k-1} = h_{k-1}(x^k) \ge \varphi_{y^j}(x^k) \ge \varphi_{y^k}(x^k) = f(x^k) \ge f(x^*) \ge \lambda_{k-1}$. Therefore $\lambda_{k-1} = f(x^k) = f(x^*)$, and so x^k is a global minimizer of the function f over S_a. □

If Algorithm 2 does not terminate after a finite iteration, then we show that the sequence $\{x^k\}$ generated by it converges to a global minimizer of the sub-topical function f on S_a. To this end, we first state and prove some results. Note that the sequence of functions $\{f_k\}_{k\geq 1}$, defined on the set $X \subseteq \mathbb{R}^n$, is called equicontinuous on X if for every $\epsilon > 0$ there exists $\delta > 0$ such that $|f_k(x) - f_k(t)| < \epsilon$ whenever $\|x - t\| < \delta$, $x \in X$, $t \in X$, and $k \geq 1$.

Lemma 3. The sequence $\{\varphi_{y^k}\}_{k\geq 1}$ generated by Algorithm 2 is equicontinuous on S_a.

Proof. Let $\epsilon > 0$ be given and let $k \geq 1$. Put $\delta = \epsilon$, and let $x = (x_1, \ldots, x_n)$, $t = (t_1, \ldots, t_n) \in S_a$ such that $\|x - t\| < \delta$. Therefore

$$|x_i - t_i| < \epsilon, \quad i = 1, \ldots, n. \tag{5}$$

Now, assume that $\min_{1\leq i\leq n}(x_i - x_i^k) = x_{i_k} - x_{i_k}^k$ and $\min_{1\leq i\leq n}(t_i - x_i^k) = t_{i'_k} - x_{i'_k}^k$ for some $i_k,\ i'_k \in \{1, \ldots, n\}$. We conclude that

$$x_{i_k} - t_{i_k} \leq \varphi_{y^k}(x) - \varphi_{y^k}(t) \leq x_{i'_k} - t_{i'_k}.$$

It follows from (5) that $|\varphi_{y^k}(x) - \varphi_{y^k}(t)| < \epsilon$. Hence the sequence $\{\varphi_{y^k}\}$ is equicontinuous on S_a. □

Lemma 4. The sequence $\{h_k\}_{k\geq n}$ generated by Algorithm 2 is equicontinuous on S_a.

Proof. Let $\epsilon > 0$ be given. Put $\delta = \epsilon$, and let $x = (x_1, \ldots, x_n)$, $t = (t_1, \ldots, t_n) \in S_a$ such that $\|x - t\| < \delta$. By Lemma 3, we have

$$|\varphi_{y^k}(x) - \varphi_{y^k}(t)| < \epsilon, \quad k = 1, 2, \ldots. \tag{6}$$

Let $k \geq n$. Now, assume that $h_k(x) = \max_{1\leq i\leq k} \varphi_{y^i}(x) = \varphi_{y^{i_0}}(x)$ and $h_k(t) = \max_{1\leq i\leq k} \varphi_{y^i}(t) = \varphi_{y^{i_1}}(t)$ for some $i_0,\ i_1 \in \{1, \ldots, k\}$. We conclude that

$$\varphi_{y^{i_1}}(x) - \varphi_{y^{i_1}}(t) \leq h_k(x) - h_k(t) = \varphi_{y^{i_0}}(x) - \varphi_{y^{i_1}}(t) \leq \varphi_{y^{i_0}}(x) - \varphi_{y^{i_0}}(t).$$

It follows from (6) that $|h_k(x) - h_k(t)| < \epsilon$. Hence the sequence $\{h_k\}_{k\geq n}$ is equicontinuous on S_a. □

Remark 4. By the definition of the function h_k, we conclude that the sequence $\{h_k\}$ in Algorithm 2 is also uniformly bounded on S_a. Indeed, we have that the sequence $\{h_k\}$ is increasing and $h_k(x) \leq f(x)$ for all $x \in S_a$ and all $k \geq n$. On the other hand, any sub-topical function is continuous and S_a is a compact set. So we get that the sequence $\{h_k\}$ is uniformly bounded on S_a.

Proposition 4. Let $\{x^k\}$ be the generated sequence by Algorithm 2 and let x^* be a limit point of this sequence. Then x^* is a global minimum point of the function $\varphi(x) = \sup_{k \geq n} h_k(x)$ over S_a.

Proof. Without loss of generality, assume that $x^k \longrightarrow x^*$. Put $\lambda_k = h_k(x^{k+1})$. We have that the sequence $\{\lambda_k\}$ is an increasing sequence and $\lambda_k \leq \varphi(x)$ for all $x \in S_a$. Thus

$$\lim_{k \longrightarrow \infty} \lambda_k \leq \inf_{x \in S_a} \varphi(x) \leq \varphi(x^*). \tag{7}$$

Now, we show that $\lim_{k \longrightarrow \infty} \lambda_k = \varphi(x^*)$. Let $\epsilon > 0$ be given. Since $\{h_k\}$ is equicontinuous on S_a and $x^k \longrightarrow x^*$, there exists $N_1 > 0$ such that for each $m \geq N_1$ we have

$$|h_k(x^m) - h_k(x^*)| < \frac{\epsilon}{2} \quad \text{for all } \ k \geq n. \tag{8}$$

On the other hand, It is clear that $h_k(x^*) \longrightarrow \varphi(x^*)$. So there exists $N_2 > n$ such that

$$|h_k(x^*) - \varphi(x^*)| < \frac{\epsilon}{2} \quad \text{for all } \ k \geq N_2. \tag{9}$$

Let $N = \max\{N_1, N_2\}$. It follows from (8) and (9) that

$$|\lambda_k - \varphi(x^*)| \leq |\lambda_k - h_k(x^*)| + |h_k(x^*) - \varphi(x^*)| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

for all $k \geq N$. This implies that $\lim_{k \longrightarrow \infty} \lambda_k = \varphi(x^*)$, and by (7) we conclude that $\varphi(x^*) = \inf_{x \in S_a} \varphi(x)$. This completes the proof. □

We will now demonstrate the convergence of Algorithm 2.

Theorem 2. Let $\{x^k\}$ be the generated sequence by Algorithm 2 and let x^* be a limit point of this sequence. Then x^* is a global optimal solution of the Problem (3).

Proof. Since $\{h_k\}$ is uniformly bounded and equicontinuous on the compact set S_a, then by [11, Theorem 7.25], this sequence has a subsequence that is uniformly convergent on S_a. On the other hand, the sequence $\{h_k\}$ is increasing, so $\{h_k\}$ converges uniformly to $\varphi(x) = \sup_{k\geq n} h_k(x)$ on S_a. It is clear that $\varphi(x) \leq f(x)$ for all $x \in S_a$ and $\varphi(x^k) = f(x^k)$ for all $k \geq n$ (indeed, by Proposition 1, for each $k \geq n$, we have $\varphi_{y^k}(x^k) = f(x^k)$ and thus $h_k(x^k) = f(x^k)$ and $\varphi(x^k) = f(x^k)$). Now, let $x^* \in S_a$ be a limit point of the sequence $\{x^k\}$ (the simplex S_a is a compact set, so $\{x^k\}$ has always a limit point $x^* \in S_a$). We have $\varphi(x) - \varphi(x^k) \leq f(x) - f(x^k)$ for all $x \in S_a$ and all $k \geq n$. This implies that

$$\varphi(x) - \varphi(x^*) \leq f(x) - f(x^*) \quad \text{for all } x \in S_a. \tag{10}$$

By Proposition 4, x^* is a minimizer of φ over S_a. So $\varphi(x) - \varphi(x^*) \geq 0$ for all $x \in S_a$. Therefore, by (10), we conclude that $f(x) - f(x^*) \geq 0$ for all $x \in S_a$. Hence x^* is a global minimizer of the function f over the set S_a. This completes the proof. □

Step 1 of Algorithm 2 involves finding the global minimum of the function h_k on the set S_a, which is the most challenging part of the algorithm. The problem in Step 1 can be formulated as follows:

$$\begin{aligned} &\min\ h_k(x) \\ &\text{subject to } x \in S_a, \end{aligned} \tag{11}$$

where

$$h_k(x) = \max_{1\leq j\leq k} \min_{1\leq i\leq n} (x_i + y_i^j),$$

$k \geq n$, $x = (x_1, \ldots, x_n) \in S_a$, and $y^j = -x^j + f(x^j)\mathbf{1}$.

We will now discuss an approach for solving the subproblem. We demonstrate that Problem (11) can be transformed into a mixed integer linear programming problem with $0-1$ variables. This technique is widely used and involves the introduction of a large positive parameter M. If we set $t = \max_{1\leq j\leq k} \min_{1\leq i\leq n}(x_i + y_i^j)$, then it is clear that Problem (11) is equivalent to the following problem:

$$\begin{aligned} &\min z = t \\ &\text{subject to } \min_{1\le i\le n}(x_i + y_i^j) \le t, \quad 1 \le j \le k, \\ &\qquad x \in S_a. \end{aligned} \tag{12}$$

Each constraint $\min_{1\le i\le n}(x_i + y_i^j) \le t$ can be expressed as a set of $n+1$ linear constraints of the following form

$$\begin{aligned} &x_i + y_i^j - Mu_i^j \le t, \quad 1 \le i \le n, \\ &\sum_{i=1}^{n} u_i^j \le n-1, \end{aligned}$$

where $u_i^j \in \{0,1\}$ for all $i,\ j$.
Hence, we conclude that Problem (12) is equivalent to the following mixed integer linear programming problem:

$$\begin{aligned} &\min z = t \\ &\text{subject to } x_i - t - Mu_i^j \le -y_i^j, \quad 1 \le i \le n, \quad 1 \le j \le k, \\ &\qquad \textstyle\sum_{i=1}^{n} u_i^j \le n-1, \quad 1 \le j \le k, \\ &\qquad \textstyle\sum_{i=1}^{n} \frac{x_i}{a_i} = 1, \\ &\qquad x_i \ge 0, \quad u_i^j \in \{0,1\}, \quad 1 \le i \le n, \quad 1 \le j \le k. \end{aligned} \tag{13}$$

Despite the fact that this technique increases the number of constraints and variables with each iteration, it also has the advantage of allowing the use of software packages designed for solving mixed integer linear programming problems, such as MATLAB, LINGO, and so on, to solve the subproblem.

Remark 5. There is a natural question that, in Problem (13), how big should M be. Let $|f(x)| \le C$ for all $x \in S_a$, and let $A = \max_{1\le i\le n}(a_i)$. Then it is clear that $|x_i + y_i^j| \le A + C$ for all i and j. This implies that, it is enough to assume that $M \ge 2A + 2C$.

At the end of this section, we present the complete version of the cutting angle method for finding the global optimal solution of Problem (3).

Algorithm 3: Cutting angle method

Initialization:

Set $x^i = a_i \mathbf{e}_i$ for $i = 1, \ldots, n$, and for each $1 \leq i \leq n$, compute the points $y^i = -x^i + f(x^i)\mathbf{1}$. Find the optimal solution y^* of Problem (13) according to $k = n$. Set $k := n + 1$ and $F := \emptyset$.

Iteration $k - n$**:**

1) Set $x^k = y^*$, $F := F \cup \{f(x^k)\}$ and compute the point $y^k = -x^k + f(x^k)\mathbf{1}$.

i) Let $f(\hat{x}^k) = \min F$.

ii) If there exists $1 \leq j \leq n$ such that $y^j_j = y^k_j$, then terminate the algorithm; x^k is a global optimal solution of Problem (3).

iii) For $k \geq n + 2$, if there exists $n + 1 \leq j \leq k - 1$ such that $y^k \leq y^j$, then terminate the algorithm; x^k is a global optimal solution of Problem (3).

2) Find the optimal solution y^* of Problem (13). Set $k := k + 1$ and go to iteration $k - n$.

Remark 6. It is clear that, if Algorithm 3 does not terminate after a finite iteration, then the sequence $\{\hat{x}^k\}$ tends to a global optimal solution of Problem (3).

4 Numerical results

In this section, we will provide three numerical examples to evaluate the effectiveness of the proposed cutting angle method. We have implemented the method using MATLAB and carried out the computations on a personal computer with the following specifications: Microprocessor: Intel® Core™ i5-7200U (2.5 GHz, up to 3.1 GHz, 3 MB cache, 2 cores); Memory: 8 GB DDR4-2133.

In each of the examples, we have selected an objective function that is sub-topical. We have then summarized the numerical results in tables, which allow us to easily evaluate the performance of the algorithm.

Example 3. Consider

$$\begin{aligned} &\min f(x_1, x_2) \\ &\text{s.t. } x_1 + x_2 = 1, \\ &\quad x_1, x_2 \geq 0, \end{aligned} \tag{14}$$

where

$$f(x_1, x_2) = \frac{1}{5} \ln(e^{3x_1} + e^{5x_2}).$$

The graph of the objective function f and functions h_i $(i = 2, 3, 4, 5)$ over the unit simplex $S_{\mathbf{1}} = \{(x_1, x_2) : x_1 + x_2 = 1,\ x_1, x_2 \geq 0\}$ are shown in Figure 1. The optimal solution and the optimal value of Problem (14) are $x^* = (0.688853, 0.311147)$ and $f^* = 0.507312$, respectively.

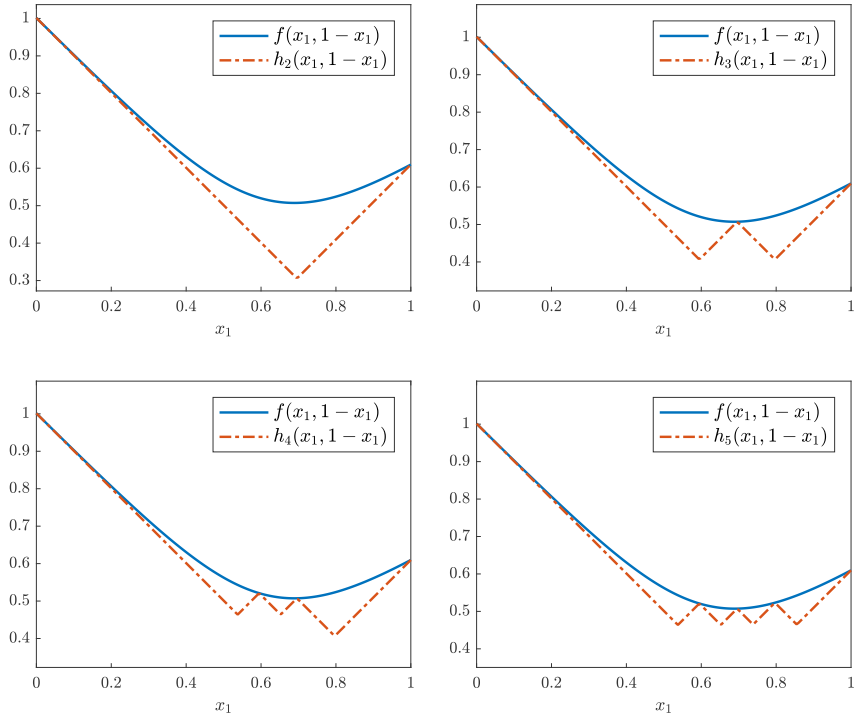


Figure 1: Graph of the objective function f of Problem (14) and the functions h_i $(i = 2, 3, 4, 5)$ over the unit simplex $S_{\mathbf{1}} = \{(x_1, x_2) : x_1 + x_2 = 1,\ x_1, x_2 \geq 0\}$.

Table 1: Obtained results by the cutting angle method for Example 3

Iter	Optimal point found $\hat{x}^k$	Optimal value found $f(\hat{x}^k)$	Time (sec)
1	$(0.695813, 0.304187)$	0.507385	0.01
13	$(0.685517, 0.314483)$	0.507329	0.19
20	$(0.690637, 0.309363)$	0.507317	0.34
34	$(0.688083, 0.311917)$	0.507314	0.73
64	$(0.688721, 0.311279)$	0.507312	2.71

Example 4. Consider

$$\begin{aligned} &\min\ f(x_1, x_2, x_3) \\ &\text{s.t. } x_1 + \tfrac{1}{2}x_2 + \tfrac{1}{3}x_3 = 1, \\ &\quad x_1, x_2, x_3 \geq 0, \end{aligned} \tag{15}$$

where

$$\begin{aligned} f(x_1, x_2, x_3) = 0.1 \max\{&0.2x_1 + 0.3x_2 + 0.5x_3, 0.1x_1 + 0.7x_2 + 0.1x_3, \\ &0.4x_1 + 0.38x_2 + 0.2x_3\} + 0.025 \ln(e^{9x_1} + e^{5x_2} + e^{12x_3}). \end{aligned}$$

Note that the function f is not differentiable on the simplex $S_a = \{(x_1, x_2, x_3) : x_1 + \frac{1}{2}x_2 + \frac{1}{3}x_3 = 1,\ x_1, x_2, x_3 \geq 0\}$. The graph of the objective function f over the simplex S_a is shown in Figure 2. The optimal solution and the optimal value of Problem (15) are $x^* = (0.5658, 0.6699, 0.2978)$ and $f^* = 0.1911$, respectively.

Table 2: Obtained results by the cutting angle method for Example 4

Iter	Optimal point found $\hat{x}^k$	Optimal value found $f(\hat{x}^k)$	Time (sec)
1	$(0.0859, 0.9609, 1.3009)$	0.4859	0.02
2	$(0.2923, 1.1673, 0.3723)$	0.2407	0.04
4	$(0.4787, 0.6702, 0.5587)$	0.2282	0.11
5	$(0.5668, 0.7583, 0.1623)$	0.1947	0.15
27	$(0.5604, 0.6496, 0.3444)$	0.1919	4.86
33	$(0.5796, 0.6689, 0.25784)$	0.1916	7.64

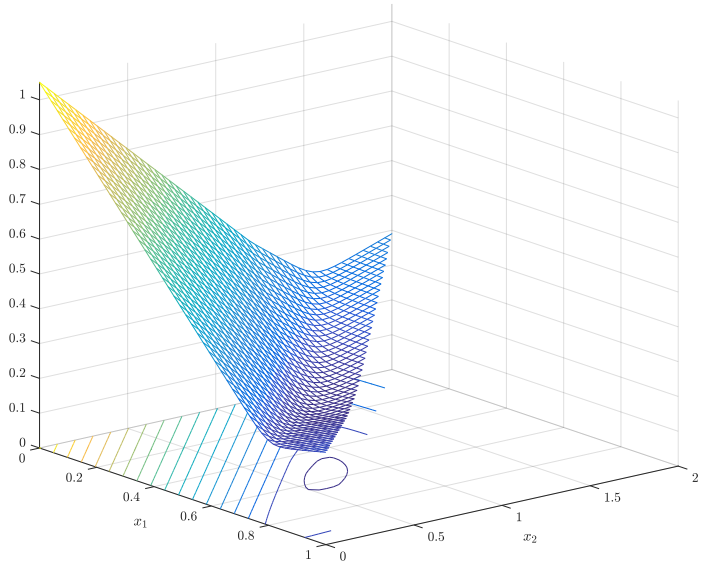


Figure 2: Graph of the objective function f of Problem (15) over the simplex $S_a = \{(x_1, x_2, x_3) : x_1 + \frac{1}{2}x_2 + \frac{1}{3}x_3 = 1,\ x_1, x_2, x_3 \geq 0\}$.

Now, we present an example that the global optimal solution is a boundary point of the feasible set of the problem.

Example 5. Consider

$$\begin{aligned} &\min\ f(x_1, x_2, x_3) \\ &\text{s.t. } x_1 + x_2 + x_3 = 1, \\ &\quad x_1, x_2, x_3 \geq 0, \end{aligned} \tag{16}$$

where

$$\begin{aligned} f(x_1, x_2, x_3) = {} & 0.1 \max\{0.2x_1 + 0.3x_2 + 0.5x_3, 0.1x_1 + 0.7x_2 + 0.1x_3\} \\ & + 0.4 \min\{0.2x_1 + 0.3x_2 + 0.5x_3, 0.1x_1 + 0.7x_2 + 0.1x_3\} \\ & + \frac{1}{60} \ln(e^{9x_1} + e^{5x_2} + e^{12x_3}). \end{aligned}$$

The graph of the objective function f over the simplex S_a is shown in Figure 3. The optimal solution and the optimal value of Problem (16) are $x^* = (0.6025, 0, 0.3975)$ and $f^* = 0.1693$, respectively (note that x^* is a boundary point of the simplex S_a).

Table 3: Obtained results by the cutting angle method for Example 5

Iter	Optimal point found $\hat{x}^k$	Optimal value found $f(\hat{x}^k)$	Time (sec)
1	$(0.3812, 0.3176, 0.3012)$	0.2199	0.03
3	$(0.6923, 0, 0.3077)$	0.1744	0.18
22	$(0.6136, 0, 0.3864)$	0.1694	3.95

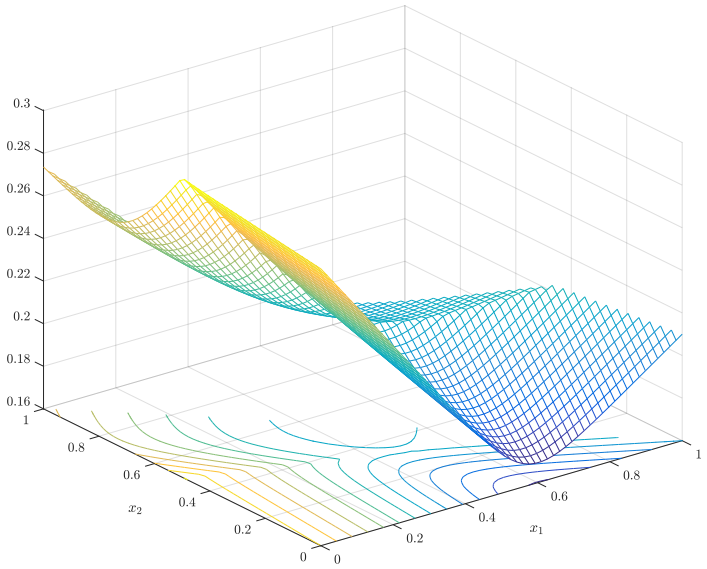


Figure 3: Graph of the objective function f of Problem (16) over the simplex $S_{\mathbf{1}} = \{(x_1, x_2, x_3) : x_1 + x_2 + x_3 = 1,\ x_1, x_2, x_3 \geq 0\}$.

5 Conclusion

This paper has proposed the cutting angle method as a means of finding the global minimum solution of a sub-topical function over a simplex. The algorithm utilized the abstract convexity of sub-topical functions to estimate the optimal value by solving mixed integer linear programming problems. The efficiency of this method has been demonstrated through numerical experiments.

Overall, the proposed cutting angle method provided a promising approach for solving sub-topical optimization problems with a simplex con-

straint. Future research can explore the extension of this algorithm to handle more complex sub-topical optimization problems with additional constraints.

Acknowledgements

The authors are very grateful to referees for their useful suggestions regarding an earlier version of this paper. The comments of the referees were very useful and they helped us to improve the paper significantly.

References

[1] Adilov, G.R., Tinaztepe, G. and Tinaztepe, R. *On the global minimization of increasing positively homogeneous functions over the unit simplex*, Int. J. Comput. Math. 87(12) (2010), 2733–2746.

[2] Andramonov, M.Y., Rubinov, A.M. and Glover, B.M. *Cutting angle methods in global optimization*, Appl. Math. Lett. 12(3) (1999), 95–100.

[3] Bagirov, A.M. and Rubinov, A.M. *Global minimization of increasing positively homogeneous functions over the unit simplex*, Ann. Oper. Res. 98 (2000), 171–187.

[4] Bakhtiari, H. and Mohebi, H. *Characterizing sub-topical functions*, Wavelet and Linear Algebra, 4(2) (2017), 13–23.

[5] Doagooei, A.R. *Sub-topical functions and plus-co-radiant sets*, Optimization 65(1) (2016), 107–119.

[6] Ferrer, A., Bagirov A. and Beliakov, G. *Solving DC programs using the cutting angle method*, J. Global. Optim. 61 (2015), 71–89.

[7] Gunawardena, J. *An introduction to idempotency*, Cambridge University Press, Cambridge, 1998.

[8] Mohebi, H. and Samet, M. *Abstract convexity of topical functions*, J. Global. Optim. 58(2) (2014), 365–375.

[9] Rubinov, A.M. *Abstract convexity and global optimization*, Kluwer Academic Publishers, Dordrecht, 2000.

[10] Rubinov, A.M. and Singer, I. *Topical and sub-topical functions, downward sets and abstract convexity*, Optimization 50 (2001), 307–351.

[11] Rudin, W. *Principles of mathematical analysis*, McGraw-Hill, New York, 1976.

[12] Singer, I. *Abstract convex analysis*, Wiley-Interscience, New York, 1997.

Iranian Journal of Numerical Analysis and Optimization
Vol. 14, No. 1, 2024, pp 219–264
https://doi.org/10.22067/ijnao.2023.84272.1311
https://ijnao.um.ac.ir/

CC BY NC

Research Article

Richardson extrapolation technique on a modified graded mesh for singularly perturbed parabolic convection-diffusion problems

K. K. Sah*, and S. Gowrisankar

Abstract

In this paper, we focus on investigating a post-processing technique designed for one-dimensional singularly perturbed parabolic convection-diffusion problems that demonstrate a regular boundary layer. We use a backward Euler numerical approach for time derivatives with uniform mesh in the temporal direction, and a simple upwind scheme is used for spatial derivatives with modified graded mesh in the spatial direction. In this study, we demonstrate the effectiveness of the Richardson extrapolation technique in enhancing the ε-uniform accuracy of simple upwinding

*Corresponding author

Received 3 September 2023; revised 2 November 2023; accepted 8 November 2023

Kishun Kumar Sah
Department of Mathematics, National Institute of Technology Patna, Patna - 800005, India. e-mail: kishuns.phd20.ma@nitp.ac.in

S. Gowrisankar
Department of Mathematics, National Institute of Technology Patna, Patna - 800005, India. e-mail: s.gowri@nitp.ac.in

How to cite this article
Sah, K. K. and Gowrisankar, S., Richardson extrapolation technique on a modified graded mesh for singularly perturbed parabolic convection-diffusion problems. *Iran. J. Numer. Anal. Optim.*, 2024; 14(1): 219-264. https://doi.org/10.22067/ijnao.2023.84272.1311

within the discrete supremum norm, as evidenced by an improvement from $O(N^{-1}\ln(1/\varepsilon)+\triangle\theta)$ to $O(N^{-2}\ln^2(1/\varepsilon)+\triangle\theta^2)$. Furthermore, to validate the theoretical findings, computational experiments are conducted for two test examples by applying the proposed technique.

AMS subject classifications (2020): 65M06, 65M12, 65M15, 65M22.

Keywords: Singularly perturbed parabolic problem; Regular boundary layer; Upwind scheme; Richardson extrapolation; Modified graded mesh; Uniform convergence.

1 Introduction

We consider the following one-dimensional singularly perturbed parabolic convection-diffusion initial-boundary value problem (IBVP) posed on the domain $\aleph = \Lambda_r \times \Lambda_\theta$, where $\Lambda_r = (0,1), \Lambda_\theta = (0,T]$:

$$\begin{cases} L_\varepsilon y(r,\theta) \equiv \left(\dfrac{\partial y(r,\theta)}{\partial\theta} - \varepsilon\dfrac{\partial^2 y(r,\theta)}{\partial r^2} + \Psi_1(r)\dfrac{\partial y(r,\theta)}{\partial r} + \Psi_2(r)y(r,\theta)\right) \\ \qquad\qquad = f(r,\theta), \qquad (r,\theta)\in\aleph, \\ y(r,0) = y_0(r), \qquad r\in\overline{\Lambda}_r, \\ y(0,\theta) = 0, \qquad \theta\in\Lambda_\theta, \\ y(1,\theta) = 0, \qquad \theta\in\Lambda_\theta, \end{cases} \tag{1}$$

where $0<\varepsilon\ll 1$ is a small parameter and the coefficients Ψ_1 and Ψ_2 are sufficiently smooth function such that

$$\Psi_1(r) > 2\alpha > 0, \quad \Psi_2(r)\geq\beta\geq 0 \quad \text{on} \quad \overline{\Lambda}_r. \tag{2}$$

Under sufficient smoothness and compatibility conditions imposed on the functions y_0 and f, the parabolic IBVP (1)–(2), in general, admits a unique solution $y(r,\theta)$, which exhibits a regular boundary layer at $r=1$. This type of problem encompasses the linearized Navier–Stokes equation, which emerges in the modeling of convection-dominated flow issues in fluid dynamics, particularly when dealing with large Reynolds numbers. In the presence

of a boundary layer, conventional numerical methods, such as standard finite difference or finite element schemes, when applied on uniform meshes, fail to produce accurate numerical solutions as the singular perturbation parameter ε approaches zero. This limitation serves as a driving force to devise ε-uniformly convergent numerical methods. Among these approaches, the fitted mesh method stands out as a satisfying and widely embraced technique for surmounting numerical challenges. By employing an especially designed layer-adapted mesh, this method successfully overcomes the shortcomings encountered with traditional approaches.

Over the past few years, numerous authors have made significant strides in the development of uniformly convergent numerical methods on Shishkin meshes, specifically tailored for singularly perturbed parabolic convection-diffusion problems. These methods address the challenges posed by such complex scenarios and have shown great promise in achieving convergence, mainly Cai and Liu [1], Clavero, Jorge, and Lisbona [3], O'Riordan, Pickett, and Shishkin [28] in the presence of regular boundary layers. However, despite these advancements, it is important to note that all of these methods exhibit first-order accuracy in both the spatial and temporal variables. Hemker [9] employed a defect correction technique to enhance the accuracy of temporal variable computations for parabolic singularly perturbed problems, specifically those lacking the convective term. Additionally, in [10], the same approach was applied to parabolic convection-diffusion problems. This technique has proved to be effective in refining the precision of the results in both cases.

Woldaregay and Duressa [35] considered the study of singularly perturbed differential-difference equations having delay and advance in the reaction term. Debela, Woldaregay, and Duressa [6] studied the singularly perturbed differential equations of convection diffusion type with nonlocal boundary conditions. Mekonnen and Duressa [15] considered the study of numerical treatment of two parametric singularly perturbed parabolic convection-diffusion problems. The scheme is developed through the Crank–Nicholson discretization method in the temporal dimension followed by fitting the B-spline collocation method in the spatial direction. Also, Hailu and Duressa [8]

focused on the study of singularly perturbed parabolic convection-diffusion equations with integral boundary conditions and a large negative shift.

Furthermore, Natesan and Deb [5] have recently introduced two novel numerical schemes that exhibit uniform convergence and higher-order accuracy in time for singularly perturbed parabolic reaction-diffusion problems. These schemes are specifically designed to handle scenarios involving only parabolic layers. In their recent work, Mukherjee and Natesan [17] proposed a novel hybrid numerical scheme for singularly perturbed parabolic convection-diffusion problems. This innovative approach demonstrates uniform convergence, achieving an impressive order of convergence one in time and nearly two in space. The results of their study pave the way for more efficient and accurate solutions to such challenging problems. Several researchers, including Clavero, Gracia, and Jorge [2], Kopteva [14], and Shishkin [30], have made significant contributions to the development of uniformly convergent methods of order two in both variables for singularly perturbed parabolic convection-diffusion problems. In their study, Mukherjee and Natesan [18] employed the Richardson extrapolation technique to enhance the ε-uniform accuracy of the simple upwinding method. This improvement was observed in the discrete supremum norm and extended the accuracy from $O(N^{-1}\ln N + \triangle t)$ to $O(N^{-2}\ln^2 N + \triangle t^2)$ on the Shishkin mesh. The focus of their investigation was one-dimensional singularly perturbed parabolic convection-diffusion problems that exhibit a regular boundary layer.

Maneesh and Natesan [33] presented an advanced numerical approach designed to handle singularly perturbed systems of parabolic convection-diffusion problems, which feature overlapping exponential boundary layers. The proposed method achieves uniform convergence of higher-order accuracy. The numerical scheme integrates the backward-Euler method for the time derivative on a uniform mesh, along with the classical upwind scheme for spatial derivatives on a piecewise-uniform Shishkin mesh. This combination yields near first-order convergence in both space and time. To further enhance precision, the authors employed the Richardson extrapolation technique, elevating the accuracy of the scheme to second-order, maintaining uniform convergence in both time and space.

Negero [20, 25] considered the study of an exponentially fitted numerical scheme and analyzed it for solving singularly perturbed two-parameter parabolic problems with large temporal lag. Negero [22] studied the numerical approximation for two-parameter singularly perturbed parabolic partial differential equations with time delay. In [23], the author considered the class of time-delayed, singularly perturbed parabolic reaction-diffusion problems. Negero [21, 24, 26] studied time delay problems and parameter-uniform robust schemes for singularly perturbed parabolic convection-diffusion problems with large time-lag.

In this current study, our objective is to employ and examine a straightforward post-processing technique on the basic upwinding solution. The primary goal is to attain a convergence order greater than one concerning both the spatial and temporal variables. Richardson extrapolation, a renowned post-processing technique, offers a superior approximation to the exact solution achieved by averaging the numerical solutions computed on two embedded meshes. This method has been extensively investigated in the literature [11, 31] to enhance the accuracy of numerical solutions for singularly perturbed elliptic reaction-diffusion equations. However, the crux of their analysis relies on the direct expansion of the upwinding solution, which proves to be intricate. Shishkin and Shishkina [32] have consistently adopted this approach for quasilinear parabolic convection-diffusion problems. However, Natividad and Stynes [19] presented a comparatively simpler and distinctive analysis for applying Richardson extrapolation to one-dimensional singularly perturbed convection-diffusion boundary-value problems (BVPs). More recently, Deb and Natesan [4] conducted an analysis of the Richardson extrapolation technique for solving singularly perturbed coupled systems of convection-diffusion BVPs.

This paper represents the inaugural analysis of Richardson extrapolation on the nonuniform graded mesh applied to singularly perturbed parabolic convection-diffusion IBVPs through error decomposition after extrapolation. Initially, we address the IBVP (1)–(2) on a nonuniform graded mesh using the classical implicit upwind finite difference scheme. Subsequently, we demonstrate that the implicit upwind scheme achieves ε-uniform convergence with an almost first-order accuracy in the discrete supremum norm. Next, we

incorporate the Richardson extrapolation technique to enhance the nearly first-order convergence achieved by the simple upwinding method, leading to an almost second-order convergence. During the error analysis, we derive separate estimates for the smooth component and the layer component of the error. Consequently, we attain the necessary ε-uniform convergence result.

The subsequent sections of this paper are organized as follows: In section 2, an apriori bound on the analytical solution is presented, and stronger bounds on the derivatives of the solution are derived through decomposition. Section 3 introduces the modified graded mesh and the classical implicit upwind scheme used for discretizing the continuous problem. Here, we establish the ε-uniform convergence result for the implicit upwind scheme and provide some technical results to be utilized later in this paper. Moving on, section 4 introduces the Richardson extrapolation technique, and the main theoretical result, demonstrating that the extrapolated solution of the upwind scheme achieves ε-uniform convergence with almost second-order accuracy to the exact solution of the continuous problem, is proven. In section 5, numerical experiments are conducted on two test examples to validate the theoretical results. Finally, section 6 summarizes the main conclusions of this study.

In this paper, we use the notation C to represent a universal positive constant that remains unaffected by changes in the perturbation parameter ε, the values of N and M (representing the number of mesh intervals in the spatial and temporal directions, respectively), as well as the mesh points. In the analysis, we use the standard supremum norm $\|\cdot\|_{\infty,D}$, which is defined by

$$\|z\|_{\infty,D} = \sup_{\kappa\in D} |z(\kappa)|,$$

for a function z defined on some domain D.

Before we analyze the problem, some of the compatibility conditions are necessary. Therefore, the following compatibility conditions at the corners for functions and its first-order derivatives are assumed to satisfy,

$$\begin{cases} y_0(0) = y_0(1) = 0, \\ -\varepsilon y_0^{''}(0) + \Psi_1(0) y_0^{'}(0) = f(0,0), \\ -\varepsilon y_0^{''}(1) + \Psi_1(1) y_0^{'}(1) = f(1,0). \end{cases} \tag{3}$$

Then (1) has a unique solution in the parabolic Holder space $C^{2+\alpha,1+\alpha/2}(\overline{\aleph})$; see [29, 7]. Moreover, the fulfillment of second-order corner compatibility conditions requires the Holder space to be $C^{4+\alpha,2+\alpha/2}(\overline{\aleph})$. These conditions can be written down explicitly in terms of the data of the problem in the following way: differentiating (1) with respect to θ, we get

$$f_\theta = y_{\theta\theta} + L_\varepsilon y_\theta + \Psi_{2\theta} y = y_{\theta\theta} + L_\varepsilon(f - L_\varepsilon y) + \Psi_{2\theta} y.$$

Hence, recalling (1) and (3), the second-order corner compatibility condition is

$$L_\varepsilon(L_\varepsilon y_0) = L_\varepsilon f - f_\theta \tag{4}$$

at the corners $(0,0)$ and $(1,0)$. Under these hypotheses, the solution y of (1) has an exponential layer along the boundary $r = 1$ of $\aleph$.

2 Bounds on the solution and its derivatives

In this section, we introduce the conventional a-priori bound for the analytical solution and further establish enhanced bounds on the derivatives of the solution for problems (1)–(2). This is achieved through a decomposition of the solution into smooth and layer components. The utilization of these bounds becomes essential in the subsequent section as we aim to prove the ε-uniform error estimate. We first present some initial reports. Let $G = \overline{\aleph}/\aleph$. The differential operator L_ε then fulfills the following minimum principle on $\overline{\aleph}$, and [29, Theorem 2.2] provides the evidence for this.

Lemma 1. [Minimum principle] Assume that a function $z \in \mathbb{C}^0(\overline{\aleph}) \cap \mathbb{C}^2(\aleph)$ satisfies $z(r,\theta) \geq 0, (r,\theta) \in G$ and $L_\varepsilon z(r,\theta) \geq 0, (r,\theta) \in \aleph$. Then we have $z(r,\theta) \geq 0,$ for all $(r,\theta) \in \overline{\aleph}$.

The following is a direct result of the aforementioned minimum principle: ε-uniform bound of the solution of the problem (1)–(2).

Lemma 2. The solution y of the IBVP (1)–(2) satisfies

$$|y(r,\theta)| \leq C, \quad (r,\theta) \in \overline{\aleph}.$$

Proof. The proof follows from Lemma 2.3 of Roos, Stynes, and Tobiska [29]. □

Subsequently, we present a priori bounds concerning the derivatives of the analytical solution y to problem (1)–(2) in the forthcoming theorem, which will be utilized in Theorem 2.

Theorem 1. For all nonnegative integers p, q satisfying $0 \leq p+q \leq 5$, the exact solution y of the IBVP (1)–(2) satisfies the estimate

$$\left|\frac{\partial^{p+q} y}{\partial r^p \partial \theta^q}(r,\theta)\right| \leq C\left(1+\varepsilon^{-p}\exp(-\alpha(1-r)/\varepsilon)\right), \quad (r,\theta)\in \aleph. \tag{5}$$

Proof. In [27], this conclusion was verified for $0 \leq p+q \leq 2$,. Under necessary compatibility conditions and sufficient smoothness on the data, the proof of the estimate (5) for higher values of p, q follows similarly from [2]. □

Let us now decompose the solution y of the IBVP (1)–(2) into the form of $y = l + m$, where l and m represent the smooth component and the layer component, respectively. We further break down the smooth component into the sum

$$l = \sum_{j=0}^{4} \varepsilon^j l_j,$$

where the functions $l_j, \quad j = 0, 1, 2, 3$, are solutions of the following first-order problems:

$$\begin{cases} \dfrac{\partial l_0}{\partial \theta} + \Psi_1 \dfrac{\partial l_0}{\partial r} + \Psi_2 l_0 = f \quad \text{in} \quad \aleph, \\ l_0(0,\theta) = y(0,\theta), \quad \theta \in (0,T], \quad l_0(r,0) = y(r,0), \quad r \in \overline{\Lambda}_r, \\ \dfrac{\partial l_j}{\partial \theta} + \Psi_1 \dfrac{\partial l_j}{\partial r} + \Psi_2 l_j = \dfrac{\partial^2 l_{j-1}}{\partial r^2} \quad \text{in} \quad \aleph, \\ l_j(0,\theta) = 0, \quad \theta \in (0,T], \quad l_j(r,0) = 0, \quad r \in \overline{\Lambda}_r, \quad j = 1,2,3, \end{cases} \tag{6}$$

and the lastly, the function l_4 satisfies

$$\begin{cases} L_\varepsilon l_4 = \dfrac{\partial^2 l_3}{\partial r^2} \quad \text{in} \quad \aleph, \\ l_4(0,\theta) = l_4(1,\theta) = 0, \quad \theta \in (0,T], \quad l_4(r,0) = 0, \quad r \in \overline{\Lambda}_r. \end{cases} \tag{7}$$

Hence, the smooth component l satisfies the following IBVP:

$$\begin{cases} L_\varepsilon l = f \quad \text{in} \quad \aleph, \\ l(0,\theta) = 0, \quad l(1,\theta) = \sum_{j=0}^{4} \varepsilon^j l_j(1,\theta), \quad \theta \in (0,T], \\ l(r,0) = y(r,0), \quad r \in \overline{\Lambda}_r, \end{cases} \tag{8}$$

and therefore, the layer component m must satisfy

$$\begin{cases} L_\varepsilon m = 0 \quad \text{in} \quad \aleph, \\ m(0,\theta) = 0, \quad m(1,\theta) = y(1,\theta) - l(1,\theta), \ \theta \in (0,T], \\ m(r,0) = 0, \quad r \in \overline{\Lambda}_r. \end{cases} \tag{9}$$

Theorem 2. For any nonnegative integers p and q that fulfill the condition $0 \le p+q \le 5$, the smooth component l and the layer component m, as defined in (8) and (9), respectively, adhere to the subsequent bounds,

$$\left\| \frac{\partial^{p+q} l}{\partial r^p \partial \theta^q} \right\|_{\infty,\aleph} \le C(1+\varepsilon^{4-p}),$$

and

$$\left| \frac{\partial^{p+q} m}{\partial r^p \partial \theta^q}(r,\theta) \right| \le C\varepsilon^{-p} \exp(-\alpha(1-r)/\varepsilon), \quad (r,\theta) \in \aleph.$$

Proof. In the initial step, we will establish more stringent bounds for the smooth component l, as defined in (8), and its derivatives. The functions $l_j, j = 0,1,2,3$, serve as solutions to the problems outlined in (6). Importantly, these functions remain unaffected by the parameter ε. As a consequence, their derivatives exhibit the property of ε-uniformly bounded behavior. Hence

$$\left\| \frac{\partial^{p+q} l_j}{\partial r^p \partial \theta^q} \right\|_{\infty,\aleph} \le C, \quad \mathrm{j} = 0,1,2,3, \quad \text{for} \quad 0 \le p+q \le 5. \tag{10}$$

Similarly, the function l_4 corresponds to the solution of a problem akin to (1). By applying estimate(5) in a manner analogous to l_4, we obtain the bounds

$$\left\| \frac{\partial^{p+q} l_4}{\partial r^p \partial \theta^q} \right\|_{\infty,\aleph} \le C(1+\varepsilon^{-p} \exp(-\alpha(1-r)/\varepsilon)), \quad \text{for} \quad 0 \le p+q \le 5. \tag{11}$$

Therefore, by combining the estimates (10) and (11), for $0 \le p+q \le 5$, we can establish the necessary bounds on the smooth component l as follows:

$$\left\| \frac{\partial^{p+q} l}{\partial r^p \partial \theta^q} \right\|_{\infty,\aleph} \leq \sum_{j=0}^{4} \varepsilon^j \left\| \frac{\partial^{p+q} l_j}{\partial r^p \partial \theta^q} \right\|_{\infty,\aleph} \leq C(1+\varepsilon^{4-p}).$$

Alternatively, we can utilize the minimum principle (Lemma 1) on $\overline{\aleph}$ with the barrier function

$$\Psi(r,\theta) = C\exp(-\alpha(1-r)/\varepsilon), \quad (r,\theta) \in \aleph.$$

By selecting a sufficiently large C, we achieve the necessary bound on m. Moreover, the bounds on the derivatives of m can be deduced, following the arguments presented in [16, 28], thus concluding the proof. □

3 Numerical discretization

Within this section, a well-suited mesh is presented for discretizing the domain, ensuring the attainment of an ε-uniformly convergent difference scheme. Additionally, a detailed exposition of the employed difference scheme for discretizing the problem (1)–(2) is provided.

3.1 Temporal Discretization

To establish the convergence of (1)–(2) at each instance, we use the uniform time grid

$$\mathbb{W}_\theta^M = \{\theta_n = n\Delta\theta,\ n = 0,\ 1,\ \ldots,\ M,\ \triangle\,\theta = T/M\},$$

Here M represents the grid points.

3.2 Spatial discretization

We generate a modified graded mesh, Λ_r^N in the interval $[0,1]$ and order to resolve the boundary layer at $r = 1$ as follows:

$$\mu_j = 1 - \vartheta_{N-1} \quad \text{for} \quad j, \tag{12}$$

where μ is defined as the following form:

$$\begin{cases} \vartheta_0 = 0, & \\ \vartheta_j = 2\varepsilon\frac{j}{N}, & 1 \le j \le \frac{N}{2}, \\ \vartheta_{j+1} = \vartheta_j(1+\rho h), & \frac{N}{2} \le j \le N-2, \\ \vartheta_N = 1, & \end{cases} \tag{13}$$

where the parameter h satisfies the following nonlinear equation:

$$\ln(1/\varepsilon) = (N/2)\ln(1+\rho h). \tag{14}$$

The above section of the parameter h ensures that there are $N/2$ grid points in the interval $[0, 1-\varepsilon]$ which are distributed gradely in the interval $[0, 1-\varepsilon]$. Numerical verification stimulate that the interval $(\vartheta_{N-1}, 1)$ is not too small in comparison with the previous one $(\vartheta_{N-2}, \vartheta_{N-1})$. In the subinterval $[1-\varepsilon, 1]$, we distribute $N/2$ points with uniform step length $2\varepsilon/N$, while in the subinterval $[0, 1-\varepsilon]$, we first find h for some fix N, by means of the nonlinear equation (14), and corresponding to that h we distribute $N/2$ points in the interval $[0, 1-\varepsilon]$. The mesh length is denoted by $h_j = \vartheta_j - \vartheta_{j-1}$, for $j = 1,\ 2,\ \ldots,\ N$.

Remark 1. The mesh size in piecewise uniform and the modified graded region is given by

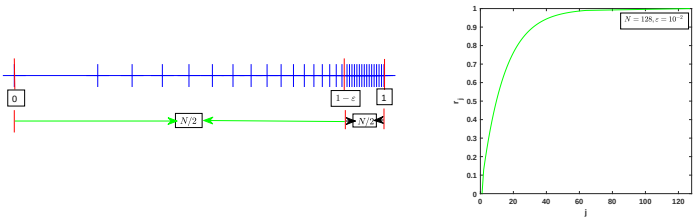
$$h_j = \begin{cases} 2\varepsilon/N & \text{for } j = 1,\ 2,\ \ldots,\ N/2, \\ \rho h \vartheta_{j-1} & \text{for } j = N/2+1,\ N/2+2,\ \ldots,\ N. \end{cases}$$

Lemma 3. The mesh defined in (13) satisfies the following estimates:

$$|h_{j+1} - h_j| \le \begin{cases} 0 & \text{for } \ j = 1,\ 2,\ \ldots,\ N/2, \\ Ch & \text{for } \ j = N/2+1,\ N/2+2,\ \ldots,\ N. \end{cases}$$

Proof. Initially, we consider $j = 1,\ 2,\ \ldots,\ N/2$. As the mesh is uniform in this portion, so there is nothing to prove.
For $j = N/2+1,\ N/2+2,\ \ldots,\ N$, we have

$$|h_{j+1} - h_j| = |\rho h \vartheta_j - \rho h \vartheta_{j-1}|$$



(a) (b)

Figure 1: Modified graded mesh layer at $r = 1$ for values of N and ε (a) $N = 32, \varepsilon = 10^{-1}$ (b) $N = 128, \varepsilon = 10^{-2}$.

$$\begin{aligned} &= \rho h|\vartheta_i - \vartheta_{j-1}| \\ &= \rho^2 h^2 \vartheta_{j-1} \\ &\leq Ch. \end{aligned}$$

Here, we have taken $0 < \rho, h < 1$. □

Lemma 4. For the modified graded mesh defined in (13), the parameter h satisfies the following bound:

$$h \leq CN^{-1} \ln(1/\varepsilon).$$

Proof. Let K_1 be the number of points ϑ_j in the partition (13) such that $\vartheta_j \leq \varepsilon$, for $j = 1, 2, \ldots, N/2$. Clearly $K_1 \leq C/h$ and K_2 are the number of points in the partition (13) such that $\vartheta_j > \varepsilon$. Let $\vartheta_{N/2+1}$ be the smallest point such that $\vartheta_j > \varepsilon$. We have to estimate the bound for K_2. Assuming $\rho h \leq 1$, we have

$$\begin{aligned} K_2 = \sum_{N/2+1}^{N} 1 &= \sum_{N/2+1}^{N} (\vartheta_{j+1} - \vartheta_j)^{-1} \int_{\vartheta_j}^{\vartheta_{j+1}} d\vartheta \\ &= \sum_{N/2+1}^{N} (h_{j+1})^{-1} \int_{\vartheta_j}^{\vartheta_{j+1}} d\vartheta \\ &= \sum_{N/2+1}^{N} (\rho h \vartheta_j)^{-1} \int_{\vartheta_j}^{\vartheta_{j+1}} d\vartheta \end{aligned}$$

$$\leq \sum_{N/2+1}^{N} (2/\rho h\vartheta_{j+1})^{-1} \int_{\vartheta_j}^{\vartheta_{j+1}} d\vartheta,$$

because $\vartheta_{j+1} < 2\vartheta_j$. For any $\vartheta \in [\vartheta_j, \vartheta_{+1}]$, we have

$$\begin{aligned} K_2 &\leq \sum_{N/2+1}^{N} 2(\rho h)^{-1} \int_{\vartheta_j}^{\vartheta_{j+1}} \frac{1}{\vartheta} d\vartheta \\ &\leq 2(\rho h)^{-1} \int_{\varepsilon}^{1} \frac{1}{\vartheta} d\vartheta \\ &\leq 2(\rho h)^{-1} \ln(1/\varepsilon). \end{aligned}$$

Recalling $N = K_1 + K_2$, we have

$$\begin{aligned} N &\leq C/\rho h + 2(\rho h)^{-1} \ln(1/\varepsilon) \\ N &\leq 1/h(C\rho + 2(\rho)^{-1} \ln(1/\varepsilon)) \\ N &\leq 1/h(C \ln(1/\varepsilon)). \end{aligned}$$

Finally, we get

$$h \leq CN^{-1} \ln(1/\varepsilon),$$

where N is the number of grid points in the r-direction. □

3.3 The classical implicit upwind scheme

Prior to explaining the scheme, we establish necessary operators for a given mesh function $z(r_j, \theta_n)$, in space and time. First, we define the forward difference operator D_r^+, the backward difference operator D_r^-, and the central difference operator D_r^0 in the spatial domain. These operators allow us to approximate derivatives with respect to the spatial variable, enabling us to capture the behavior of the function at different points. Additionally, we introduce the backward difference operator D_θ^- in the temporal domain represented by

$$\begin{cases} D_r^+ z(r_j,\theta_n) = \dfrac{z(r_{j+1},\theta_n) - z(r_j,\theta_n)}{h_{j+1}}, \\ D_r^- z(r_j,\theta_n) = \dfrac{z(r_j,\theta_n) - z(r_{j-1},\theta_n)}{h_j}, \\ D_r^0 z(r_j,\theta_n) = \dfrac{z(r_{j+1},\theta_n) - z(r_{j-1},\theta_n)}{\hat{h_j}}, \\ D_\theta^- z(r_j,\theta_n) = \dfrac{z(r_j,\theta_n) - z(r_j,\theta_{n-1})}{\triangle\theta}, \end{cases} \tag{15}$$

respectively, and define the second-order finite difference operator δ_r^2 in space by

$$\delta_r^2 z(r_j,\theta_n) = \frac{2\left(D_r^+ z(r_j,\theta_n) - D_r^- z(r_j,\theta_n)\right)}{\hat{h_j}},$$

where $\hat{h_j} = h_j + h_{j+1} \quad j = 1,\ldots,N-1$ and $\quad h_j = r_j - r_{j-1}, \quad j = 1,\ldots,N.$

Let $\aleph_\varepsilon^{N,M} = \overline{\aleph}_\varepsilon^{N,M} \bigcap \aleph$ and let $\Upsilon_\varepsilon^{N,M} = \overline{\aleph}_\varepsilon^{N,M} \setminus \aleph_\varepsilon^{N,M}$. For the discretization of the continuous problem (1), the following classical implicit upwind finite difference scheme is used on the mesh $\overline{\aleph}_\varepsilon^{N,M}$:

$$\begin{cases} L_\varepsilon^{N,M} Y^{N,\triangle\theta} \equiv (D_\theta^- - \varepsilon\delta_r^2 + \Psi_1 D_r^- + \Psi_2) Y^{N,\triangle\theta} = f \quad \text{in} \quad \aleph_\varepsilon^{N,M}, \\ Y^{N,\triangle\theta} = y, \quad \text{on} \quad \mathrm{G}_\varepsilon^{N,M}. \end{cases} \tag{16}$$

Demonstrating the discrete minimum principle of the finite difference operator $L_\varepsilon^{N,M}$ establishes its well-known property, ultimately leading to the ε-uniform stability of the difference operator $L_\varepsilon^{N,M}$.

Lemma 5. [Discrete Minimum Principle] Assume that a mesh function Usatisfies $U \geq 0$ on $G_\varepsilon^{N,M}$. Then $L_\varepsilon^{N,M} U \geq 0$ in $\aleph_\varepsilon^{N,M}$ implies that $U \geq 0$ at each point of $\overline{\aleph}_\varepsilon^{N,M}$.

Proof. Let s and l be indices such that $U_s^l = \min\limits_{(j,\theta)} U_j^\theta$, for $U_j^\theta \in \overline{\aleph}_\varepsilon^{N,M}$. Assume that $U_s^l < 0$. It is easy to see that $(s,l) \in \{1,2,\ldots,N\} \times \{1,2,\ldots,M\}$, because otherwise $U_s^l \geq 0$. It follows that $U_{s+1}^l - U_s^l > 0$ and $U_{s-1}^l - U_s^l > 0$. Thus $L_\varepsilon^{N,M} U_s^l < 0$, which is a contradiction. Therefore $U_s^l \geq 0$. The indices s and l being arbitrary, we obtain $U_j^\theta \geq 0$ in $\overline{\aleph}_j^\theta$. □

The subsequent theorem asserts the ε-uniform convergence of the numerical scheme (16) on the modified graded mesh $\overline{\aleph}_\varepsilon^{N,M}$, demonstrating nearly first-order accuracy.

Theorem 3. [Error due to up-winding] Consider the solution y to the continuous problem defined by (1)–(2), and let $Y^{N,\triangle\theta}$ represent the solution corresponding to the discrete problem outlined in (16). We can then examine the error linked to the discrete solution $Y^{N,\triangle\theta}$ at the time instance θ_n that satisfies

$$\left|y(r_j,\theta_n)-Y^{N,\triangle\theta}(r_j,\theta_n)\right| \le C\big(N^{-1}\ln(1/\varepsilon)+\triangle\theta\big), \quad \text{for } 1\le j\le N-1.$$

Proof. We obtained this result with the help of article Mukherjee and Natesan; see [18, Appendix A]. □

The primary objective of this article is to enhance the accuracy of the discrete solution $Y^{N,\triangle\theta}$ for the problem (16) through a post-processing technique. The aim is to achieve an ε-uniform order of accuracy greater than one, considering both the spatial and temporal variables for the IBVP (1)–(2). To accomplish this, we will employ the Richardson extrapolation technique. Prior to introducing this technique, we present some essential technical lemmas that will be utilized in section 4.

Lemma 6. On $\overline{\Lambda}_r^{N,\varepsilon}=\{r_j\}_0^N$, define the mesh function

$$R_j=\prod_{k=1}^{j}\left(1+\frac{\alpha h_k}{\varepsilon}\right),$$

(with the usual convention that $R_0=1$). Then for $1\le j\le N-1$,

$$L_\varepsilon^{N,M}R_j \ge \left(\frac{C_1}{\varepsilon+\alpha h_j}\right)R_j, \tag{17}$$

for some constant C_1. Moreover, for $N/2+1\le j\le N-1$, we have

$$L_\varepsilon^{N,M}R_j \ge C_2\varepsilon^{-1}R_j, \tag{18}$$

for some constant C_2.

Proof. Firstly, $R_j-R_{j-1}=\dfrac{\alpha h_j}{\varepsilon}R_{j-1}$ we have

$$\begin{aligned} L_\varepsilon^{N,M}R_j &= -\frac{2\alpha}{h_j+h_{j+1}}(R_j-R_{j-1})+\Psi_j\frac{\alpha}{\varepsilon}R_{j-1}+\Psi_jR_i \\ &\ge \frac{\alpha}{\varepsilon}R_{j-1}\left[\Psi_j-\frac{2\alpha h_j}{h_j+h_{j+1}}\right] \end{aligned}$$

$$\geq \frac{C\alpha}{\varepsilon + \alpha h_j} R_i \quad for\ 1 \leq j \leq N-1.$$

Therefore, (17) is proved and (18) is an easy consequence of it, since $h/\varepsilon < 4/\alpha$. □

Lemma 7. For the modified graded mesh $\overline{\Lambda}_r^{N,\varepsilon} = \{r_j\}_0^N$, the following inequalities hold true:

(i)
$$\exp(-\alpha(1-r_j)/\varepsilon) \leq \prod_{k=j+1}^{N} \left(1 + \frac{\alpha h_k}{\varepsilon}\right)^{-1}, \quad \text{for all } j. \tag{19}$$

(ii) There exists a constant C such that
$$\prod_{k=j+1}^{N} \left(1 + \frac{\alpha h_k}{\varepsilon}\right)^{-1} \leq CN^{-4(1-j/N)}, \quad \text{for } N/2 \leq j \leq N-1. \tag{20}$$

Proof. The proof of (i) and (ii) follows from [34]. □

4 Extrapolation of $Y^{N,\triangle\theta}$

Within this section, we introduce the extrapolation technique and concurrently present a thorough error analysis. This analysis involves the decomposition of the solution pertaining to the discrete problem outlined in (16). Ultimately, we establish an ε-uniform error estimate that is closely linked with the extrapolated solution.

4.1 Extrapolation technique

To improve the accuracy of the numerical solution $Y^{N,\triangle\theta}$ by extrapolation, we shall solve the discrete problem (16) on the fine mesh $\overline{\aleph}_\varepsilon^{2N,2M} = \overline{\Lambda}_r^{2N,\varepsilon} \times \mathbb{W}_\theta^{2M}$, where $\Lambda_r = (0,1)$ with $2N$ mesh interval in the spatial direction and $2M$ mesh interval in the θ-direction, where $\overline{\Lambda}_r^{2N,\varepsilon}$ is a Modified graded mesh having the same transition point $(1-\varepsilon)$ as $\overline{\Lambda}_r^{N,\varepsilon}$ and is obtained by bisecting each mesh interval of $\overline{\Lambda}_r^{N,\varepsilon}$. Clearly, $\overline{\aleph}_\varepsilon^{N,M} = \{(r_j, \theta_n)\} \subset \overline{\aleph}_\varepsilon^{2N,2M} = \{(\tilde{r}_j, \tilde{\theta}_n)\}$

and the following are the corresponding mesh widths in $\overline{\aleph}_{\varepsilon}^{2N,2M}$:

$$\tilde{r}_j - \tilde{r}_{j-1} = \begin{cases} \overline{H}/2, & \text{for } \tilde{r}_j \in \overline{\Lambda}^{2N,\varepsilon} \bigcap [0, 1-\varepsilon], & \text{where } \overline{H} = 2\varepsilon/N, \\ \overline{h}/2, & \text{for } \tilde{r}_j \in \overline{\Lambda}^{2N,\varepsilon} \bigcap [1-\varepsilon, 1], & \text{where } \overline{h} = \rho h \vartheta_{j-1}, \\ \tilde{\theta}_n - \tilde{\theta}_{n-1} = \triangle\theta/2 & & \text{for } \tilde{\theta}_n \in \mathbb{W}_{\theta}^{2M}. \end{cases}$$

Let $\tilde{Y}^{2N,\triangle\theta/2}$ be the solution of the discrete problem (16) on the mesh $\overline{\aleph}_{\varepsilon}^{2N,2M}$. Now, it follows from Theorem 3 that

$$\begin{aligned} Y^{N,\triangle\theta}(r_j,\theta_n) - y(r_j,\theta_n) &= C_3\bigg(N^{-1}\ln(1/\varepsilon) + \triangle\theta\bigg) + \mathbb{R}_{N,\triangle\theta}(x_j,\theta_n) \\ &= C_3\bigg(N^{-1}(\alpha\varepsilon/2\varepsilon) + \triangle\theta\bigg) + \mathbb{R}_{N,\triangle\theta}(r_j,\theta_n), \\ &\qquad (r_j,\theta_n) \in \overline{\aleph}_{\varepsilon}^{N,M}, \end{aligned} \tag{21}$$

where C_3 is some fixed constant and the remainder term $\mathbb{R}_{N,\triangle\theta}$ is $O\bigg(N^{-1}\ln(1/\varepsilon)+\triangle\theta\bigg)$. keeping in the mind that $\tilde{Y}^{2N,\triangle\theta/2}$ is obtaining using the same transition point $1-\varepsilon$, we have

$$\begin{aligned} \tilde{Y}^{2N,\triangle\theta/2}(\tilde{r}_j,\tilde{\theta}_n) - y(\tilde{r}_j,\tilde{\theta}_n) &= C_3\bigg((2N)^{-1}(\alpha\varepsilon/2\varepsilon) + (\triangle\theta/2)\bigg) + \tilde{\mathbb{R}}_{2N,\triangle\theta/2}(\tilde{r}_j,\tilde{\theta}_n), \\ &\qquad (\tilde{r}_j,\tilde{\theta}_n) \in \overline{\aleph}_{\varepsilon}^{2N,2M}, \end{aligned} \tag{22}$$

where the remainder term $\mathbb{R}_{2N,\triangle\theta/2}$ is $O\bigg(N^{-1}\ln(1/\varepsilon) + \triangle\theta\bigg)$. Now, the elimination of the $O(N^{-1})$ and $O(\triangle\theta)$ terms from (21) and (22) lead to the following approximation:

$$\begin{aligned} & y(r_j,\theta_n) - \bigg(2\tilde{Y}^{2N,\triangle\theta/2}(r_j,\theta_n) - Y^{N,\triangle\theta}(r_j,\theta_n)\bigg) \\ &= -2\tilde{\mathbb{R}}_{2N,\triangle\theta/2}(r_j,\theta_n) + \mathbb{R}_{N,\triangle\theta}(r_j,\theta_n) \\ &= O\bigg(N^{-1}\ln(1/\varepsilon) + \triangle\theta\bigg), \qquad (r_j,\theta_n) \in \overline{\aleph}_{\varepsilon}^{N,M}. \end{aligned}$$

Hence, we will employ the subsequent straightforward extrapolation formula,

$$Y_{extp}^{N,\triangle\theta}(r_j,\theta_n) = 2\tilde{Y}^{2N,\triangle\theta/2}(r_j,\theta_n) - Y^{N,\triangle\theta}(r_j,\theta_n), \quad (r_j,\theta_n) \in \overline{\aleph}_{\varepsilon}^{N,M}, \tag{23}$$

which is expected to yield an approximation of $y(r_j, \theta_n)$ is more accurate than either $Y^{N,\triangle\theta}(r_j, \theta_n)$ or $\tilde{Y}^{2N,\triangle\theta/2}(r_j, \theta_n)$.

4.2 Solution decomposition

To obtain the estimate of the nodal error $\left|y(r_j, \theta_n) - Y^{N,\triangle\theta}_{extp}(r_j, \theta_n)\right|$ after extrapolation of $Y^{N,\triangle\theta}$, we decompose the solution $Y^{N,\triangle\theta}$ on the mesh $\overline{\aleph}^{N,M}$ into the sum

$$Y^{N,\triangle\theta} = R^{N,\triangle\theta} + S^{N,\triangle\theta},$$

where the smooth component $R^{N,\triangle\theta}$ and the layer component $S^{N,\triangle\theta}$ are, respectively, the solutions of the following discrete problems:

$$\begin{cases} L^{N,M}_{\varepsilon} R^{N,\triangle\theta} = f \ \text{ in } \ \aleph^{N,M}_{\varepsilon}, R^{N,\triangle\theta} = l \ \text{ on } \ G^{N,M}_{\varepsilon}, \\ L^{N,M}_{\varepsilon} S^{N,\triangle\theta} = 0 \ \text{ in } \ \aleph^{N,M}_{\varepsilon}, S^{N,\triangle\theta} = m \ \text{ on } \ G^{N,M}_{\varepsilon}, \end{cases} \tag{24}$$

Similarly, we decompose the solution $\tilde{Y}^{2N,\triangle\theta/2}$ into the smooth component $\tilde{R}^{2N,\triangle\theta/2}$ and the layer component $\tilde{S}^{2N,\triangle\theta/2}$ on the fine mesh $\overline{\aleph}^{2N,2M}$ as

$$\tilde{Y}^{2N,\triangle\theta/2} = \tilde{R}^{2N,\triangle\theta/2} + \tilde{S}^{2N,\triangle\theta/2}.$$

Then, the errors can be written in the following form:

$$Y^{N,\triangle\theta} - y = \left(R^{N,\triangle\theta} - l\right) + \left(S^{N,\triangle\theta} - m\right).$$

4.3 Extrapolation of $R^{N,\triangle\theta}$

Let

$$\eta_1(r,\theta) = \frac{1}{2}\Psi_1(r)\frac{\partial^2 l}{\partial r^2(r,\theta)} \ \text{ and } \ \eta_2(r,\theta) = \frac{1}{2}\frac{\partial^2 l}{\partial r^2(r,\theta)}, \quad (r,\theta) \in \aleph.$$

Lemma 8. Assume that $\varepsilon \leq N^{-1}$. Then, the local truncation error related to the smooth component fulfills the following condition:

$$L_{\varepsilon}^{N,M}\left(R^{N,\triangle\theta}-l\right)(r_j,\theta_{n+1})$$
$$=h_j\eta_1(r_j,\theta_{n+1})+\triangle\theta\eta_2(r_j,\theta_{n+1})+O\left(\overline{H}^2+\triangle\theta^2\right),\quad \text{for } 1\le j\le N-1.$$

Proof. By utilizing the bounds on the derivatives of l as provided in Theorem 2 and $\varepsilon \le N^{-1} \le \overline{H}$, we can readily derive the following outcome from Taylor's expansion:

$$L_{\varepsilon}^{N,M}\left(R^{N,\triangle\theta}-l\right)(r_j,\theta_{n+1})$$
$$=\frac{\varepsilon}{3\hat{h}_j}\left[h_{j+1}^2\frac{\partial^3 l}{\partial r^3}(\omega_1,\theta_{n+1})-h_j^2\frac{\partial^3 l}{\partial r^3}(\omega_2,\theta_{n+1})\right]+\frac{h_j}{2}\Psi_1(r_j)\frac{\partial^2 l}{\partial r^2}(r_j,\theta_{n+1})$$
$$-\frac{h_j^2}{3!}\Psi_1(r_j)\frac{\partial^3 l}{\partial r^3}(\omega_3,\theta_{n+1})-\frac{\triangle\theta}{2}\frac{\partial^2 l}{\partial\theta^2}(r_j,\theta_{n+1})-\frac{\triangle\theta}{3!}\frac{\partial^3 l}{\partial\theta^3}(r_j,\tilde{\theta}),$$

for some $\omega_1\in(r_j,r_{j+1}),\omega_2,\omega_3\in(r_{j-1},r_j)$ and $\tilde{\theta}\in(\theta_n,\theta_{n+1})$. □

Now, following the classical approach of Keller [12], we define the functions Q_i, where $i=1,2$, as the solutions of the following IBVPs:

$$\begin{cases} L_{\varepsilon}Q_i=\eta_i \ \text{ in } \ \aleph,\\ Q_i(r,0)=0,\quad r\in\overline{\Lambda}_r,\\ Q_i(0,\theta)=Q_i(1,\theta)=0,\quad \theta\in(0,T],\quad i=1,2.\end{cases}\tag{25}$$

Next, decompose Q_i as $Q_i=\Theta_i+\Phi_i,\ \ i=1,2$, where the smooth component Θ_i and the layer component Φ_i satisfy the following IBVPs:

$$\begin{cases} L_{\varepsilon}\Theta_i=\eta_i,\quad L_{\varepsilon}\Phi_i=0 \ \text{ in } \ \aleph,\\ \Theta_i(r,0)=\Phi_i(r,0)=0,\quad \mathrm{r}\in\overline{\Lambda}_r,\\ \Theta_i(0,\theta)=\Phi_i(0,\theta)=0,\quad \Theta_i(1,\theta)=-\Phi_i(1,\theta)\ \ \theta\in(0,T],\quad i=1,2.\end{cases}\tag{26}$$

Theorem 4. For all nonnegative integers p,q satisfying $0\le p+q\le 2$, the smooth components $\Theta_i, i=1,2$, defined in (26) satisfy the following bounds:

$$\left|\left|\frac{\partial^{p+q}\Theta_i}{\partial r^p\partial\theta^q}\right|\right|_{\infty,\aleph}\le C,\quad \left|\left|\frac{\partial^3\Theta_i}{\partial r^3}\right|\right|_{\infty,\aleph}\le C\varepsilon^{-1}.$$

Proof. firstly, following the approach of [27] and using the bounds on the derivatives of l given in Theorem 2, we obtain that the solution Q_i of the IBVP (25) satisfies the estimate

$$\left|\frac{\partial^{p+q}Q_i}{\partial r^p \partial\theta^q}(r,\theta)\right| \leq C\left(1+\varepsilon^{-p}\exp(-\alpha(1-r)/\varepsilon\right), \quad (r,\theta)\in\aleph, \text{ for } 0\leq p+q\leq 2. \tag{27}$$

This shows that away from the side $r=1$, the derivatives of Q_i are bounded independent of ε; see, for example, [33]. □

Remark 2. It is to be noted that the bounds obtained in Theorem 2 on $\dfrac{\partial^{p+q}l}{\partial r^3}$, for $0\leq p+q\leq 4$, are sufficient to derive the estimate (27) for the solution Q_i of the IBVP (25).

Lemma 9. Assume that $\varepsilon\leq N^{-1}$. Then,

$$\begin{aligned}&\left(R^{N,\triangle\theta}-l\right)(r_j,\theta_{n+1})\\ &\quad= h_j\Theta_1(r_j,\theta_{n+1})+\triangle\theta\Theta_2(r_j,\theta_{n+1})+O\left(N^{-2}+\triangle\theta^2\right), \quad \text{for } 1\leq j\leq N-1.\end{aligned}$$

Proof. Assume that $1\leq j\leq N-1$. Then, applying Lemma 8 and (26), we get

$$\left(R^{N,\triangle\theta}-l\right)(r_j,\theta_{n+1}=h_jL_\varepsilon\Theta_1(r_j,\theta_{n+1})+\triangle\theta L_\varepsilon\Theta_2(r_j,\theta_{n+1})+O\left(\overline{H}+\triangle\theta^2\right). \tag{28}$$

Now, following the Taylor's expansion, it can be deduced that for $i=1,2$,

$$\begin{aligned}\left|\left(L_\varepsilon-L_\varepsilon^{N,M}\right)\Theta_i(r_i,\theta_{n+1})\right| \leq \Bigg[&\frac{\varepsilon}{3}(h_j+h_{j+1})\left\|\frac{\partial^3\Theta_i}{\partial r^3}\right\|_\infty\\ &+\frac{h_j}{2}\Psi_1(r_j)\left\|\frac{\partial^2\Theta_i}{\partial r^2}\right\|_\infty+\frac{\triangle\theta}{2}\left\|\frac{\partial^2\Theta_i}{\partial\theta^2}\right\|_\infty\Bigg].\end{aligned}$$

In accordance with Theorem 4 and the inequality $h_j\leq\overline{H}$ for all j, where $\overline{H}=2\varepsilon/N$, it follows that

$$\begin{aligned}&\left|h_j\left(L_\varepsilon-L_\varepsilon^{N,M}\right)\Theta_1(r_{j,\theta_{n+1}})+\triangle\theta\left(L_\varepsilon-L_\varepsilon^{N,M}\right)\Theta_2(r_{j,\theta_{n+1}})\right|\\ &\quad\leq C\left(\overline{H}^2+\overline{H}\triangle\theta+\triangle\theta^2\right)\leq C\left(\overline{H}^2+\triangle\theta^2\right).\end{aligned} \tag{29}$$

Hence, from the combination of (28) and (29), we get

$$L_{\varepsilon}^{N,M}\Big[R^{N,\triangle\theta}-l-h_j\Theta_1-\triangle\theta\Theta_2\Big](r_j,\theta_{n+1})=O\Big(\overline{H}^2+\triangle\theta^2\Big),\quad 1\le j\le N-1. \tag{30}$$

Provide the definitions for the following pair of discrete functions:

$$\upsilon^{\pm}(r_j,\theta_n)=C_4\Big(N^{-2}+\triangle\theta^2\Big)(1+r_j)\pm\xi(r_j,(\theta_n),\quad \text{for } 0\le j\le N \text{ and } n\ge 0,$$

where

$$\xi(r_j,\theta_n)=\begin{cases}\big[R^{N,\triangle\theta}-l-h_j\Theta_1-\triangle\theta\Theta_2\big](r_j,\theta_n), & \text{for } 0\le j\le N-1,\\ 0, & \text{for } j=0,N.\end{cases} \tag{31}$$

Now, for $1\le j\le N-1$, using the inequality $\overline{H}\le 2N^{-1}$, where $\overline{H}=2\varepsilon/N$, we have

$$L_{\varepsilon}^{N,M}\Big(\Big(N^{-2}+\triangle\theta^2\Big)(1+r_j)\Big)\ge\alpha\Big(N^{-2}+\triangle\theta^2\Big)\ge\Big(\overline{H}^2+\triangle\theta^2\Big)/4. \tag{32}$$

Therefore, by utilizing (30)and (32), one can select the constant C_4 to be suitably large, ensuring that $L_{\varepsilon}^{N,M}\upsilon^{\pm}\ge 0$ within $\aleph_{\varepsilon}^{N,M}$. Additionally, in accordance with (24), (26), and (31), it is established that $\upsilon^{\pm}\ge 0$ over the domain $G_{\varepsilon}^{N,M}$. Consequently, by employing Lemma 5 of the discrete minimum principle across the region $\aleph_{\varepsilon}^{N,M}$, we achieve

$$\left|\Big[R^{N,\triangle\theta}-l-h_j\Theta_1-\triangle\theta\Theta_2\Big](r_j,\theta_{n+1})\right|\le C\Big(N^{-2}+\triangle\theta^2\Big),\ \text{for } 1\le j\le N-1,$$

and hereby, the desired result follows. □

We deduce the estimate for the smooth part of the error after extrapolation, in the following lemma.

Lemma 10. Assume that $\varepsilon\le N^{-1}$. Then, the error after extrapolation associated to the smooth component $R^{N,\triangle\theta}$ satisfies

$$\left|l(r_j,\theta_{n+1})-R_{extp}^{N,\triangle\theta}(r_j,\theta_{n+1})\right|\le C\Big(N^{-2}+\triangle\theta^2\Big),\quad \text{for}\ \ 1\le j\le N-1.$$

Proof. Since the mesh widths of $\overline{\aleph}_{\varepsilon}^{2N,2M}$ are half of those of $\overline{\aleph}_{\varepsilon}^{N,M}$, applying Lemma 9 on the fine mesh $\overline{\aleph}_{\varepsilon}^{2N,2M}$ we have

$$\left(\tilde{R}^{2N,\triangle\theta/2}(r_j,\theta_{n+1}-l\right)(r_j,\theta_{n+1})$$
$$= \begin{cases} (\overline{H}/2)\Theta_1(r_j,\theta_{n+1}) + (\triangle\theta/2)\Theta_2(r_j,\theta_{n+1}) + O\left(N^{-2}+\triangle\theta^2\right), \\ \qquad\qquad \text{for} \;\; 1 \le j \le N/2, \\ (\overline{h}/2)\Theta_1(r_j,\theta_{n+1}) + (\triangle\theta/2)\theta_2(r_j,\theta_{n+1}) + O\left(N^{-2}+\triangle\theta^2\right), \\ \qquad\qquad \text{for} \;\; N/2+1 \le j \le N-1, \end{cases} \tag{33}$$

where $\overline{H} = 2\varepsilon/N$ and $\overline{h} = \rho h \vartheta_{j-1}$. Therefore, according to the extrapolation formula (23), from Lemma 9 and (33), it immediately follows that

$$\begin{aligned} & l(r_j,\theta_{n+1}) - R^{N,\triangle\theta}_{extp}(r_j,\theta_{n+1}) \\ &= l(r_j,\theta_{n+1}) - \left(2\tilde{R}^{2N,\triangle\theta/2}(r_j,\theta_{n+1}) - R^{N,\triangle\theta}(r_j,\theta_{n+1})\right) \\ &= -2\left(\tilde{R}^{2N,\triangle\theta/2} - l\right)(r_j,\theta_{n+1}) + \left(R^{N,\triangle\theta} - l\right)(r_j,\theta_{n+1}) \\ &= O\left(N^{-2}+\triangle\theta^2\right), \quad \text{for} \;\; 1 \le j \le N-1. \end{aligned}$$

□

4.4 Extrapolation of $S^{N,\triangle\theta}$

Lemma 11. The error after extrapolation associated to the layer component $S^{N,\triangle t}$ satisfies

$$\left| m(r_j,\theta_{n+1}) - S^{N,\triangle\theta}_{extp}(r_j,\theta_{n+1}) \right| \le CN^{-2}, \quad \text{for} \;\; 1 \le j \le N/2.$$

Proof. Let $1 \le j \le N/2$. Now, following the argument resulting to [34] over $\overline{\aleph}^{N,M}$ and using (20), we can show that

$$\left| S^{N,\triangle\theta}(r_j,\theta_{n+1}) \right| \le C \prod_{k=j+1}^{N} \left(1 + \frac{\alpha h_k}{\varepsilon}\right)^{-1} \le CN^{-2}.$$

Next, from (19) and Theorem 2 we obtain $|m(r_j,\theta_{n+1}| \le CN^{-2}$. Hence, we have

$$\left|\left(S^{N,\triangle\theta}-m\right)(r_j,\theta_{n+1})\right|\leq CN^{-2}.$$

Similarly, $\left|\left(\tilde{S}^{2N,\triangle\theta/2}-m\right)(r_j,\theta_{n+1})\right|\leq CN^{-2}$ and hereby, the required extrapolated error bound is obtained from the extrapolation formula (23). □

Let $\varrho_1(r,\theta)=\dfrac{2\varepsilon}{\alpha}\Psi_1(r)\dfrac{\partial^2 m}{\partial r^2}(r,\theta)$, and let $\varrho_2(r,\theta)=\dfrac{1}{2}\dfrac{\partial^2 m}{\partial\theta^2}(r,\theta),\ (r,\theta)\in\aleph$.

Lemma 12. For $N/2+1\leq j\leq N-1$, the local truncation error associated to the layer component satisfies

$$\begin{aligned}L_\varepsilon^{N,M}\left(S^{N,\triangle\theta}-m\right)(r_j,\theta_{n+1})=&\left(N^{-1}\ln(1/\varepsilon)\right)\varrho_1(r_j,\theta_{n+1})+\triangle\theta\varrho_2(r_j,\theta_{n+1})\\&+O\Big(\varepsilon^{-1}\exp(-\alpha(1-r_{j+1})/\varepsilon)N^{-2}\ln^2(1/\varepsilon)\\&+\exp(-\alpha(1-r_j)/\varepsilon)\triangle\theta^2\Big).\end{aligned}$$

Proof. Let $N/2+1\leq j\leq N-1$, Then, from the Taylor's expansion and using Theorem 2, for some $\omega_1\in(r_j,r_{j+1}),\omega_2,\omega_3\in(r_{j-1},r_j)$, and $\tilde{\theta}\in(\theta_n,\theta_{n+1})$, we have

$$\begin{aligned}&L_\varepsilon^{N,M}\left(S^{N,\triangle\theta}-m\right)(r_j,\theta_{n+1})\\&=\frac{\varepsilon h^2}{4!}\left[\frac{\partial^4 m}{\partial r^4}(\omega_1,\theta_{n+1})+\frac{\partial^4 m}{\partial r^4}(\omega_2,\theta_{n+1})\right]+\frac{h}{2}\Psi_1(r_j)\frac{\partial^2 m}{\partial r^2}(r_j,\theta_{n+1})\\&\quad-\frac{h^2}{3!}\Psi_1(r_j)\frac{\partial^3 m}{\partial r^3}(\omega_3,\theta_{n+1})+\frac{\triangle\theta}{2}\frac{\partial^2 m}{\partial\theta^2}(r_j,\theta_{n+1})-\frac{\triangle\theta^2}{3!}\frac{\partial^3 m}{\partial\theta^3}(r_j,\tilde{\theta})\\&=\frac{2\varepsilon}{\alpha}\left(N^{-1}\ln(1/\varepsilon)\right)\Psi_1(r_j)\frac{\partial^2 m}{\partial r^2}(r_j,\theta_{n+1})+\frac{\triangle\theta}{2}\frac{\partial^2 m}{\partial\theta^2}(x_j,t_{n+1})\\&\quad+O\Big(\varepsilon^{-1}\exp(-\alpha(1-r_{j+1})/\varepsilon)N^{-2}\ln^2(1/\varepsilon)+\exp(-\alpha(1-r_j)/\varepsilon)\triangle\theta^2\Big).\end{aligned}$$

□

Let the functions $\mathbb{Q}_i, i=1,2$, represent the solutions of the following IBVPs:

$$\begin{cases} L_\varepsilon \mathbb{Q}_i = \varrho_i \ \text{ in } \ \Omega = (1-\varepsilon, 1) \times (0, T], \\ \mathbb{Q}_i(r, 0) = 0, \quad r \in [1-\varepsilon, 1], \\ \mathbb{Q}_i(1-\varepsilon, \theta) = \mathbb{Q}_i(1, \theta) = 0, \quad \theta \in (0, T], \quad i = 1, 2. \end{cases} \tag{34}$$

The bounds on the derivatives of $\mathbb{Q}_i, i = 1, 2$, are obtained by using Lemmas 13–14. We provide a detailed proof for the function $\mathbb{Q}_1$ and show that the proof for $\mathbb{Q}_2$ follows in a similar manner.

Lemma 13. The bounds for the temporal derivatives of $\mathbb{Q}_1$ are as follows:

$$\left| \frac{\partial^q \mathbb{Q}_1}{\partial \theta^q}(r, \theta) \right| \leq C \exp(-\alpha(1-r)/\varepsilon), \quad (r, \theta) \in \Omega, \quad \text{for } \ 0 \leq q \leq 3.$$

Proof. From (34), we have $\mathbb{Q}_1 \equiv 0$ on $\Upsilon_\Omega = \overline{\Omega} \setminus \Omega$. Also, from Theorem 2, we obtain

$$|L_\varepsilon \mathbb{Q}_1(r, \theta)| = |\varrho_1| \leq C\varepsilon^{-1} \exp(-\alpha(1-r)/\varepsilon), \quad (r, \theta) \in \Omega.$$

Now, consider the barrier function

$$\psi(r, \theta) = C \exp(-\alpha(1-r)/\varepsilon), \quad (r, \theta) \in \Omega, \tag{35}$$

where C is sufficiently large. It follows that $L_\varepsilon \psi(r, \theta) \geq |L_\varepsilon \mathbb{Q}_1(r, \theta)|$ in Ω and $\psi \geq \mathbb{Q}_1$ on Υ_Ω. Since L_ε satisfies the minimum principle Lemma 1 on $\overline{\Omega}$, there we have $\mathbb{Q}_1 \leq C \exp(-\alpha(1-r)/\varepsilon)$, in Ω.

Next, we shall obtain the bound on $\dfrac{\partial \mathbb{Q}_1}{\partial \theta}$. On the sides $r = 1-\varepsilon$ and $r = 1$ of $\overline{\Omega}$, we have $\mathbb{Q}_1 \equiv 0$ and so $\dfrac{\partial \mathbb{Q}_1}{\partial \theta} \equiv 0$. On the side $\theta = 0$, we have $\mathbb{Q}_1 \equiv 0$, and so $\dfrac{\partial \mathbb{Q}_1}{\partial r} \equiv \dfrac{\partial^2 \mathbb{Q}_1}{\partial r^2} \equiv 0$. Hence, from (34), we obtain

$$\frac{\partial \mathbb{Q}_1}{\partial \theta}(r, 0) = \frac{2\varepsilon}{\alpha} \Psi_1(r) \frac{\partial^2 m}{\partial r^2}(r, 0), \quad r \in [1-\varepsilon, 1]. \tag{36}$$

Since on the side $\theta = 0$, we have $m \equiv 0$, which implies that $\dfrac{\partial^2 m}{\partial r^2} \equiv 0$, and hence from (36), we obtain $\dfrac{\partial \mathbb{Q}_1}{\partial \theta}(r, 0) = 0$, $\mathrm{r} \in [1-\varepsilon, 1]$. Therefore, it is clear that

$$\frac{\partial \mathbb{Q}_1}{\partial \theta}(r, 0) \equiv 0, \quad \text{on } \ \Upsilon_\Omega.$$

Now, differentiating the equation given in (34) with respect to θ, we get

$$\frac{\partial^2 \mathbb{Q}_1}{\partial \theta^2} - \varepsilon \frac{\partial^3 \mathbb{Q}_1}{\partial \theta \partial r^2} + \Psi_1(r) \frac{\partial^2 \mathbb{Q}_1}{\partial \theta \partial r} + \Psi_2(r) \frac{\partial \mathbb{Q}_1}{\partial \theta} = \frac{2\varepsilon}{\alpha} \Psi_1(r) \frac{\partial^3 m}{\partial \theta \partial r^2} \quad \text{in} \quad \Omega, \quad (37)$$

and employing Theorem 2 in (37) yields that

$$\left| L_\varepsilon \left(\frac{\partial \mathbb{Q}_1}{\partial \theta} \right)(r, \theta) \right| \leq C \varepsilon^{-1} \exp(-\alpha(1-r)/\varepsilon), \quad (r, \theta) \in \Omega.$$

Therefore, choosing the barrier function $\psi(r, \theta)$ given in (20), we obtain that

$$L_\varepsilon \psi(r, \theta) \geq \left| L_\varepsilon \left(\frac{\partial \mathbb{Q}_1}{\partial \theta} \right)(r, \theta) \right| \text{ in } \Omega \text{ and } \psi \geq \left| \frac{\partial \mathbb{Q}_1}{\partial \theta} \right| \text{ on } \Upsilon_\Omega.$$

Thus, applying the minimum principle on $\overline{\Omega}$, we have

$$\left| \frac{\partial \mathbb{Q}_1}{\partial \theta} \right| \leq C \exp(-\alpha(1-r)/\varepsilon) \quad \text{in} \quad \Omega.$$

Similarly, we shall obtain the bound on $\dfrac{\partial^2 \mathbb{Q}_1}{\partial \theta^2}$. On the side $r = 1 - \varepsilon$ and $r = 0$ of $\overline{\Omega}$ Hence,

$$\left| \frac{\partial^2 \mathbb{Q}_1}{\partial \theta^2}(r, \theta) \right| \leq C \exp(-\alpha(1-r)/\varepsilon), \quad (r, \theta) \in \Omega.$$

Finally, by employing a similar approach as for the second-order derivative $\dfrac{\partial^2 \mathbb{Q}_1}{\partial \theta^2}$, one can derive the bound for $\dfrac{\partial^3 \mathbb{Q}_1}{\partial \theta^3}$, thereby completing the proof. □

Lemma 14. The following mixed derivative of $\mathbb{Q}_1$, bound on the derivative of $\dfrac{\partial \mathbb{Q}_1}{\partial \theta}$ and spatial derivative of $\mathbb{Q}_1$ satisfies the following bound:

$$\begin{cases} (A). \quad \left| \dfrac{\partial^3 \mathbb{Q}_1}{\partial r \theta^2}(r, \theta) \right| \leq C \varepsilon^{-1} \exp(-\alpha(1-r)/\varepsilon), \quad (r, \theta) \in \Omega \\ (B). \quad \left| \dfrac{\partial^p}{\partial r^p} \left(\dfrac{\partial \mathbb{Q}_1}{\partial \theta} \right)(r, \theta) \right| \leq C \varepsilon^{-p} \exp(-\alpha(1-r)/\varepsilon), \quad (r, \theta) \in \Omega \text{ for p} = 1, 2. \\ (C). \quad \left| \dfrac{\partial^p \mathbb{Q}_1}{\partial r^p}(r, \theta) \right| \leq C \varepsilon^{-p} \exp(-\alpha(1-r)/\varepsilon), \quad (r, \theta) \in \Omega \text{ for } 1 \leq p \leq 3. \end{cases}$$

Proof. For (A) from Lemma 13, we have

$$-\varepsilon \frac{\partial}{\partial r} \left(\frac{\partial^3 \mathbb{Q}_1}{\partial r \partial \theta^2} \right) + \Psi_1(r) \frac{\partial^3 \mathbb{Q}_1}{\partial r \partial \theta^2} + \Psi_2(r) \frac{\partial^2 \mathbb{Q}_1}{\partial r^2} = J_1 \quad \in \Omega, \qquad (38)$$

where $J_1(r,\theta) = -\varepsilon\dfrac{\partial^3 \mathbb{Q}_1}{\partial\theta^3} + \dfrac{2\varepsilon}{\alpha}\Psi_1(r)\dfrac{\partial^4 m}{\partial\theta^2\partial r^2}$. From Lemma 13 and Theorem 2, we have

$$|J_1(r,\theta)| \le C\varepsilon^{-1}\exp(-\alpha(1-r)/\varepsilon), \quad (r,\theta)\in\Omega. \tag{39}$$

Let $\theta \in [0,T]$. be fixed. To obtain the desired bound on $\dfrac{\partial^3 \mathbb{Q}_1}{\partial r\partial\theta^2}$, we apply the argument presented by Kellogg and Tsan [13] on the line segment $\{(r,\theta), r \in [1-\varepsilon, 1]\}$. This requires to use (38), inequality (39), and the previously established bound on $\dfrac{\partial^2 \mathbb{Q}_1}{\partial\theta^2}$.

For (B) from Lemma 13, rewriting (37), we get the following form:

$$-\varepsilon\frac{\partial}{\partial r}\left(\frac{\partial^2 \mathbb{Q}_1}{\partial r\partial\theta}\right) + \Psi_1(r)\frac{\partial^2 \mathbb{Q}_1}{\partial r\partial\theta} + \Psi_2(r)\frac{\partial \mathbb{Q}_1}{\partial} = J_2 \quad \in \Omega, \tag{40}$$

$J_2(r,\theta) = -\varepsilon\dfrac{\partial^2 \mathbb{Q}_1}{\partial\theta^2} + \dfrac{2\varepsilon}{\alpha}\Psi_1(r)\dfrac{\partial^3 m}{\partial\theta\partial r^2}$. From Lemmas 13 and 14 (A) and Theorem 2, then we get

$$\left|\frac{\partial^p J_2}{\partial r^p}(r,\theta)\right| \le C\varepsilon^{-1}\exp(-\alpha(1-r)/\varepsilon), \quad (r,\theta)\in\Omega. \tag{41}$$

Now, let us consider the fixed time $\theta \in [0,T]$. Applying the argument presented by Kellogg and Tsan [13] to the line segment $\{(r,\theta), r \in [1-\varepsilon, 1]\}$, using (40) inequality (40) and the bound on $\dfrac{\partial \mathbb{Q}_1}{\partial\theta}$, we can obtain the desired bound for $\dfrac{\partial^2 \mathbb{Q}_1}{\partial r\partial\theta}$. Similarly, differentiating (40) with respect to r and use of the inequality (41), the obtained bound on $\dfrac{\partial \mathbb{Q}_1}{\partial\theta}$, $\dfrac{\partial^2 \mathbb{Q}_1}{\partial r\partial\theta}$ directly implies the desired bound on $\dfrac{\partial^3 \mathbb{Q}_1}{\partial r^2\partial\theta}$.

For (C) from Lemma 12, rewriting (34) we get the following form:

$$-\varepsilon\frac{\partial}{\partial r}\left(\frac{\partial \mathbb{Q}_1}{\partial r}\right) + \Psi_1(r)\frac{\partial \mathbb{Q}_1}{\partial r} + \Psi_2\mathbb{Q}_1 = J_3 \in \Omega, \tag{42}$$

where $J_3(r,\theta) = -\dfrac{\partial \mathbb{Q}_1}{\partial\theta} + \dfrac{2\varepsilon}{\alpha}\Psi_1\dfrac{\partial^2 m}{\partial r^2}$. From Lemmas 13 and 14 (B), we have

$$\left|\frac{\partial^p J_3}{\partial r^p}(r,\theta)\right| \le C\varepsilon^{-1}\exp(-\alpha(1-r)/\varepsilon), \quad (r,\theta)\in\Omega \ \text{ for } \ 0\le p\le 2. \tag{43}$$

Let us now fix $\theta \in [0,T]$ and proceed to apply the argument presented by Kellogg and Tsan [13] to the line segment $\{(r,\theta), r \in [1-\varepsilon, 1]\}$. By utilizing (42), inequality (43), and the given bound on $\mathbb{Q}_1$, we can derive the desired bound for $\dfrac{\partial \mathbb{Q}_1}{\partial r}$.

The bound on $\dfrac{\partial^2 \mathbb{Q}_1}{\partial r^2}$ can be deduced in a similar manner as in the previous cases. By employing (43), along with the bounds on $\mathbb{Q}_1$ and $\dfrac{\partial \mathbb{Q}_1}{\partial r}$, we can establish the desired bound for $\dfrac{\partial^2 \mathbb{Q}_1}{\partial r^2}$. Lastly, to obtain the bound on $\dfrac{\partial^3 \mathbb{Q}_1}{\partial r^3}$, one can differentiate with respect to r and employ (42) in a similar fashion. □

Subsequently, we consolidate the preceding lemmas into the following theorem.

Theorem 5. For all nonnegative integers p, q, satisfying $0 \le p+q \le 3$, the solution $\mathbb{Q}_i, i = 1,2$, of (34) satisfies the following bounds:

$$\left| \frac{\partial^{p+q} \mathbb{Q}_i}{\partial r^p \theta^q}(r,\theta) \right| \le C\varepsilon^{-p} \exp(-\alpha(1-r)/\varepsilon), \quad (r,\theta) \in \Omega.$$

Lemma 15. For $N/2+1 \le j \le N-1$, we obtain

$$\Big(S^{N,\triangle\theta} - m\Big)(r_j,\theta_{n+1}) = \Big(N^{-1}\ln(1/\varepsilon)\Big)\mathbb{Q}_1(r_j,\theta_{n+1}) + \triangle\theta \mathbb{Q}_2(r_j,\theta_{n+1}) + O\Big(N^{-2}\ln^2(1/\varepsilon) + \triangle\theta^2\Big).$$

Proof. Let $N/2+1 \le j \le N-1$.
Then, Lemma 12 and (34) imply that

$$\begin{cases} L_\varepsilon^{N,M}\Big(S^{N,\triangle\theta} - m\Big)(r_j,\theta_{n+1}) \\ = \Big(N^{-1}\ln(1/\varepsilon)\Big) L_\varepsilon \mathbb{Q}_1(r_j,\theta_{n+1}) + \triangle\theta L_\varepsilon \mathbb{Q}_2(r_j,\theta_{n+1}) \\ \quad + O\Big(\varepsilon^{-1}\exp(-\alpha(1-r)/\varepsilon) N^{-2}\ln^2(1/\varepsilon) + \exp(-\alpha(1-r_j)/\varepsilon) \triangle \theta^2\Big). \end{cases} \tag{44}$$

Again, from the Taylor's expansion and Theorem 5 it follows that for $l = 1,2$,

$$\Big(L_\varepsilon - L_\varepsilon^{N,M}\Big)\mathbb{Q}_i(r_j,\theta_{n+1})$$

$$= O\Big(\varepsilon^{-1}\exp(-\alpha(1-r_{j+1})/\varepsilon)N^{-1}\ln(1/\varepsilon) + \exp(-\alpha(1-r_j)/\varepsilon)\triangle\theta\Big),$$

and so

$$\begin{aligned}
&\left|\Big(N^{-1}\ln(1/\varepsilon)\Big)\Big(L_\varepsilon - L_\varepsilon^{N,M}\Big)\mathbb{Q}_l(r_j,\theta_{n+1}) + \triangle\theta\Big(L_\varepsilon - L_\varepsilon^{N,M}\Big)\mathbb{Q}_2(r_j,\theta_{n+1})\right| \\
&\quad\le C\Big[\varepsilon^{-1}\exp(-\alpha(1-r_{j+1})/\varepsilon)N^{-2}\ln^{-2}(1/\varepsilon) + \exp(-\alpha(1-r_j)/\varepsilon)\triangle\theta^2 \\
&\qquad + \Big(\varepsilon^{-1}\exp(-\alpha(1-r_{j+1})/\varepsilon) + \exp(-\alpha(1-r_j)/\varepsilon)\Big)\Big(N^{-1}\ln(1/\varepsilon)\Big)\triangle\theta\Big] \\
&\quad\le C\varepsilon^{-1}\exp(-\alpha(1-r_{j+1})/\varepsilon)\Big(N^{-2}\ln^2(1/\varepsilon) + \triangle\theta^2\Big).
\end{aligned} \tag{45}$$

Therefore, combining (44) and (45), we have

$$\begin{aligned}
&\left|L_\varepsilon^{N,M}\Big[S^{N,\triangle\theta} - m - \Big(N^{-1}\ln(1/\varepsilon)\Big)\mathbb{Q}_1 - \triangle\theta\mathbb{Q}_2\Big](r_i,\theta_{n+1})\right| \\
&\qquad\le C_5\Big(\varepsilon^{-1}\exp(-\alpha(1-r_{j+1})/\varepsilon)\Big(N^{-2}\ln^2(1/\varepsilon) + \triangle\theta^2\Big)\Big), \\
&\qquad\qquad \text{for}\quad N/2+1 \le j \le N-1,
\end{aligned} \tag{46}$$

for some constant C_5. Now, consider the barrier function

$$\chi_j^n = C_6\Big[N^{-2}(1+r_j) + \Big(N^{-2}\ln^2(1/\varepsilon) + \triangle\theta^2\Big)R_j\prod_{k=1}^{N}\Big(1+\frac{\alpha h_k}{\varepsilon}\Big)^{-1}\Big], \quad \text{for } N/2 \le j \le N.$$

For $N/2+1 \le j \le N-1$, employing (18), we get

$$L_\varepsilon^{N,M}\chi_j^{n+1} \ge C_2C_6\varepsilon^{-1}\Big(N^{-2}\ln^2(1/\varepsilon) + \triangle\theta^2\Big)\prod_{k=j+1}^{N}\Big(1+\frac{\alpha h_k}{\varepsilon}\Big)^{-1}, \tag{47}$$

Also, for $j \ge N/2$, from (19), we get

$$\begin{aligned}
\exp(-\alpha(1-r_{j+1})/\varepsilon) &\le \exp(\lambda h/\varepsilon)\prod_{k=j+1}^{N}\Big(1+\frac{\alpha h_k}{\varepsilon}\Big)^{-1} \\
&\le \exp(\lambda_1)\prod_{k=j+1}^{N}\Big(1+\frac{\alpha h_k}{\varepsilon}\Big)^{-1}.
\end{aligned} \tag{48}$$

Hence, for $N/2+1 \leq j \leq N-1$, it is easy to show from (46) – (48) that

$$L_{\varepsilon}^{N,M}\chi_j^{n+1} \geq \left| L_{\varepsilon}^{N,M}\left[S^{N,\triangle\theta} - m - \left(N^{-1}\ln(1/\varepsilon)\right)\mathbb{Q}_1 - \triangle\theta\mathbb{Q}_2\right](r_j,\theta_{n+1})\right|, \tag{49}$$

provided $C_6 \geq \exp(\lambda_1)C_5/C_2$. Again, from the proof of Lemma 11, we get

$$\left|\left(S^{n,\triangle\theta} - m\right)(1-\varepsilon,\theta_{n+1})\right| \leq C_7 N^{-2},$$

for some constant C_7. Hence, by (34),

$$\chi_{N/2}^{n+1} \geq C_6 N^{-2} \geq \left|\left[S^{N,\triangle\theta} - m - \left(N^{-1}\ln(1/\varepsilon)\right)\mathbb{Q}_1 - \triangle\theta\mathbb{Q}_2\right](1-\varepsilon,\theta_{n+1})\right|, \tag{50}$$

provided $C_6 \geq C_7$. Also, recalling (24) and (34), we have

$$\chi_N^{n+1} \geq 0 = \left|\left[S^{N,\triangle\theta} - m - \left(N^{-1}\ln(1/\varepsilon)\right)\mathbb{Q}_1 - \triangle\theta\mathbb{Q}_2\right](1,\theta_{n+1})\right|. \tag{51}$$

Therefore, it follows from (49)–(51) that one can choose

$$C_6 = \max\{\exp(\lambda_1)C_5/C_2, C_7\}$$

so that χ_j^n is a barrier function for

$$\pm\left[S^{N,\triangle\theta} - m - \left(N^{-1}\ln(1/\varepsilon)\right)\mathbb{Q}_1 - \triangle\theta\mathbb{Q}_2\right](r_j,\theta_n).$$

Thus, by the discrete minimum principle over $\overline{\aleph}_{\varepsilon}^{N,M}\bigcap\left(\left[1-\varepsilon,1\right]\times\left[0,T\right]\right)$, we have

$$\begin{aligned}
&\left|\left[S^{N,\triangle\theta} - m - \left(N^{-1}\ln(1/\varepsilon)\right)\mathbb{Q}_1 - \triangle\theta\mathbb{Q}_2\right](r_j,\theta_{n+1})\right| \\
&\quad\leq \chi_j^{n+1} \\
&\quad\leq C\left(N^{-2}\ln^2(1/\varepsilon) + \triangle\theta^2\right),
\end{aligned}$$

for $N/2+1 \leq j \leq N-1$, and this completes the proof. □

Lemma 16. The error after extrapolation associated to the layer component $S^{N,\triangle\theta}$ satisfies

$$\left|m(r_j,\theta_{n+1})-S^{N,\triangle\theta}_{\exp}(r_j,\theta_{n+1})\right|\leq C\bigg(N^{-2}\ln^2(1/\varepsilon)+\triangle\theta^2\bigg),$$

for $N/2+1\leq j\leq N-1$.

Proof. Let $N/2+1\leq j\leq N-1$. from Lemma 15, we have

$$\bigg(S^{N,\triangle\theta}-m\bigg)(r_j,\theta_{n+1})=N^{-1}\mathbb{Q}_1(r_j,\theta_{n+1})+\triangle\theta\mathbb{Q}_2(r_j,\theta_{n+1})+O\bigg(N^{-2}+\triangle\theta^2\bigg). \tag{52}$$

Next, since the fine mesh $\overline{\aleph}_\varepsilon^{2N,2M}$ has the same transition point $1-\varepsilon$ as that of $\overline{\aleph}_\varepsilon^{N,M}$, Lemma 15 implies that

$$\begin{aligned}&\bigg(\tilde{S}^{2N,\triangle\theta/2}-m\bigg)(r_j,\theta_{n+1})\\&\quad=(2N)^{-1}\mathbb{Q}_1(r_j,\theta_{n+1})+\triangle\theta/2\mathbb{Q}_2(r_j,\theta_{n+1})+O\bigg(N^{-2}+\triangle\theta^2\bigg).\end{aligned} \tag{53}$$

Henceforth, eliminating $O(N^{-1})$ and $O(\triangle\theta)$ terms from (52) and (53), the required extrapolated error bound is obtained. □

4.5 Convergence result of the solution $Y^{N,\triangle\theta}_{extp}$

The main result of this paper is presented in the following theorem.

Theorem 6. [Error after extrapolation] Assume that $\varepsilon\leq N^{-1}$. Let y be the solution of the continuous problem (1) – (2) and let $Y^{N,\triangle\theta}_{extp}$ be the solution obtained via the Richardson extrapolation technique by solving the discrete problem (16) on two nested meshes $\overline{\aleph}_\varepsilon^{N,M}$ and $\overline{\aleph}_\varepsilon^{2N,2M}$. Then, the error associated with the solution $Y^{N,\triangle\theta}_{extp}$ at time level θ_n satisfies

$$\left|y(r_j,\theta_n)-Y^{N,\triangle\theta}_{extp}(r_j,\theta_n)\right|\leq C\bigg(N^{-2}\ln^2(1/\varepsilon)+\triangle\theta^2\bigg),\quad\text{for}\quad 1\leq j\leq N-1. \tag{54}$$

Proof. For each $(r_j,\theta_n)\in\overline{\aleph}_\varepsilon^{N,M}$, we have

$$y(r_j,\theta_n)-Y^{N,\triangle\theta}_{extp}=\bigg(l(r_j,\theta_n)-R^{N,\triangle\theta}_{extp}\bigg)+\bigg(m(r_j,\theta_n)-S^{N,\triangle\theta}_{extp}\bigg).$$

Therefore, the result (54) follows immediately after combining Lemma 10 for the smooth component and Lemmas 11 and (16 for the layer component. □

5 Numerical results

In this section, we showcase the numerical findings acquired through the utilization of the Richardson extrapolation technique, aimed at corroborating the theoretical outcomes asserted in the preceding section. To accomplish this objective, we conduct extensive numerical experiments on a modified graded mesh $\overline{\aleph}_{\varepsilon}^{N,M}$. For all the test examples, we carefully select the transition parameter $\rho = 0.9$, along with $\alpha = 1$ and a time step size of $\triangle\theta = 1/N$. These parameters are consistently applied throughout the numerical investigations.

Example 1. Let us examine the parabolic IBVP presented below:

$$\begin{cases} \dfrac{\partial y}{\partial \theta} - \varepsilon \dfrac{\partial^2 y}{\partial r^2} + \big(1 + r(1-r)\big)\dfrac{\partial y}{\partial r} = f(r,\theta), 7(r,\theta) \in (0,1) \times (0,1], \\ y(r,0) = y_0(r), \quad 0 \le r \le 1, \\ y(0,\theta) = 0, \quad y(1,\theta) = 0, \quad 0 < \theta \le 1. \end{cases} \tag{55}$$

Example 2. Let us examine the parabolic IBVP presented below:

$$\begin{cases} \dfrac{\partial y}{\partial \theta} - \varepsilon \dfrac{\partial^2 y}{\partial r^2} + \dfrac{\partial y}{\partial r} = f(r,\theta), & (r,\theta) \in (0,1) \times (0,1], \\ y(r,0) = y_0(r), & 0 \le r \le 1, \\ y(0,\theta) = 0, \quad y(1,\theta) = 0, & 0 < \theta \le 1, \end{cases} \tag{56}$$

where the initial data $y_0(r)$ and the source function $f(r,\theta)$ have been chosen to fit the exact solution for the IBVP (55) and (56), which is given by the expression

$$y(r,\theta) = \exp(-\theta)\left(\overline{m}1 + \overline{m}2r - \exp\left(-\frac{1-r}{\varepsilon}\right)\right),$$

where the constants are defined as $\overline{m}1 = \exp\left(-\frac{1}{\varepsilon}\right)$ and $\overline{m}2 = 1 - \exp\left(-\frac{1}{\varepsilon}\right)$. Once the exact solution is known for each ε, the maximum point-wise error $\mathbb{E}_{\varepsilon}^{N,\triangle\theta}$ before and after extrapolation is calculated using

$$\max_{(r_j,\theta_n)\in\overline{\aleph}_\varepsilon^{N,M}} \left| y(r_j,\theta_n) - Y^{N,\triangle\theta}(r_j,\theta_n) \right|$$

and

$$\max_{(r_j,\theta_n)\in\overline{\aleph}_\varepsilon^{N,M}} \left| y(r_j,\theta_n) - Y_{extp}^{N,\triangle\theta}(r_j,\theta_n) \right|,$$

respectively, where $y(r_j,\theta_n), Y^{N,\triangle\theta}(r_j,\theta_n)$ and $Y_{extp}^{N,\triangle\theta}(r_j,\theta_n)$, respectively, denote the exact solution, the upwind numerical solution and the extrapolated solution obtained on the mesh $\overline{\aleph}_\varepsilon^{N,M}$ with N mesh-points in the spatial direction and M mesh-points in the θ-direction such that $\triangle\theta = T/M$ is the uniform time step. In addition, the corresponding order of convergence is determined by

$$\mathbb{P}_\varepsilon^{N,\triangle\theta} = \log_2\left(\frac{\mathbb{E}_\varepsilon^{N,\triangle\theta}}{\mathbb{E}_\varepsilon^{2N,\triangle\theta/2}}\right).$$

Now, for each N and $\triangle\theta$, we define $\mathbb{E}_\varepsilon^{N,\triangle\theta} = \max_\varepsilon \mathbb{E}_\varepsilon^{N,\triangle\theta}$ as the ε-uniform maximum point-wise error and the corresponding local ε-uniform order of convergence is defined by

$$\mathbb{P}^{N,\triangle\theta} = \log_2\left(\frac{\mathbb{E}^{N,\triangle\theta}}{\mathbb{E}^{2N,\triangle\theta/2}}\right).$$

Tables 1 and 3 present the calculated maximum point-wise errors $\mathbb{E}_\varepsilon^{N,\triangle\theta}$, as well as the corresponding order of convergence $\mathbb{P}^{N,\triangle\theta}$ for various values of ε and N before and after extrapolation, respectively, for Example 1 and (55). Similarly, Tables 2 and 4 showcase the corresponding results for Example 2 and (56). The results presented in Tables 1-4 demonstrate a consistent decrease in the computed ε-uniform errors $\mathbb{E}^{N,\triangle\theta}$ for Examples 1 and 2 as the number of grid points, N, increases. Besides, the comparison of numerical results obtained by the proposed scheme and results in [18] are tabulated in Tables 5 and 6 for Example 1. From these tables, one can conclude that the proposed scheme gives better results than the scheme considered in [18]. Figures 2–5 exhibit the solution profile for Examples 1 and 2 for various values of ε and N. Figures 2 and 3 provide a solution profile before Richardson extrapolation whereas Figures 4 and 5 provide a solution profile after Richardson extrapolation. This observation indicates that the implicit upwind scheme (16) exhibits ε-uniform convergence before and after extrapolation. Additionally, Figure 6 depicts the maximum point-wise errors for Examples 1 and 2. The

plots clearly illustrate that Richardson extrapolation enhances the order of convergence of the upwind scheme from $O\big(N^{-1}\ln(1/\varepsilon)\big)$ to $O\big(N^{-2}\ln^2(1/\varepsilon)\big)$ for $\triangle\theta = 1/N$. This finding validates the theoretical bounds established in (16) and (33).

6 Conclusion

In this article, our focus revolved around devising an efficient numerical approach for tackling one-dimensional singularly perturbed parabolic problems. Specifically, we leveraged the Richardson extrapolation technique to enhance the performance of the upwind scheme in addressing the one-dimensional singularly perturbed parabolic convection-diffusion IBVP presented in $(1)-(2)$. The methodology began by discretizing the domain using a modified graded mesh. This mesh structure was defined by $N \times M$ points. Subsequently, we applied the classical implicit upwind finite difference scheme to solve the IBVP effectively. To validate the efficacy of our approach, we furnished rigorous proof detailing the ε-uniform error estimation. This proof substantiated the convergence of the upwind scheme in an ε-uniform manner, displaying nearly first-order accuracy. Furthermore, by maintaining the transition parameter at a constant value of $\rho = 0.9$ within the modified graded mesh framework, we extended our analysis to the fine mesh configuration, which is characterized by $2N \times 2M$ mesh-points. Finally, we amalgamated the outcomes obtained from both coarse and fine mesh computations. This amalgamation served as a foundation for the implementation of the Richardson extrapolation technique. Both theoretical and computational examinations affirmed that this extrapolation technique propels the convergence of the basic upwinding scheme from almost first-order accuracy to a nearly second-order accuracy, all within the confines of the discrete supremum norm.

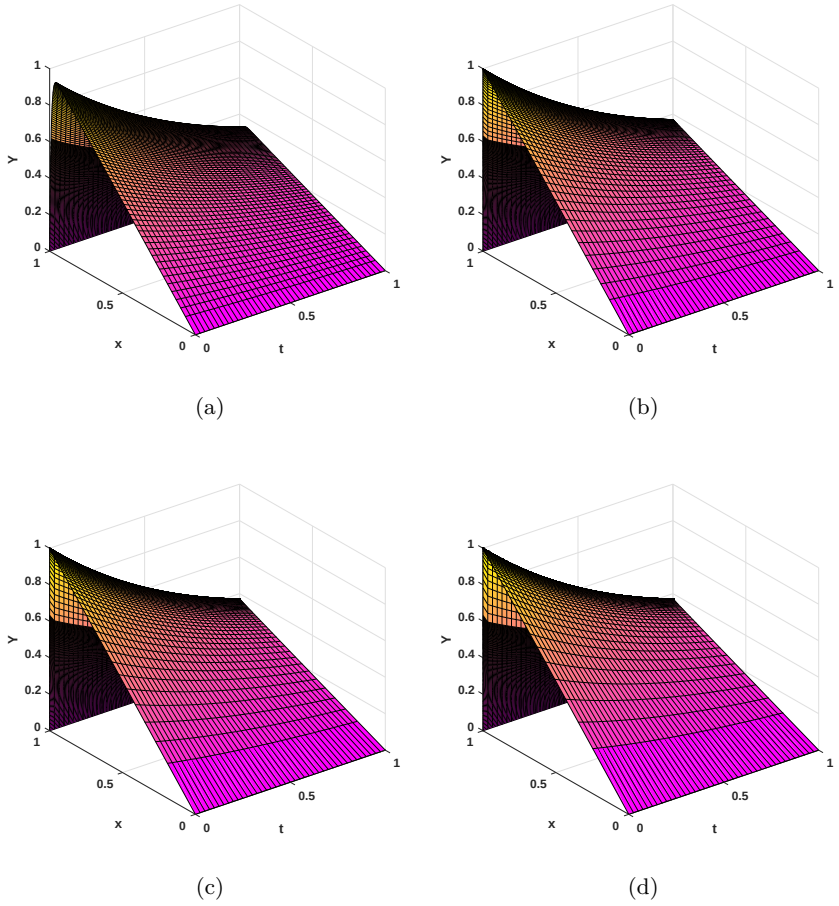


Figure 2: Solution profile for Example 1 using simple upwind scheme before Richardson extrapolation technique with different value of ε and N. (a) Solution of $N = 128, \varepsilon = 10^{-2}$ (b) Solution of $N = 128, \varepsilon = 10^{-4}$ (c) Solution of $N = 128, \varepsilon = 10^{-6}$ (d) Solution of $N = 128, \varepsilon = 10^{-8}$.

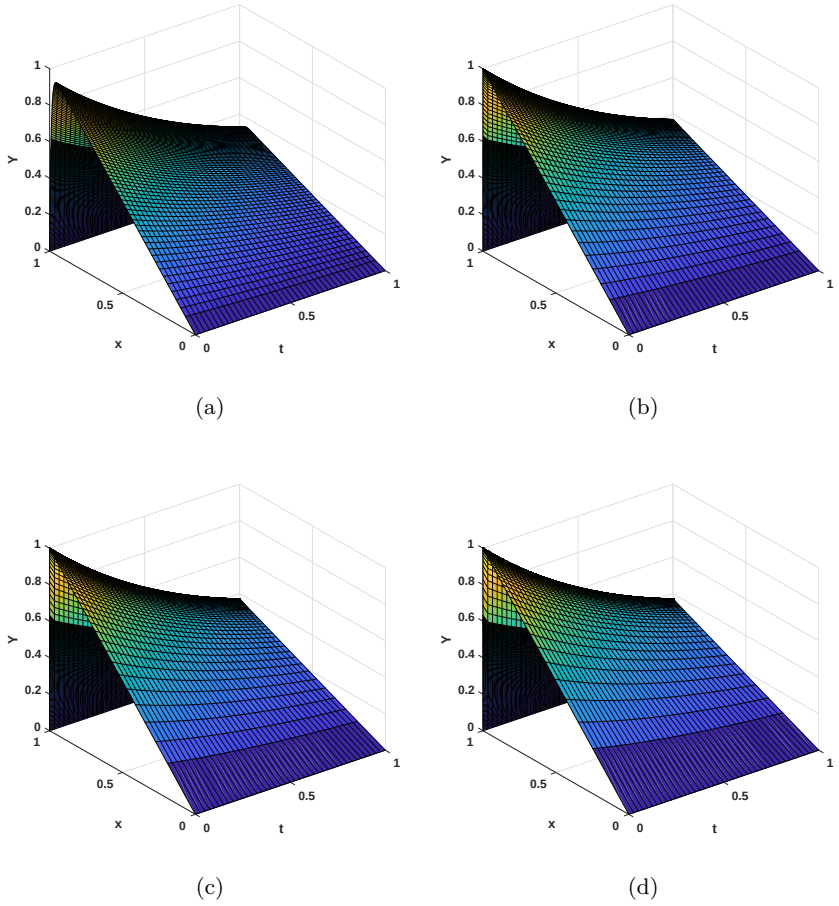
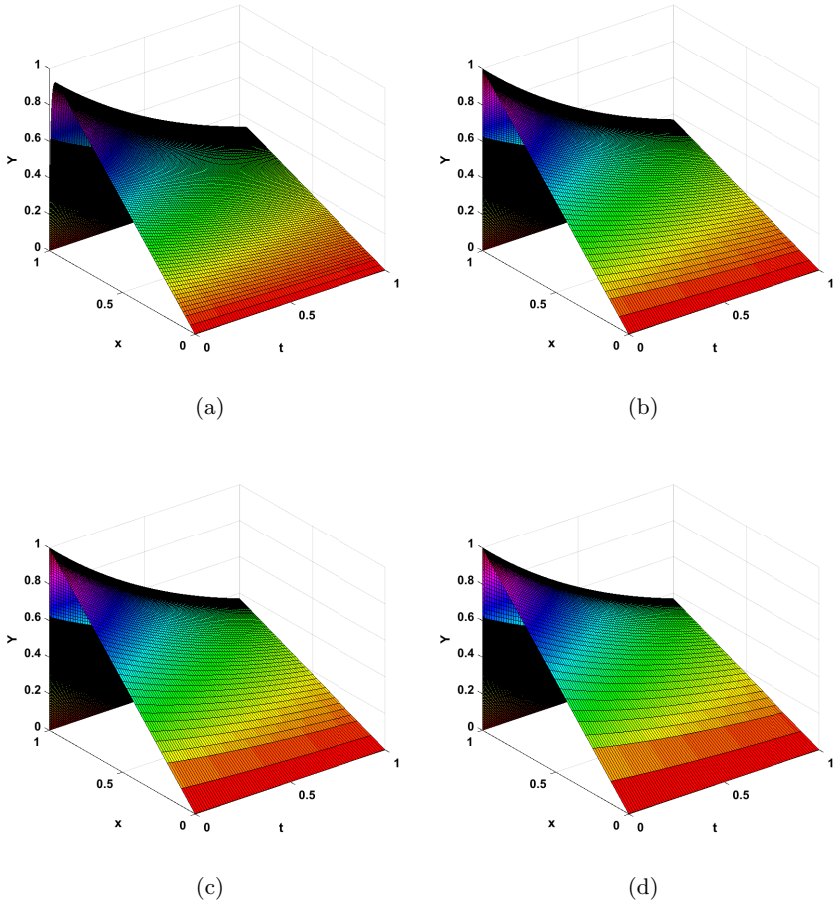


Figure 3: Solution profile for Example 2 using simple upwind scheme before Richardson extrapolation technique with different value of ε and N. (a) Solution of $N = 128, \varepsilon = 10^{-2}$ (b) Solution of $N = 128, \varepsilon = 10^{-4}$ (c) Solution of $N = 128, \varepsilon = 10^{-6}$ (d) Solution of $N = 128, \varepsilon = 10^{-8}$.



(a) (b)

(c) (d)

Figure 4: Solution profile for Example 1 using simple upwind scheme and after Richardson extrapolation technique with different value of ε and N. (a) Solution of $N = 128, \varepsilon = 10^{-2}$ (b) Solution of $N = 128, \varepsilon = 10^{-4}$ (c) Solution of $N = 128, \varepsilon = 10^{-6}$ (d) Solution of $N = 128, \varepsilon = 10^{-8}$.

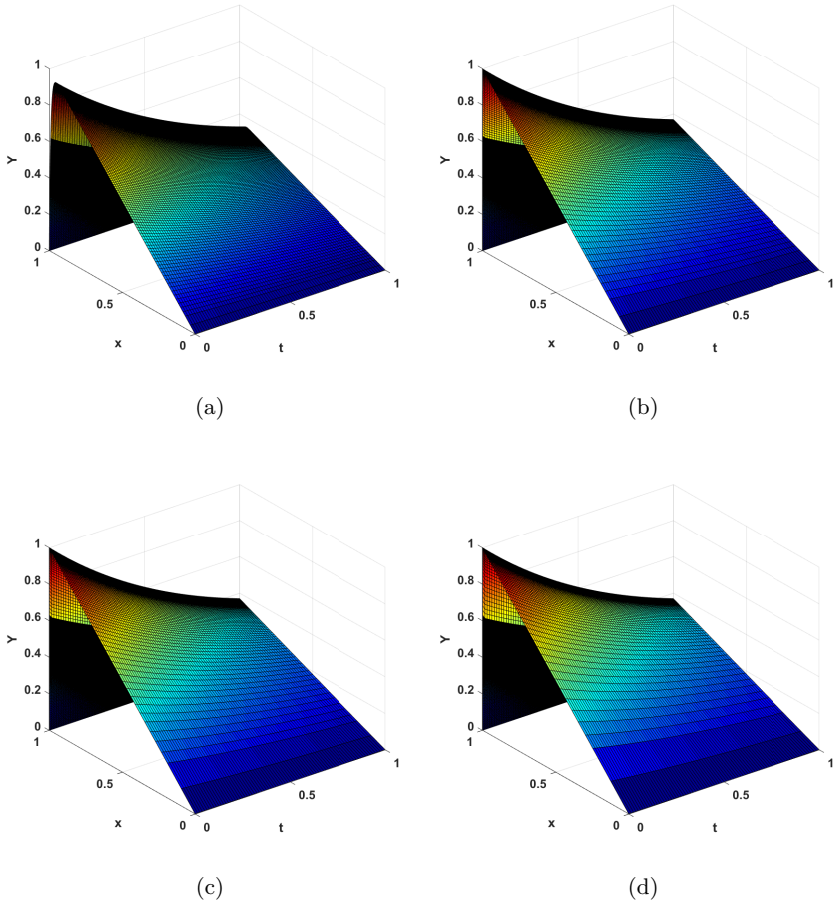


Figure 5: Solution profile for Example 2 using simple upwind scheme and after Richardson extrapolation technique with different value of ε and N. (a) Solution of $N = 128, \varepsilon = 10^{-2}$ (b) Solution of $N = 128, \varepsilon = 10^{-4}$ (c) Solution of $N = 128, \varepsilon = 10^{-6}$ (d) Solution of $N = 128, \varepsilon = 10^{-8}$.

Table 1: Maximum point-wise errors and the corresponding order of convergence before extrapolation for Example 1 on Modified graded mesh

	Number of Intervals $N/\Delta\theta$					
ε	$128/\frac{1}{10}$	$256/\frac{1}{20}$	$512/\frac{1}{40}$	$1024/\frac{1}{80}$	$2048/\frac{1}{160}$	$4096/\frac{1}{320}$
10^{-2}	$1.3707e-02$	$7.5039e-03$	$3.9359e-03$	$2.0188e-03$	$1.0231e-03$	$5.1505e-04$
	0.8692	0.9310	0.9632	0.9805	0.9902	
10^{-4}	$3.2822e-02$	$1.7760e-02$	$9.1788e-03$	$4.7257e-03$	$2.3904e-03$	$1.2045e-03$
	0.8860	0.9523	0.9578	0.9832	0.9889	
10^{-6}	$4.8001e-02$	$2.6420e-02$	$1.3866e-02$	$7.1029e-03$	$3.5934e-03$	$1.8063e-03$
	0.8614	0.9302	0.9650	0.9831	0.9923	
10^{-8}	$6.2334e-02$	$3.4617e-02$	$1.8304e-02$	$9.4150e-03$	$4.7736e-03$	$2.4034e-03$
	0.8485	0.9193	0.9591	0.9799	0.9900	
10^{-10}	$7.4394e-02$	$4.2641e-02$	$2.2683e-02$	$1.1706e-02$	$5.9463e-03$	$2.9968e-03$
	0.8030	0.9106	0.9544	0.9772	0.9886	
$\mathbb{E}^{N,\triangle\theta}$	$7.4394e-02$	$4.2641e-02$	$2.2683e-02$	$1.1706e-02$	$5.9463e-03$	$2.9968e-03$
$\mathbb{P}^{N,\triangle\theta}$	0.8030	0.9106	0.9544	0.9772	0.9886	

Table 2: Maximum point-wise errors and the corresponding order of convergence before extrapolation for Example 2 on Modified graded mesh

	Number of Intervals $N/\Delta\theta$					
ε	$128/\frac{1}{10}$	$256/\frac{1}{20}$	$512/\frac{1}{40}$	$1024/\frac{1}{80}$	$2048/\frac{1}{160}$	$4096/\frac{1}{320}$
10^{-2}	$1.3635e-02$	$7.4761e-03$	$3.9204e-03$	$2.0104e-03$	$1.0185e-03$	$5.1265e-04$
	0.8669	0.9313	0.9636	0.9811	0.9903	
10^{-4}	$3.2821e-02$	$1.7761e-02$	$9.1790e-03$	$4.7258e-03$	$2.3905e-03$	$1.2045e-03$
	0.8859	0.9523	0.9578	0.9832	0.9889	
10^{-6}	$4.7999e-02$	$2.6420e-02$	$1.3865e-02$	$7.1029e-03$	$3.5934e-03$	$1.8063e-03$
	0.8613	0.9302	0.9650	0.9831	0.9923	
10^{-8}	$6.2332e-02$	$3.4617e-02$	$1.8304e-02$	$9.4150e-03$	$4.7736e-03$	$2.4034e-03$
	0.8485	0.9193	0.9591	0.9799	0.9900	
10^{-10}	$7.4392e-02$	$4.2641e-02$	$2.2683e-02$	$1.1706e-02$	$5.9463e-03$	$2.9968e-03$
	0.8029	0.9106	0.9544	0.9772	0.9886	
$\mathbb{E}^{N,\triangle\theta}$	$7.4392e-02$	$4.2641e-02$	$2.2683e-02$	$1.1706e-02$	$5.9463e-03$	$2.9968e-03$
$\mathbb{P}^{N,\triangle\theta}$	0.8029	0.9106	0.9544	0.9772	0.9886	

Table 3: Maximum point-wise errors and the corresponding order of convergence after extrapolation for Example 1 on Modified graded mesh

	Number of Intervals $N/\Delta\theta$					
ε	$128/\frac{1}{10}$	$256/\frac{1}{20}$	$512/\frac{1}{40}$	$1024/\frac{1}{80}$	$2048/\frac{1}{160}$	$4096/\frac{1}{320}$
10^{-2}	$1.8300e-03$	$6.7491e-04$	$2.1138e-04$	$6.1563e-05$	$1.6744e-05$	$4.3943e-06$
	1.4390	1.6748	1.7797	1.8784	1.9299	
10^{-4}	$1.3132e-03$	$4.8091e-04$	$2.4003e-04$	$1.6554e-04$	$1.2649e-04$	$8.9042e-05$
	1.4492	1.0025	1.5360	1.3881	1.5065	
10^{-6}	$2.5422e-03$	$7.2236e-04$	$1.9374e-04$	$5.1776e-05$	$1.4992e-05$	$5.6506e-06$
	1.8153	1.8986	1.9038	1.7881	1.4077	
10^{-8}	$4.3611e-03$	$1.2692e-03$	$3.3970e-04$	$8.8050e-05$	$2.2424e-05$	$5.6808e-06$
	1.7808	1.9015	1.9479	1.9733	1.9809	
10^{-10}	$6.5786e-03$	$1.9550e-03$	$5.2916e-04$	$1.3762e-04$	$3.5092e-05$	$8.8610e-06$
	1.7506	1.8854	1.9430	1.9715	1.9856	
$\mathbb{E}^{N,\triangle\theta}$	$6.5786e-03$	$1.9550e-03$	$5.2916e-04$	$1.3762e-04$	$3.5092e-05$	$8.8610e-06$
$\mathbb{P}^{N,\triangle\theta}$	1.7506	1.8854	1.9430	1.9715	1.9856	

Table 4: Maximum point-wise errors and the corresponding order of convergence after extrapolation for Example 2 on Modified graded mesh

	Number of Intervals $N/\Delta\theta$					
ε	$128/\frac{1}{10}$	$256/\frac{1}{20}$	$512/\frac{1}{40}$	$1024/\frac{1}{80}$	$2048/\frac{1}{160}$	$4096/\frac{1}{320}$
10^{-2}	$1.8144e-03$	$6.6235e-04$	$2.0693e-04$	$6.0125e-05$	$1.6335e-05$	$4.2857e-06$
	1.4538	1.6784	1.7831	1.8800	1.9304	
10^{-4}	$1.3142e-03$	$4.8110e-04$	$2.4012e-04$	$1.6561e-04$	$1.2653e-04$	$8.9056e-05$
	1.4498	1.0026	1.5360	1.3882	1.5067	
10^{-6}	$2.5430e-03$	$7.2242e-04$	$1.9375e-04$	$5.1777e-05$	$1.4991e-05$	$5.6505e-06$
	1.8156	1.8987	1.9038	1.7882	1.4077	
10^{-8}	$4.3619e-03$	$1.2692e-03$	$3.3971e-04$	$8.8050e-05$	$2.2424e-05$	$5.6808e-06$
	1.7810	1.9016	1.9479	1.9733	1.9809	
10^{-10}	$6.5793e-03$	$1.9550e-03$	$5.2917e-04$	$1.3762e-04$	$3.5092e-05$	$8.8617e-06$
	1.7507	1.8854	1.9430	1.9715	1.9855	
$\mathbb{E}^{N,\triangle\theta}$	$6.5793e-03$	$1.9550e-03$	$5.2917e-04$	$1.3762e-04$	$3.5092e-05$	$8.8617e-06$
$\mathbb{P}^{N,\triangle\theta}$	1.7507	1.8854	1.9430	1.9715	1.9855	

Table 5: Comparison of maximum point-wise errors and the corresponding order of convergence before extrapolation for Example 1

	Number of Intervals $N/\Delta\theta$					
ε	$128/\frac{1}{10}$	$256/\frac{1}{20}$	$512/\frac{1}{40}$	$1024/\frac{1}{80}$	$2048/\frac{1}{160}$	$4096/\frac{1}{320}$
10^{-2}	$1.3707e-02$	$7.5039e-03$	$3.9359e-03$	$2.0188e-03$	$1.0231e-03$	$5.1505e-04$
	0.8692	0.9310	0.9632	0.9805	0.9902	
10^{-4}	$3.2822e-02$	$1.7760e-02$	$9.1788e-03$	$4.7257e-03$	$2.3904e-03$	$1.2045e-03$
	0.8860	0.9523	0.9578	0.9832	0.9889	
10^{-6}	$4.8001e-02$	$2.6420e-02$	$1.3866e-02$	$7.1029e-03$	$3.5934e-03$	$1.8063e-03$
	0.8614	0.9302	0.9650	0.9831	0.9923	
$\mathbb{E}^{N,\triangle\theta}$	$4.8001e-02$	$2.6420e-02$	$1.3866e-02$	$7.1029e-03$	$3.5934e-03$	$1.8063e-03$
$\mathbb{P}^{N,\triangle\theta}$	0.8614	0.9302	0.9650	0.9831	0.9923	

Results in [18]

	Number of Intervals N					
ε	32	64	128	256	512	1024
10^{-2}	$4.3463e-02$	$2.9874e-02$	$1.8761e-02$	$1.1330e-02$	$6.5980e-03$	$3.7392e-03$
	0.5408	0.6711	0.7276	0.7799	0.8192	
10^{-4}	$4.2785e-02$	$2.9447e-02$	$1.8513e-02$	$1.1186e-02$	$6.5158e-03$	$3.6933e-03$
	0.5389	0.6695	0.7269	0.7796	0.8190	
10^{-6}	$4.2779e-02$	$2.9443e-02$	$1.8511e-02$	$1.1184e-02$	$6.5150e-03$	$3.6928e-03$
	0.5389	0.6695	0.7269	0.7796	0.8190	
$\mathbb{E}^{N,\triangle\theta}$	$4.2779e-02$	$2.9443e-02$	$1.8511e-02$	$1.1184e-02$	$6.5150e-03$	$3.6928e-03$
$\mathbb{P}^{N,\triangle\theta}$	0.5389	0.6695	0.7269	0.7796	0.8190	

Table 6: Comparison of maximum point-wise errors and the corresponding order of convergence after extrapolation for Example 1

	Number of Intervals $N/\Delta\theta$					
ε	$128/\frac{1}{10}$	$256/\frac{1}{20}$	$512/\frac{1}{40}$	$1024/\frac{1}{80}$	$2048/\frac{1}{160}$	$4096/\frac{1}{320}$
10^{-2}	$1.8300e-03$	$6.7491e-04$	$2.1138e-04$	$6.1563e-05$	$1.6744e-05$	$4.3943e-06$
	1.4390	1.6748	1.7797	1.8784	1.9299	
10^{-4}	$1.3132e-03$	$4.8091e-04$	$2.4003e-04$	$1.6554e-04$	$1.2649e-04$	$8.9042e-05$
	1.4492	1.0025	1.5360	1.3881	1.5065	
10^{-6}	$2.5422e-03$	$7.2236e-04$	$1.9374e-04$	$5.1776e-05$	$1.4992e-05$	$5.6506e-06$
	1.8153	1.8986	1.9038	1.7881	1.4077	
$\mathbb{E}^{N,\triangle\theta}$	$2.5422e-03$	$7.2236e-04$	$1.9374e-04$	$5.1776e-05$	$1.4992e-05$	$5.6506e-06$
$\mathbb{P}^{N,\triangle\theta}$	1.8153	1.8986	1.9038	1.7881	1.4077	
Results in [18]						
	Number of Intervals N					
ε	32	64	128	256	512	1024
10^{-2}	$6.7626e-03$	$2.9552e-03$	$1.2260e-03$	$4.4704e-03$	$1.5112e-04$	$4.8606e-05$
	1.1943	1.2693	1.4555	1.5646	1.6365	
10^{-4}	$6.7299e-03$	$2.9308e-03$	$1.2146e-03$	$4.4264e-04$	$1.4950e-04$	$4.8044e-05$
	1.1993	1.2708	1.4563	1.5660	1.6377	
10^{-6}	$6.7297e-03$	$2.9306e-03$	$1.2145e-03$	$4.4260e-04$	$1.4949e-04$	$4.8039e-05$
	1.1993	1.2708	1.4563	1.5660	1.6378	
$\mathbb{E}^{N,\triangle\theta}$	$6.7297e-03$	$2.9306e-03$	$1.2145e-03$	$4.4260e-04$	$1.4949e-04$	$4.8039e-05$
$\mathbb{P}^{N,\triangle\theta}$	1.1993	1.2708	1.4563	1.5660	1.6378	

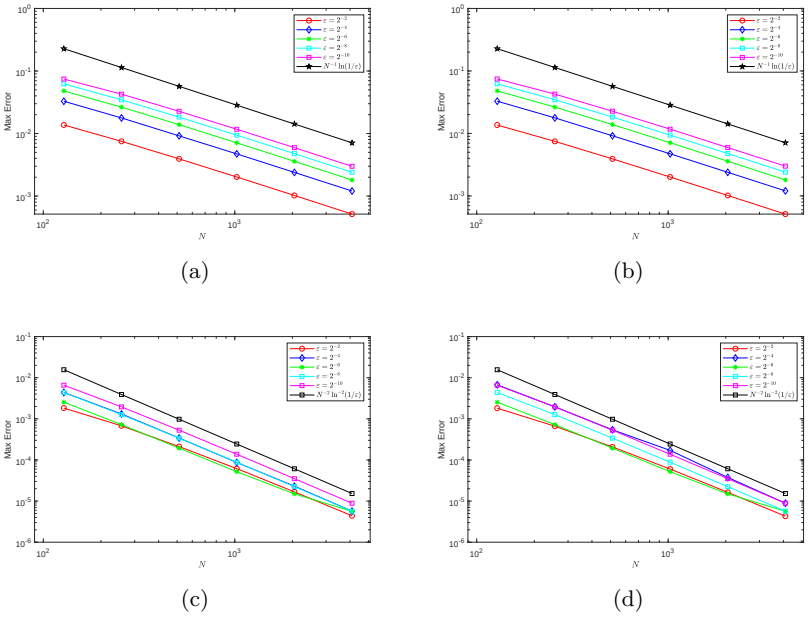


Figure 6: Before extrapolation and after extrapolation Loglog plot of the maximum point-wise errors for Example 1 and Example 2. (a) Before extrapolation for Example 1 (b) Before extrapolation for Example 2 (c) After extrapolation for Example 1 (d) After extrapolation for Example 2.

References

[1] Cai, X. and Liu, F. *A Reynolds uniform scheme for singularly perturbed parabolic differential equation*, ANZIAM Journal 47 (2005), C633–C648.

[2] Clavero, C., Gracia, J.L. and Jorge, J.C. *High-order numerical methods for one-dimensional parabolic singularly perturbed problems with regular layers*, Numer. Methods Partial. Differ. Equ. 21(1) (2005), 149–169.

[3] Clavero, C., Jorge, J.C., and Lisbona, F. *A uniformly convergent scheme on a nonuniform mesh for convection-diffusion parabolic problems*, J. Comput. Appl. Math. 154(2) (2003), 415–429.

[4] Deb, B.S. and Natesan, S. *Richardson extrapolation method for singularly perturbed coupled system of convection-diffusion boundary-value problems*, CMES Comput. Model. Eng. Sci. 38(2) (2008), 179–200.

[5] Deb, R. and Natesan, S. *Higher-order time accurate numerical methods for singularly perturbed parabolic partial differential equations*, Int. J. Comput. Math. 86(7) (2009), 1204–1214.

[6] Debela, H.G., Woldaregay, M.M. and Duressa, G.F. *Robust numerical method for singularly perturbed convection-diffusion type problems with non-local boundary condition*, Math. Model. Anal. 27(2) (2022), 199–214.

[7] Friedman, A. *Partial differential equations of parabolic type*, Courier Dover Publications, 2008.

[8] Hailu, W.S. and Duressa, G.F. *Accelerated parameter-uniform numerical method for singularly perturbed parabolic convection-diffusion problems with a large negative shift and integral boundary condition*, Results Appl. Math. 18 (2023), 100364.

[9] Hemker, P.W., Shishkin, G.I. and Shishkina, L.P. *ε-uniform schemes with high-order time-accuracy for parabolic singular perturbation problems*, IMA journal of numerical analysis 20(1) (2000), 99–121.

[10] Hemker, P.W., Shishkin, G.I. and Shishkina, L.P. *High-order time-accurate schemes for singularly perturbed parabolic convection-diffusion problems with robin boundary conditions*, Comput. Methods Appl. Math. 2(1) (2002), 3–25.

[11] Hemker, PW. Shishkin, GI. and Shishkina, LP. *High-order accurate decomposition of richardson's method for a singularly perturbed elliptic reaction-diffusion equation*, Comput. Math. Math. Phys. 44(2) (2004), 309–316.

[12] Keller, H.B. *Numerical solution of two point boundary value problems*, Society for industrial and applied mathematics, 1976.

[13] Kellogg, R.B. and Tsan, A. *Analysis of some difference approximations for a singular perturbation problem without turning points*, Math. Comput. 32(144) (1987), 1025–1039.

[14] Kopteva, N.V. *On the uniform in small parameter convergence of a weighted scheme for the one-dimensional time-dependent convection-diffusion equation*, Comput. Math. Math. Phys. 37 (1997), 1173–1180.

[15] Mekonnen, T.B. and Duressa, G.F. *Uniformly convergent numerical method for two-parametric singularly perturbed parabolic convection-diffusion problems*, J. Appl. Comput. Mech. 7(2) (2021), 535–545.

[16] Miller, J.J.H., O'riordan, E. and Shishkin, G.I. *Fitted numerical methods for singular perturbation problems: error estimates in the maximum norm for linear problems in one and two dimensions*, World scientific, 1996.

[17] Mukherjee, K. and Natesan, S. *Parameter-uniform hybrid numerical scheme for time-dependent convection-dominated initial-boundary-value problems*, Computing 84(3) (2009), 209–230.

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

[19] Natividad, M.C. and Stynes, M. *Richardson extrapolation for a convection–diffusion problem using a shishkin mesh*, Appl. Numer. Math. 45(2-3) (2003), 315–329.

[20] Negero, N.T. *A uniformly convergent numerical scheme for two parameters singularly perturbed parabolic convection–diffusion problems with a large temporal lag*, Results Appl. Math. 16 (2022), 100338.

[21] Negero, N.T. *A fitted operator method of line scheme for solving two-parameter singularly perturbed parabolic convection-diffusion problems with time delay*, J. Math. Model. 11(2) (2023), 395–410.

[22] Negero, N.T. *A parameter-uniform efficient numerical scheme for singularly perturbed time-delay parabolic problems with two small parameters*, Partial Differ. Equ. Appl. Math. 7 (2023), 100518.

[23] Negero, N.T. *A robust fitted numerical scheme for singularly perturbed parabolic reaction–diffusion problems with a general time delay*, Results Phys. 51 (2023), 106724.

[24] Negero, N.T. *A robust uniformly convergent scheme for two parameters singularly perturbed parabolic problems with time delay*, Iran. J. Num. Anal. Optim. 13 (4) (2023), 627–645.

[25] Negero, N.T. *Fitted cubic spline in tension difference scheme for two-parameter singularly perturbed delay parabolic partial differential equations*, Partial Differ. Equ. Appl. Math. (2023), 100530.

[26] Negero, N.T. and Duressa, G.F. *Parameter-uniform robust scheme for singularly perturbed parabolic convection-diffusion problems with large time-lag*, Comput. Methods Differ. Equ. 10(4) (2022), 954–968.

[27] Ng-Stynes, M.J., O'Riordan, E. and Stynes, M. *Numerical methods for time-dependent convection-diffusion equations*, J. Comput. Appl. Math. 21(3) (1988), 289–310.

[28] O'Riordan, E., Pickett, M. and Shishkin, G. *Parameter-uniform finite difference schemes for singularly perturbed parabolic diffusion-convection-reaction problems*, Math. Comput. 75(255) (2006), 1135–1154.

[29] Roos, H-G., Stynes, M. and Tobiska, L. *Robust numerical methods for singularly perturbed differential equations: convection-diffusion-reaction and flow problems*, Springer Science & Business Media, 24, 2008.

[30] Shishkin, G.I. *Robust novel high-order accurate numerical methods for singularly perturbed convection-diffusion problems*, Math. Model. Anal. 10(4) (2005), 393–412.

[31] Shishkin, G.I. and Shishkina, L.P. *A higher-order richardson method for a quasilinear singularly perturbed elliptic reaction-diffusion equation*, Differential Equations 41(7) (2005), 1030–1039.

[32] Shishkin, G.I. and Shishkina, L.P. *The Richardson extrapolation technique for quasilinear parabolic singularly perturbed convection-diffusion*

equations, In Journal of Physics: Conference Series 55 (2006), 019, IOP Publishing.

[33] Singh, M.K. and Natesan, S. *Richardson extrapolation technique for singularly perturbed system of parabolic partial differential equations with exponential boundary layers*, Appl. Math. Comput. 333 (2018), 254–275.

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

[35] Woldaregay, M.M. and Duressa, G.F. *Exponentially fitted tension spline method for singularly perturbed differential difference equations*, Iran. J. Numer. Anal. Optim. 11(2) (2021), 261–282.

Iranian Journal of Numerical Analysis and Optimization
Vol. 14, No. 1, 2024, pp 265–290
https://doi.org/10.22067/ijnao.2023.84198.1307
https://ijnao.um.ac.ir/

CC BY NC

Research Article

Noisy label relabeling by nonparallel support vector machine

A. Sahleh and M. Salahi*,

Abstract

In machine learning, models are derived from labeled training data where labels signify classes and features define sample attributes. However, noise from data collection can impair the algorithm's performance. Blanco, Japón, and Puerto proposed mixed-integer programming (MIP) models within support vector machines (SVM) to handle label noise in training datasets. Nonetheless, it is imperative to underscore that their models demonstrate an observable escalation in the number of variables as sample size increases. The nonparallel support vector machine (NPSVM) is a binary classification method that merges the strengths of both SVM and twin SVM. It accomplishes this by determining two nonparallel hyperplanes by solving two optimization problems. Each hyperplane is strategically positioned to be closer to one of the classes while maximizing its distance

*Corresponding author

Received 30 August 2023; revised 4 December 2023; accepted 31 December 2023

Ali Sahleh
Department of Applied Mathematics, Faculty of Mathematical Sciences, Rasht, Iran. e-mail: alisahleh@gmail.com

Maziar Salahi
Department of Applied Mathematics, Faculty of Mathematical Sciences, University of Guilan, Rasht, Iran. e-mail: salahim@guilan.ac.ir

How to cite this article
Sahleh, A. and Salahi, M., Noisy label relabeling by nonparallel support vector machine. *Iran. J. Numer. Anal. Optim.*, 2024; 14(1): 265-290. https://doi.org/10.22067/ijnao.2023.84198.1307

from the other class. In this paper, to take advantage of NPSVM's features, NPSVM-based relabeling (RENPSVM) MIP models are developed to deal with the label noises in the dataset. The proposed model adjusts observation labels and seeks optimal solutions while minimizing computational costs by selectively focusing on class-relevant observations within an ϵ-intensive tube. Instances exhibiting similarities to the other class are excluded from this ϵ-intensive tube. Experiments on 10 UCI datasets show that the proposed NPSVM-based MIP models outperform their counterparts in accuracy and learning time on the majority of datasets.

AMS subject classifications (2020): Primary 6BT09; Secondary 90C11.

Keywords: Label noise; SVM; Mixed-integer program; Nonparallel SVM.

1 Introduction

Support Vector Machine (SVM) [8, 34, 35] is a renowned technique employed for binary classification in diverse domains, such as abnormal recognition [22], stock market prediction [1], and pose estimation [36]. Despite its proficient performance, SVM encounters substantial computational demands when solving the Quadratic Programming Problem (QPP) for large datasets. In response to this challenge, Jayadeva, Khemchandani, and Chandra [19] introduced the Twin SVM (TWSVM), a method that utilizes two nonparallel hyperplanes. These hyperplanes are positioned closer to each of the two classes while maintaining a minimum unit distance from samples of the other class. In contrast to SVM, TWSVM tackles two smaller QPPs, thereby mitigating the training time complexity. The TWSVM framework has been extended through various adaptations, including the Wavelet TWSVM by Ding et al. [11, 12] with glowworm swarm optimization, an enhanced K-nearest neighbor TWSVM by Nasiri and Mir [26] addressing noise and outliers, and an automatic TWSVM by Jimenez-Castano, Alvarez-Meza, and Orozco-Gutierrez [20] for imbalanced datasets using kernel representation. Although TWSVM offers valuable attributes, it encounters a challenge in computing the inverses of specific matrices as part of its model training process. This task becomes impractical or even infeasible for siz-

able datasets when utilizing conventional methods. Conversely, the standard SVM can efficiently solve large-scale problems through algorithms like Sequential Minimal Optimization (SMO). To address this concern, Nonparallel SVM (NPSVM) [33] is introduced, which integrates SVM's benefits into TWSVM. This integration incorporates the utilization of the SMO algorithm [21, 28] and the concept of semi-sparseness [33], collectively enhancing the overall performance of the model.

The existence of label noise within datasets can have a substantial impact on the accuracy and generalizability of supervised learning algorithms. Instances with incorrect labels may originate from diverse origins, such as human errors, label switching, or intentional introduction of noise [25, 23, 24, 27]. It has been the subject of several research studies. Angluin and Laird [2] introduced a noise model that establishes a sample for learning in noisy environments. They also suggested computationally feasible learning algorithms for noisy domains and explored extending these concepts to broader contexts. However, a drawback of their model is the question of whether there are domains where approximately correct identification is computationally feasible without noise. However, it becomes computationally infeasible even with moderate levels of noise. Another study by Xiao et al. [38] devised an optimal attack strategy and used heuristic methods for practical computation. Biggio, Nelson, and Laskov [4] introduced an algorithmic strategy that effectively manages adversarial alterations of labels. This technique involves the adjustment of the kernel matrix when labels are independently flipped with equal probabilities. Another alternative, as presented in a prior work [15], entails the process of detecting and eliminating inaccurately labeled instances. This involves the selection of samples considered dubious and necessitating additional scrutiny. Obtaining labels with reduced levels of noise might entail increased time and expenses. Nevertheless, this endeavor holds the potential to significantly augment classification accuracy. To address this challenge, Duan and Wu [13] proposed a novel learning approach that leverages both noisy and less noisy labels extracted from a limited portion of the training dataset. This methodology involves the estimation of noise rate parameters and the inference of precise labels by utilizing a noise model built upon flipping probabilities and a logistic regres-

sion classifier. While the methods presented in [4, 15] effectively tackle label noise in data, they do so through a two-phase process. The aspect related to label noise is addressed before the training process, and the models do not handle label noise simultaneously with the primary task. Thulasidasan et al. [31] introduced an innovative approach to mitigate label noise within the context of Deep Neural Network (DNN) classification. Their methodology involves the introduction of a new loss function, enabling the DNN to decide to abstain from classifying certain samples, thereby avoiding confusion. At the same time, this approach improves the classification performance of samples that are not abstained from. The proposed method holds substantial promise for considerably augmenting the accuracy and robustness of DNN classifiers in real-world practical applications. In [7], the authors presented a technique for estimating the level of label noise and showed that implementing importance reweighting can enhance classification accuracy when dealing with label noise and evaluates the reliability of two classification approaches: Convolutional neural networks and convolutional neural networks with importance reweighting. Despite the merits of models in [31, 7], they suffer from the explosion of parameters as the number of layers increases for some tasks, such as natural language processing and computer vision, which results in demanding high computation resources. Blanco, Japón, and Puerto [5] introduced a unique approach for constructing optimal classification trees that consider the presence of noisy labels in the training data. Their method combines margin-based classifiers with outlier detection techniques to improve performance. It utilizes two main components: (1) the splitting rules of the tree are designed to maximize class separation margins, following SVM principles, and (2) during tree construction, some training sample labels can be adjusted to identify and address label noise. These elements are integrated to create the final optimal classification tree. Bertsimas et al. [3] introduced a robust optimization approach for addressing label noise by introducing a new variable representing the probability of mislabeling for each training point. They also imposed a constraint to limit the total number of mislabeled points below a specified threshold, considering worst-case scenarios. However, a limitation of their method is its primary focus on constructing a classifier robustly, concentrating on worst-case scenarios, and

controlling label noise with a budget hyperparameter. Also, their approach prioritizes worst-case scenarios and does not explore all possible parameter vectors (w, b) for different scenarios.

Recently, in [6], the authors have formulated SVM-based mixed-integer programming (MIP) models to effectively handle classification tasks in the presence of noisy labels. In contrast to the existing techniques, their methodology involves a simultaneous process of constructing an SVM-based classifier while adjusting the labels of observations to achieve an optimal solution. A significant advantage of their approach lies in its capability to derive separating hyperplanes that conventional SVM methods cannot achieve. However, it is important to highlight that the Relabel SVM-based (RESVM) model introduced in [6] exhibits a noteworthy increase in the number of variables as the number of samples rises. To tackle this issue, they further proposed a clustering-based relabeling (CRESVM) model by employing clustering and classification in the SVM framework.

In this paper, the relabeling idea is employed within NPSVM to take benefit of its features. Each proposed MIP model considers instances within the class represented by the ϵ-intensive tube. If any of these instances share similarities with the other class, the samples belonging to the original class are excluded from the ϵ-insensitive tube. In most datasets, RENPSVM exhibits fewer linear constraints and variables in comparison to RESVM. Moreover, while the CRESVM model introduced in [6] possesses fewer linear constraints and variables than both RESVM and RENPSVM, its demerit is the utilization of nonlinear constraints. Besides the above, the structure of NPSVM allows parallel implementation of the proposed MIP models, leading to faster learning times on most datasets. The main contributions of this paper are summarized as follows:

(1) Expanding the NPSVM models into MIP models to address label noise in a manner that not only adjusts the labels of observations but also achieves an optimal solution simultaneously.

(2) Minimizing the computational cost by avoiding the consideration of all observations as potential candidates for the label changes in the model.

Instead, we focus on instances related to the class that the model aims to represent within an epsilon (ϵ)-intensive tube.

(3) Parallelization of the proposed MIP models, which results in faster learning times on the majority of datasets.

(4) Computational experiments conducted on 10 UCI datasets reveal that RENPSVM outperforms RESVM and CRESVM in terms of classification accuracy while demonstrating similar learning times to RESVM and CRESVM.

(5) The outcomes of evaluating our algorithms on diverse real-world datasets demonstrate that our suggestions exhibit greater resilience against attacks than the recent relabel models approach mentioned in [6].

The rest of this paper is organized as follows. Section 2 briefly reviews TWSVM and NPSVM. In Section 3, we delve into our proposed model and provide a comparison of the number of constraints and variables with RESVM and CRESVM models in [6]. Moving on to Section 4, computational experiments are conducted on 10 UCI datasets to illustrate the efficiency of the proposed models in comparison to those outlined in [6]. Also, this section encompasses two statistical tests aimed at highlighting differences between the proposed model and those from [6]. Ultimately, Section 5 presents concluding remarks.

2 Background

Consider a classification problem with the dataset $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_l, y_l)\}$, where $x_i \in R^n$ and $y_i \in \{+1, -1\}$ for $i = 1, \ldots, l$ denote samples and labels of samples, respectively. We further symbolize the sets of indices associated with positive and negative classes as I^+ and I^-, respectively. This is defined as

$$I^+ = \{i \,|y_i = +1\}, \quad I^- = \{i \,|y_i = -1\}.$$

In this section, we present a concise overview of the TWSVM and NPSVM models.

2.1 TWSVM

Consider $A \in \mathbb{R}^{m_1 \times n}$ and $B \in \mathbb{R}^{m_2 \times n}$ as the data matrices containing points belonging to I^+ and I^-, respectively. The TWSVM functions as a binary classifier, establishing two nonparallel hyperplanes via solving two smaller QPPs compared to a large one in SVM as follows:

$$\begin{aligned} \min_{w_1,\, b_1,\, \xi_2} \ & \tfrac{1}{2}||Aw_1 + e_1 b_1||^2 + c_1 e_1^T \xi_2 \\ \text{s.t. } & -(Bw_1 + e_2 b_1) + \xi_2 \geq e_2, \\ & \xi_2 \geq 0, \end{aligned} \tag{1}$$

and

$$\begin{aligned} \min_{w_2,\, b_2,\, \xi_1} \ & \tfrac{1}{2}||Bw_2 + e_2 b_2||^2 + c_2 e_2^T \xi_1 \\ \text{s.t. } & (Aw_2 + e_1 b_2) + \xi_1 \geq e_1, \\ & \xi_1 \geq 0, \end{aligned} \tag{2}$$

where c_1 and c_2 are predetermined trade-off factors between the error variable vectors ξ_1 and ξ_2. Also, e_1 and e_2 are vectors of ones with appropriate dimensions. The first term in the objective function of (1) (or (2)) aims to maintain the hyperplane in proximity to the points of one class (I^+), while the constraints work to ensure that the hyperplane remains at a unit distance from the points of the other class (I^-). The Wolfe dual forms of (1) and (2) are given by

$$\begin{aligned} \max_{\alpha} \ & e_2^T \alpha - \tfrac{1}{2} \alpha^T G (H^T H)^{-1} G^T \alpha \\ \text{s.t. } & 0 \leq \alpha \leq c_1 e_2 \end{aligned} \tag{3}$$

and

$$\begin{aligned} \max_{\beta} \ & e_2^T \beta - \tfrac{1}{2} \beta^T P (Q^T Q)^{-1} P^T \beta \\ \text{s.t. } & 0 \leq \beta \leq c_2 e_1, \end{aligned} \tag{4}$$

where $G = [B; e_2]$, $H = [A; e_1]$, $P = [A; e_1]$, and $Q = [B; e_2]$. As we see, dual models involve using the inverse of $G^T G$ and $H^T H$, which are multiplied by Lagrangian multipliers $\alpha \in \mathbb{R}^{m_1}$ and $\beta \in \mathbb{R}^{m_2}$. Finally, the nonparallel hyperplanes are obtained from the solutions α and β of (3) and (4), respectively, through

$$z_1 = -(H^T H)^{-1} G^T \alpha, \quad \text{where } \ z_1 = [w_1^T \ b_1] \tag{5}$$

and

$$z_1 = -(Q^T Q)^{-1} P^T \beta, \quad \text{where } z_2 = [w_2^T \ b_2]. \tag{6}$$

While TWSVM handles smaller QPPs compared to SVM, it is not without drawbacks [32]. Firstly, TWSVM, in addressing its primal problems, exclusively minimizes empirical risk, neglecting the essential aspect of minimizing structural risk present in conventional SVMs. Secondly, to handle singularity concerns, TWSVM employs approximations by substituting inverse matrices, leading to solutions that are only approximative. Thirdly, the computational complexity of TWSVM is impeded by the necessity to compute inverse matrices, rendering it impractical for extensive datasets. Moreover, TWSVM is confined to linear classification and lacks a straightforward extension to nonlinear scenarios. The demand for swift solvers, such as the SMO algorithm used for standard SVMs, adds another layer of complexity. Lastly, TWSVM compromises sparsity by employing a quadratic loss function, resulting in a situation where the majority of points in a class exert substantial influence on each decision function, consequently forfeiting the advantages associated with sparsity.

2.2 The NPSVM

The NPSVM, which is a generalized version of TWSVM, provides a more comprehensive formulation than TWSVM and determines two nonparallel hyperplanes using a similar approach. The key difference is that NPSVM represents each class within ϵ-insensitive tubes (Figure 1) and inherits the advantages of SVM that TWSVM lacks, such as utilizing the SMO algorithm and avoiding the computation of matrix inverses during its model training process. The NPSVM solves the following two QPPs:

$$\begin{aligned} \min_{w_1,\, ,b_1,\, \eta_1,\, \eta_2,\, \xi_1} \quad & \tfrac{1}{2}||w_1||^2 + c_1 e_1^T(\eta_1 + \eta_2) + c_2 e_2^T \xi_1 \\ \text{s.t. } & (w_1^T x_i + b_1) \le \epsilon_1 + \eta_{1i}, && i \in I^+, \\ & -(w_1^T x_i + b_1) \le \epsilon_1 + \eta_{2i}, && i \in I^+, \\ & w_1^T x_i + b_1 \le -1 + \xi_{1i} && i \in I^-, \\ & \xi_1,\ \eta_1,\ \eta_2 \ge 0, \end{aligned} \tag{7}$$

and

$$\begin{aligned}\min_{w_2,\, b_2,\, \eta_1',\, \eta_2',\, \xi_2} \quad & \tfrac{1}{2}||w_2||^2 + c_3 e_2^T(\eta_1' + \eta_2') + c_4 e_1^T \xi_1 \\ \text{s.t. } & w_2^T x_i + b_2 \le \epsilon_2 + \eta_{1i}', && i \in I^-, \\ & -(w_2^T x_i + b_2) \le \epsilon_2 + \eta_{2i}', && i \in I^-, \\ & w_2^T x_i + b_2 \ge 1 - \xi_{2i} && i \in I^+, \\ & \xi_2, \eta_1',\ \eta_2' \ge 0,\end{aligned} \tag{8}$$

where $c_i > 0$ $(i = 1, \ldots, 4)$ are trade-off factors for error variables η_i and ξ_i $(i = 1, 2)$. The aim of (7) is to maximize the margin between the hyperplanes of ϵ-intensive tube, which can be mathematically expressed as $\frac{2\epsilon}{\|w_1\|}$. The first and second set of constraints ensure that the positive class is largely concentrated within the ϵ-band situated between the hyperplanes $w_1^T x + b_1 = \epsilon$ and $(w_1^T x) + b_1 = -\epsilon$. The third set of constraints push away negative class from the hyperplane $w_1^T x + b_1 = -1$ as far as possible. Similar description holds for problem (8).

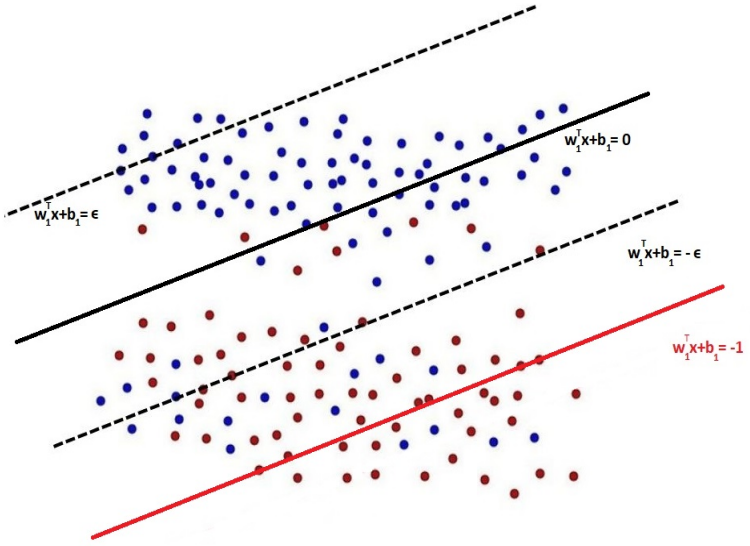


Figure 1: Illustration of NPSVM.

3 Relabel NPSVM

In contrast to TWSVM, NPSVM eliminates the need for specific matrix inversions during model training, making it particularly advantageous when dealing with substantial datasets, where traditional methods may become extremely challenging or unfeasible. However, akin to the standard SVM, NPSVM retains efficiency in addressing large-scale problems by utilizing the SMO algorithm. In NPSVM, the introduction of an ϵ-insensitive loss function naturally incorporates a regularization term. This characteristic distinguishes it from the initial TWSVM or improved TBSVM, with these latter models being special cases of the more general NPSVM. Notably, NPSVM reverts to the initial TWSVM or TBSVM when the corresponding parameters are appropriately chosen. Additionally, the transition from semi-sparseness to complete sparseness is promoted within the NPSVM framework [33]. In this section, we explore the implementation of the relabeling approach within NPSVM, aiming to bolster its robustness against label noise in datasets.

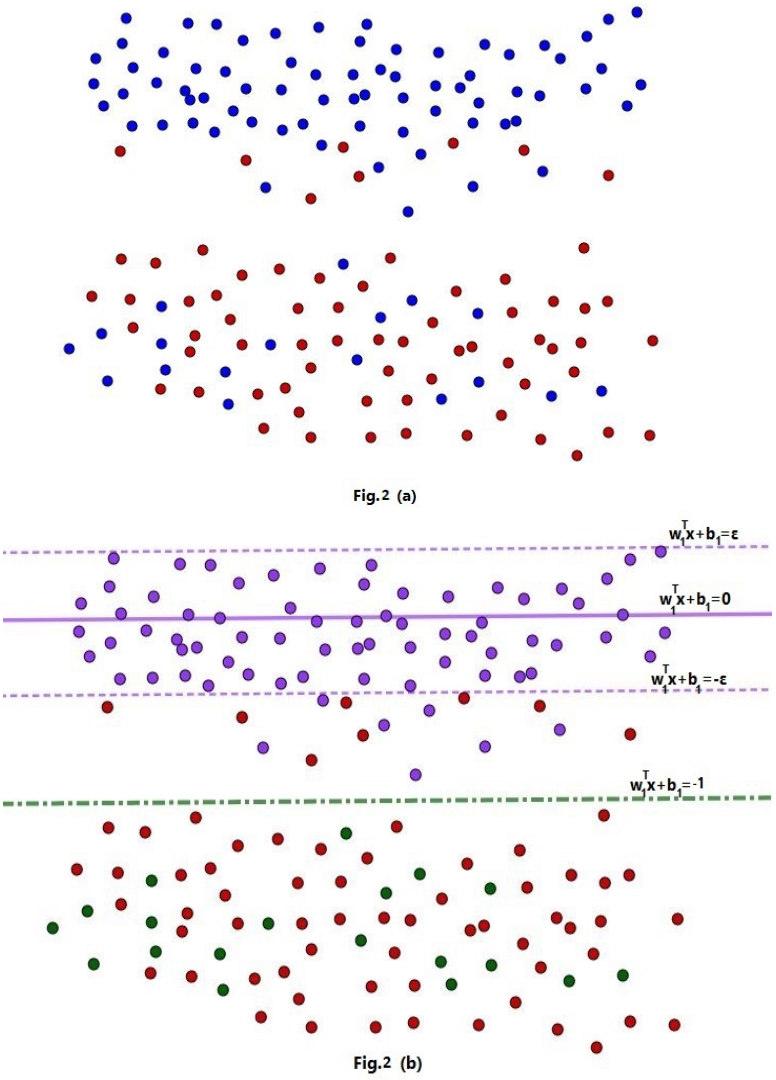


Figure 2: Original data dataset Figure 2 (a). Optimal separating hyperplanes with (9) Figure 2 (b). Instances from the positive dataset that remain within the ϵ-intensive tube are colored purple, while instances that are excluded from the respective constraint are colored green.

To apply relabeling on NPSVM, initially, we aim for the positive class to be predominantly positioned within the ϵ-intensive tube while maximizing its distance from the other class. The gap between the hyperplanes of ϵ-

intensive tube is controlled via the loss function. This results in enhancing the alignment between the nonparallel hyperplanes and the classes they represent. Due to the presence of label noise, there are some instances that appear to belong to a specific class based on their labels, but they bear a resemblance to a different class (the blue color in Figure 2 (a)). To determine whether the samples of the positive class are to be included within the ϵ-insensitive tube or not, a binary variable vector is incorporated into the model. This vector determines which instances are removed from the ϵ-insensitive tube (green color in Figure 2 (b)) and which instances remain inside it (purple color in Figure 2 (b)). This binary variable vector is integrated into both the constraints and the objective function, effectively preventing the extensive relabeling of observations. The first relabeling NPSVM (RENPSVM) model is formulated as follows:

$$\begin{aligned}
\min_{w_1,\, ,b_1,\, \eta_1,\, \eta_2,\, \xi_1,\, \theta_1} \quad & \tfrac{1}{2}||w_1||^2 + c_1 e_1^T(\eta_1+\eta_2) + c_2 e_3^T \xi_1 + c_3 e_1^T \theta_1 \\
\text{s.t. } & (w_1^T x_i + b_1)\theta_{1i} \le \epsilon_1 + \eta_{1i}, && i \in I^+, \\
& -(w_1^T x_i + b_1)\theta_{1i} \le \epsilon_1 + \eta_{2i}, && i \in I^+, \\
& (w_1^T x_i + b_1)(1-\theta_{1i}) \le -1 + \xi_{1i} + M_1(1-\theta_{1i}), && i \in I^+, \\
& w_1^T x_i + b_1 \le -1 + \xi_i && i \in I^-, \\
& \theta_{1i} \in \{0,1\}, && i \in I^+, \\
& w_1 \in \mathcal{R}^n,\ b_1 \in \mathcal{R}, \\
& \eta_1,\ \eta_2,\ \xi_1 \ge 0,
\end{aligned} \tag{9}$$

where $c_i > 0$ $(i = 1, 2, 3)$ are trade-off parameters and M_1 is large positive constant that is chosen such that its associated constraint becomes redundant when $\theta = 0$. The second term in objective function $\eta_1 + \eta_2$ controls the error for the gap between the hyperplanes $w_1^T x + b_1 = \epsilon$ and $w_1^T x + b_1 = -\epsilon$. In the third set of constraints, we strive to distance the negative class from the hyperplane $w_1^T x + b_1 = -1$. The error vector ξ_1 is assessed using the soft margin loss function. The binary variable vector θ_1 determines whether the samples of the positive class are to be included within the ϵ-insensitive tube or not. Typically, when $\theta_{1i} = 1$, it signifies that the ith instance belongs to the positive class. This is represented by the first three sets of constraints in (9). The final term in the objective function serves to avoid extensive reassignment

of labels to observations, a situation that might result in producing ineffectual classifiers. The second RENPSVM model corresponding to other classes is as follows:

$$\begin{aligned}
\min_{w_2,\ ,b_2,\ \eta_1',\ \eta_2',\ \xi_2,\theta_2} \quad & \tfrac{1}{2}||w_2||^2 + c_4 e_2^T(\eta_1' + \eta_2') + c_5 e_3^T \xi_2 + c_6 e_2^T \theta_2 \\
\text{s.t. } & (w_2^T x_i + b_2)\theta_{2i} \le \epsilon_2 + \eta_1', && i \in I^-, \\
& -(w_2^T x_i + b_2)\theta_{2i} \le \epsilon_2 + \eta_2', && i \in I^-, \\
& (w_2^T x_i + b_2)(1-\theta_{2i}) \ge 1 - \xi_{2i} - M_2(1-\theta_{2i}), && i \in I^-, \\
& w_2^T x_i + b_2 \ge 1 - \xi_{2i}, && i \in I^+, \\
& \theta_{2i} \in \{0,1\}, && i \in I^-, \\
& w_2 \in \mathcal{R}^n,\ b_2 \in \mathcal{R}, \\
& \eta_1',\ \eta_2',\ \xi_2 \ge 0.
\end{aligned} \tag{10}$$

Both models (9) and (10) exhibit nonlinearity in the constraints. To linearize those constraints, we introduce variables α, α_0, β, β_0 as

$$\beta_i = w_1\theta_{1i}, \quad \beta_{0i} = b_1\theta_{1i}, \tag{11}$$

$$\alpha_i = w_2\theta_{2i}, \quad \alpha_{0i} = b_2\theta_{2i}. \tag{12}$$

Now by adding the following constraints to (9) and (10)

$$\begin{aligned}
& w_1 - M_3\theta_i \le \beta_i \le w_1 + M_3\theta_i, && i \in I^+, \\
& -M_3(1-\theta_{1i}) \le \beta_i \le M_3(1-\theta_{1i}), && i \in I^+, \\
& b_1 - M_3\theta_{1i} \le \beta_{0i} \le b_1 + M_3\theta_{1i}, && i \in I^+, \\
& -M_3(1-\theta_{1i}) \le \beta_{0i} \le M_3(1-\theta_{1i}), && i \in I^+,
\end{aligned}$$

and

$$\begin{aligned}
& w_2 - M_4(\theta_{2i}) \le \alpha_i \le w_2 + M_4(\theta_{2i}), && i \in I^-, \\
& -M_4(1-\theta_{2i}) \le \alpha_i \le M_4(1-\theta_{2i}), && i \in I^-, \\
& b_2 - M_4(\theta_{2i}) \le \alpha_{0i} \le b_2 + M_4(\theta_{2i}), && i \in I^-, \\
& -M_4(1-\theta_{2i}) \le \alpha_{0i} \le M_4(1-\theta_{2i}), && i \in I^-.
\end{aligned}$$

We obtain the following problems:

$$
\begin{aligned}
\min_{w_1,\, ,b_1,\, \eta_1,\, \eta_2,\, \beta,\, \beta_0,\, \xi_1,\, \theta_1} \quad & \tfrac{1}{2}||w_1||^2 + c_1 e_1^T(\eta_1+\eta_2) + c_2 e_3^T \xi_1 + c_3 e_1^T \theta_1 \\
\text{s.t. } & \beta_i^T x_i + \beta_{0i} \le \epsilon_1 + \eta_{1i}, && i \in I^+, \\
& -(\beta_i^T x_i + \beta_{0i}) \le \epsilon_1 + \eta_{2i}, && i \in I^+, \\
& w_1^T x_i + b_1 - (\beta_i^T x_i + \beta_{0i}) \le -1 + \xi_{1i} + M_1(1-\theta_{1i}), && i \in I^+, \\
& w_1^T x_i + b_1 \le -1 + \xi_i && i \in I^-, \\
& w_1 - M_3\theta_i \le \beta_i \le w_1 + M_3\theta_i, && i \in I^+, \\
& -M_3(1-\theta_{1i}) \le \beta_i \le M_3(1-\theta_{1i}), && i \in I^+, \\
& b_1 - M_3\theta_{1i} \le \beta_{0i} \le b_1 + M_3\theta_{1i}, && i \in I^+, \\
& -M_3(1-\theta_{1i}) \le \beta_{0i} \le M_3(1-\theta_{1i}), && i \in I^+, \\
& \beta_i \in \mathcal{R}^n,\ \beta_{0i} \in \mathcal{R}, && i \in I^+, \\
& \theta_{1i} \in \{0,1\}, && i \in I^+, \\
& w_1 \in \mathcal{R}^n,\ b_1 \in \mathcal{R}, \\
& \eta_1,\ \eta_2,\ \xi_1 \ge 0,
\end{aligned}
$$

and

$$
\begin{aligned}
\min_{w_2,\, ,b_2,\, \eta_1',\, \eta_2',\, \alpha,\, \alpha_0,\, \xi_2, \theta_2} \quad & \tfrac{1}{2}||w_2||^2 + c_4 e_2^T(\eta_1'+\eta_2') + c_5 e_3^T \xi_2 + c_6 e_2^T \theta_2 \\
\text{s.t. } & \alpha_i^T x_i + \alpha_{0i} \le \epsilon_2 + \eta_1', && i \in I^-, \\
& -(\alpha_i^T x_i + \alpha_{0i}) \le \epsilon_2 + \eta_2', && i \in I^-, \\
& w_2^T x_i + b_2 - (\alpha_i^T + \alpha_{0i}) \ge 1 - \xi_{2i} - M_2(1-\theta_{2i}), && i \in I^-, \\
& w_2^T x_i + b_2 \ge 1 - \xi_{2i}, && i \in I^+, \\
& w_2 - M_4(\theta_{2i}) \le \alpha_i \le w_2 + M_4(\theta_{2i}), && i \in I^-, \\
& -M_4(1-\theta_{2i}) \le \alpha_i \le M_4(1-\theta_{2i}), && i \in I^-, \\
& b_2 - M_4(\theta_{2i}) \le \alpha_{0i} \le b_2 + M_4(\theta_{2i}), && i \in I^-, \\
& -M_4(1-\theta_{2i}) \le \alpha_{0i} \le M_4(1-\theta_{2i}), && i \in I^-, \\
& \alpha_{2i} \in \mathcal{R}^n,\ \alpha_{0i} \in \mathcal{R}, && i \in I^-, \\
& \theta_2 \in \{0,1\}, && i \in I^-, \\
& w_2 \in \mathcal{R}^n,\ b_2 \in \mathcal{R}, \\
& \eta_1',\ \eta_2',\ \xi_2 \ge 0,
\end{aligned}
$$

where M_i, with $i = 1, \ldots, 4$, represent significant positive constants. As known, MIP models are NP-hard problems, and solving large-scale MIP problems can be computationally challenging and often requires sophisticated optimization algorithms and heuristics to find near-optimal solutions within reasonable time frames [37]. To compare the proposed MIP mod-

els with those in [6], their number of variables and constraints of models are provided in Table 1. According to this table, it becomes evident that the number of variables for each model of RENPSVM is lower than that of RESVM when $\frac{m_i}{l} \leq \frac{n+2}{n+4}$ (for $i = 1, 2$), which is the case for datasets where the number of features exceeds seven. Also, each RENPSVM MIP model has less linear constraints than RESVM when $\frac{m_i}{l} \leq \frac{n+1}{n+2}$ (for $i = 1, 2$), which is the case for datasets where $n \geq 7$. It should also be noted that despite the fact that CRESVM has less linear constraints compared to both RESVM and RENPSVM, its demerit is that is has nonlinear constraints.

Table 1: Number of variables and constraints of RESVM, CRESVM, and RENPSVM

	RESVM	CRESVM	RENPSVM (13)	RENPSVM (13)
Variables	$ln + 3l + n + 1$	$4l + 3n + 1$	$m_1n + 4m_1 + l + n + 1$	$m_2n + 4m_2 + l + n + 1$
Linear constraints	$6l + 4ln$	$5l$	$4m_1n + 9m_1 + m_2 + l$	$4m_2n + 9m_2 + m_1 + l$
Nonlinear constraints	0	$2l$	0	0

4 Computational experiments

To demonstrate the effectiveness of RENPSVM, we conducted experiments using a set of 10 UCI datasets, as detailed in Table 2. To evaluate the models' resilience to label noise, we executed three distinct experiments for each dataset. For the Vertebral dataset, two scenarios are considered. First (Vertebral1), distinguishing patients as either Normal (100) or those with Disk Hernia (60); and second (Vertebral2), categorizing patients as either Normal (100) or Abnormal, with Abnormal encompassing individuals with Disk Hernia (60) or Spondylolisthesis (150). These experiments encompassed the original datasets, along with two scenarios involving the introduction of random label flips in the training data at percentages of 20% and 50%. The implementation of all models is carried out in MATLAB 2020 (64-bit) on a computer equipped with an Intel Core i5 processor and 4 GB of RAM.

Also, RESVM, CRESVM, and RENPSVM models are solved using the CVX-Mosek [16]. For all models, the hyperparameters c_i $(i = 1, \ldots, 4)$ are chosen from the set $\{2^i|, i = -8, \ldots, 8\}$, taking into consideration their impact on the models' performance. To mitigate issues like overfitting and bias across all datasets, we employed a 10-fold cross-validation methodology. This technique partitions the dataset into ten equally sized subsets, as recommended by [14]. Subsequently, the models are trained on nine of these subsets, while the remaining subset is utilized to compute the prediction error of the models. This process is repeated for each of the ten subsets. Finally, the average classification accuracy is computed using the following formula:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN},$$

where TP, TN, FP, and FN denote the number of true positive, true negative, false positive and false negative, respectively. Computational results are summarized in Table 3.

Table 2: Characteristics of datasets

Datasets	Samples	Positive	Negative	Features	Classes
Car	1594	1210	384	7	2
Haberman	306	225	81	3	2
Cancer	699	458	241	9	2
Vertebral	310	60	100	7	3
Hayes-Ruth	102	51	51	5	2
Diabetes	768	500	268	8	2
Ionosphere	351	225	126	34	2
Votes	435	267	168	16	2
Heart	270	260	120	13	2

Table 3: Performance comparison of all models

		Flip percentage		
Datasets	Method	0.0	0.2	0.5
		Accuracy(Time)		
Car	RESVM	100(**53.40**)	79.98(**51.38**)	51.38(**56.04**)
	CRESVM	100(68.61)	78.41(82.67)	**52.01**(86.69)
	RENPSVM	100(161.89)	**79.99**(99.37)	51.25(73.93)
Haberman	RESVM	71.29(449.52)	63.36(405.02)	48.93(116.95)
	CRESVM	**75.438(13.79)**	59.57(290.36)	49.18(188.52)
	RENPSVM	72.87(104.42)	**66.025(30.28)**	**55.89 (28.32)**
Cancer	RESVM	94.85(**40.15**)	58.50(125.75)	48.93(116.95)
	CRESVM	96.28(311.55)	69.07(311.35)	48.074 (309.69)
	RENPSVM	**97**(45.92)	**77.97(42.02)**	**53.06(36.01)**
Vertebral1	RESVM	100(23.055)	77.07(108.57)	55.01 (108.75)
	CRESVM	100(114.80)	76.87(**9.23**)	55.01 (**21.57**)
	RENPSVM	100(**16.12**)	**80**(16.21)	**56.26**(30.89)

Vertebral2	RESVM	**81.29**(347.71)	68.81(366.69)	55.005(108.76)
	CRESVM	76.7742(311.88)	64.39(25.41)	51.62(96.93)
	RENPSVM	79.67(**18.47**)	**69.032(18.06)**	**55.447(18.82)**
Hayes-roth	RESVM	51.09(111.36)	50.12(107.37)	52.90(108.51)
	CRESVM	53.51(308.04)	53.92(51.91)	49.22(47.81)
	RENPSVM	**62.54(32.57)**	**62.30(14.33)**	**54.63(13.99)**
Diabet	RESVM	65.10(115.84)	55.73(114.52)	51.53(115.80)
	CRESVM	69.52(309.61)	55.75(281.17)	50.51 (305.35)
	RENPSVM	**75.65(43.79)**	**56.51(34.75)**	**51.69(36.04)**
Ionosphere	RESVM	84.88(316.98)	56.13(318.21)	51.85(316.65)
	CRESVM	82.9(312.50)	**69.22(40.40)**	45.48(**27.55**)
	RENPSVM	**85.18(68.81)**	68.39(65.44)	**53.85**(65.21)
Votes	RESVM	**95.88**(310.98)	76.99(310.41)	50.33 (319.23)
	CRESVM	94.49(299.53)	74.41(308.84)	44.17(311.22)
	RENPSVM	95.62(**35.25**)	**78.88(32.73)**	**54.28 (27.02)**

Heart	RESVM	78.89(09.23)	54.81(309.49)	53.387 (359.49)
	CRESVM	83.70(28.73)	**74.41(308.84)**	52.96 (**72.06**)
	RENPSVM	**85.18(21.81)**	72.59(326.27)	**55.93**(319.05)

According to Table 3, for the original datasets, RENPSVM exhibits superior accuracy compared to the other models across all datasets except for Votes and Vertebral2. Additionally, it demonstrates equivalent accuracy to the other models for datasets such as Car and Vertebral1. In the aspect of learning time, RENPSVM outperforms CRESVM and RESVM, except for Car. Additionally, RENPSVM secures the second-best position in terms of learning time for Haberman and Cancer datasets. When considering a label flip scenario of 20%, it becomes evident that the accuracy of RENPSVM surpasses that of the other models for all datasets except Heart and Ionosphere. However, in terms of learning time, RENPSVM generally outperforms other models, except for Car, Vertebral1, Ionosphere, and Heart. Among these, Vertebral and Ionosphere are notable as being the second-best in terms of learning time. When dealing with a label flip rate of 50%, among all the datasets, only Car does not exhibit the highest accuracy with the proposed model. Turning to learning time, the proposed model demonstrates superiority over all other models, except for Car, Vertebral1, Ionosphere, and Heart, which secure the position of being the second-best performers. By analyzing the comprehensive results presented in Table 3, it becomes evident that as the percentage of flipped labels increases, the proposed model exhibits superior accuracy compared to the referenced models and demonstrates enhanced robustness.

Next, the modified Friedman test is initially conducted to assess whether distinctions exist among the three models. Following this, the Nemenyi post-hoc test is utilized to enable the comparison of multiple methods, offering pairwise assessments between them. This post-hoc analysis assists in deter-

mining the presence of significant differences between the considered methods.

To determine whether the results of the three models in Table 3 differ or not, the modified Friedman test is conducted. The Friedman test, being a nonparametric statistical test, does not rely on assumptions about the underlying data distribution [9]. For each dataset, individual ranks are assigned to all algorithms, with the top-performing algorithm receiving rank 1, the second-best algorithm receiving rank 2, and so forth. In cases of ties, average ranks are employed. Let r_{ij} denote the rank of the jth algorithm on the ith dataset. The test examines the average rank for each algorithm, denoted as $\bar{r}j = \sum_{i=1}^{n} r_{ij}$. To account for the potential conservatism of the Friedman test, a modified version of it is calculated as outlined by [18].

$$F_f = \frac{(N-1)\mathcal{X}_F^2}{N(k-1)-\mathcal{X}_F^2}, \tag{13}$$

where $\mathcal{X}_F^2$ is equal to $\frac{12N}{k(k+1)}(\sum_{j=1}^{k}\bar{r}_j^2 - \frac{k(k+1)^2}{4})$, N represents the number of datasets, and k denotes the number of methods. Furthermore, F_f is distributed according to the F-distribution with degrees of freedom $(k-1,(k-1)(N-1))$. The average accuracy ranks corresponding to Table 3 are presented in a tabular format as shown in Table 4. The critical value at a significance level of $\alpha = 0.1$ for $F_f(2,18)$ is determined to be 3.63. Considering the average ranks (Table 4), the $\mathcal{X}_F^2$ values for scenarios with label flip percentages of 0%, 20%, and 50% are 3.8, 9.8, and 10.05, respectively. The corresponding F_f values are 2.1111, 8.6471, and 9.0905. Given that the F_f values for the 20% and 50% scenarios exceed the critical value of $F_f(3,16) = 3.63$, and considering that the rank of RENPSVM is lower than that of RESVM and CRESVM, it can be inferred that there exists a significant distinction between RENPSVM and the models introduced in [6] for these particular scenarios.

Table 4: Average accuracy rank of all models

	Method	Flip percentage 0.0	0.2	0.5
Average rank	RESVM	2.3	2.5	2.25
	CRESVM	2.2	2.3	2.55
	RENPSVM	1.5	1.2	1.2

The Nemenyi post-hoc test serves as a statistical technique utilized for the comparison of multiple methods, offering pairwise comparisons to ascertain the presence of significant distinctions. To execute this test for pairwise comparisons, we compute a parameter known as the critical difference (CD). The CD is determined by considering the number of datasets, the number of methods, the chosen significance level, and the average rank associated with each model from Table 4. When the difference in average ranks between the two methods exceeds the CD value, it can be inferred that a noteworthy and statistically significant difference exists between those two methods. The CD value is calculated as follows:

$$CD = q_{\alpha=0.1}\sqrt{\frac{k(k+1)}{6N}},$$

where the parameter q represents the critical value, while k signifies the number of models, and N denotes the number of datasets. For a significance level of 0.1 and considering four methods, the critical value extracted from the Nemenyi distribution table amounts to $q = 2.3122$. Substituting these values into the above equation yields a computed CD value of 1.0340. The difference between the average ranks of the two models is represented as Ξ. In the scenario where the label flip percentage is 0%, we encounter the following conditions:

$$\Xi(\text{RENPSVM} - \text{RESVM}) = |1.5 - 2.3| = 0.8 < CD(1.0340),$$

$$\Xi(\text{RENPSVM} - \text{CRESVM}) = |1.5 - 2.2| = 0.7 < CD(1.0340),$$

When the flip percentage of labels is 20% we have:

$$\Xi(\text{RENPSVM} - \text{RESVM}) = |1.2 - 2.5| = 1.3 > CD(1.0340),$$
$$\Xi(\text{RENPSVM} - \text{CRESVM}) = |1.2 - 2.3| = 1.1 > CD(1.0340).$$

Finally, in the case when the flip percentage of labels is 50%, we have

$$\Xi(\text{RENPSVM} - \text{RESVM}) = |1.2 - 2.25| = 1.05 > CD(1.0340),$$
$$\Xi(\text{RENPSVM} - \text{CRESVM}) = |1.2 - 2.55| = 1.35 > CD(1.0340).$$

Based on the preceding results, it is evident that a significant difference exists between RENPSVM and the other models, except for the original dataset.

5 Conclusions

In this paper, we have introduced MIP models based on NPSVM for the purpose of relabeling noisy data. Our approach effectively refines observation labels while simultaneously achieving an optimal solution. We achieve significant reductions in computational costs by strategically avoiding the consideration of all observations as potential candidates for label adjustments in the model. Instead, we concentrate on instances associated with the class that the model aims to represent within an ϵ-intensive tube. The inherent structure of NPSVM allows for parallel execution of the proposed MIP models, resulting in accelerated learning times across the majority of datasets. Our findings indicate that, for datasets with a number of features exceeding seven, each RENPSVM MIP model has fewer linear constraints and variables compared to RESVM, subject to specific conditions. This holds true for the majority of datasets. Additionally, the CRESVM model also exhibits fewer linear constraints and variables compared to both RESVM and RENPSVM, although it introduces the trade-off of incorporating nonlinear constraints. The effectiveness of our proposed models is evaluated through experiments conducted on 10 UCI datasets. The outcomes showcased that RENPSVM models exhibit better performance in terms of classification accuracy and

learning, akin to RESVM and CRESVM, respectively, for most datasets and as the percentage of flipped labels increases, the proposed RENPSVM model demonstrates superior accuracy compared to the referenced models and showcases enhanced robustness. Moreover, we employed the modified Friedman test and Nemenyi post-hoc test to assess the influence of label noise on our model's performance relative to other methods. The tests revealed that a notable distinction between RENPSVM and other models exists, except for the original datasets. For future work, one may consider extending the proposed model to multi-class classification, either through a one-vs-one-vs-rest approach [29] or by adapting it into a regression model. This adaptation can be particularly useful for handling label noise in target values, a common challenge in regression tasks [30]. Also, when dealing with datasets containing a large number of features, the computational cost can become prohibitively high. In such cases, it is efficient to derive hyperplane classifiers using the dual problem formulation [10]. Therefore, studying label noise using dual models might be another interesting future research direction.

6 Acknowledgements

The authors would like to thank the editor and reviewers for their useful comments and suggestions.

References

[1] Abdi, A., Nabi, R.M., Sardasht, M. and Mahmood R. *Multiclass classifiers for stock price prediction: A comparison study*, J. Harbin Inst. Technol. 54(3) (2022), 32–39.

[2] Angluin, D. and Laird, P. *Learning from noisy examples*, Mach. Learn. 2 (1988), 343–370.

[3] Bertsimas, D., Dunn, J., Pawlowski, C. and Zhuo, Y.D. *Robust classification*, INFORMS Journal on Optimization 1(1) (2019), 2–34.

[4] Biggio, B., Nelson, B. and Laskov, P. *Support vector machines under adversarial label noise*, Asian conference on machine learning (2011), 97–112.

[5] Blanco, V., Japón, A. and Puerto, J. *Robust optimal classification trees under noisy labels*, Adv. Data Anal. Classif. 16(1) (2022), 155–179.

[6] Blanco, V., Japón, A. and Puerto, J. *A mathematical programming approach to SVM-based classification with label noise*, Comput. Ind. Eng. 172 (2022), 108611.

[7] Chen, Z., Song, A., Wang, Y., Huang, X. and Kong, Y. *A noise rate estimation method for image classification with label noise*, J. Phys. Conf. Ser. IOP Publishing 2433(1) (2023), 012039.

[8] Cortes, C. and Vapnik, V.N. *Support vector networks*, Mach. Learn. 20(3) (1995), 273–297.

[9] Demsar, J. *Statistical comparisons of classifiers over multiple data sets*, J. Mach. Learn Res. 7 (2006), 1–30.

[10] Deng, N., Tian, Y. and Zhang, C. *Support vector machines: Optimization based theory, algorithms, and extensions*, CRC press, 2012.

[11] Ding, S., Zhang, N., Zhang, X. and Wu, F. *Twin support vector machine: Theory, algorithm and applications*, Neural Comput. Appl. 28(11) (2017), 3119–3130.

[12] Ding, S., Zhao, X., Zhang, J., Zhang, X. and Xue, Y. *A review on multi-class TWSVM*, Artif. Intell. Rev. 52(2) (2019), 775–801.

[13] Duan, Y. and Wu, O. *Learning with auxiliary less-noisy labels*, IEEE Trans. Neural Netw. Learn. Syst. 28(7) (2018), 1716–1721.

[14] Duda, R.O., Hart, P.E. and Stork, D.G. *Pattern Classification*, John Wiley & Sons, 2012.

[15] Ekambaram, R., Fefilatyev, S., Shreve, M., Kramer, K., Hall, L.O. and Goldgof, D.B. *Active cleaning of label noise*, Pattern Recognit. 51 (2016), 463–480.

[16] Grant, M., Boyd, S. and Ye, Y. *Cvx: Matlab software for disciplined convex programming*, version 2.0 beta, 2013.

[17] Hassani, S.F., Eskandari, S. and Salahi, M. *CInf-FS: An efficient infinite feature selection method using K-means clustering to partition large feature spaces*, Pattern Anal. Appl. (2023), 1–9.

[18] Iman, R.L. and Davenport, J.M. *Approximations of the critical region of the fbietkan statistic*, Commun. Stat. Theory Methods 9 (6) (1980), 571–595.

[19] Jayadeva, Khemchandani, R. and Chandra, S. *Twin support vector machines for pattern classification*, Trans. Pattern Anal. Mach. Intell. 29(5) (2007), 905–910.

[20] Jimenez-Castano, C., Alvarez-Meza, A. and Orozco-Gutierrez, A. *Enhanced automatic twin support vector machine for imbalanced data classification*, Pattern Recognit. 107 (2020), 107442.

[21] Keerthi, S.S., Shevade, S.K., Bhattacharyya, C. and Murthy, K.R.K. *Improvements to platt's SMO algorithm for SVM classifier design*, Neural Comput. 13(3) (2001), 637–649.

[22] Kshirsagar, A.P. and Shakkeera, L. *Recognizing Abnormal Activity Using MultiClass SVM Classification Approach in Tele-health Care*, IOT with Smart Systems: Proceedings of ICTIS 2021, Springer Singapore, 2 (2022), 739–750.

[23] Lachenbruch, P.A. *Discriminant analysis when the initial samples are misclassified*, Technometrics 8(4) (1966), 657–662.

[24] Lachenbruch, P.A. *Note on initial misclassification effects on the quadratic discriminant function*, Technometrics 21(1) (1979), 129–132.

[25] McLachlan, G.J. *Asymptotic results for discriminant analysis when the initial samples are misclassified*, Technometrics 14(2) (1972), 415–422.

[26] Nasiri, J.A. and Mir, A.M. *An enhanced KNN-based twin support vector machine with stable learning rules*, Neural Comput. Appl. 16 (2020), 12949–12969.

[27] Okamoto, S. and Yugami, N. *An average-case analysis of the k-nearest neighbor classifier for noisy domains*, 15th International Joint Conference on Artificial Intelligence (IJCAI) (1997), 238–245.

[28] Platt, J. *Fast Training of Support Vector Machines using Sequential Minimal Optimization*, MIT Press, 1998.

[29] Sahleh, A., Salahi, M. and Eskandari, S. *Multi-class nonparallel support vector machine*, Prog. Artif. Intell. (2023), 1–15.

[30] Tanveer, M., Rajani, T., Rastogi, R., Shao, Y.H. and Ganaie, M.A. *Comprehensive review on twin support vector machines*, Ann. Oper. Res. (2022), 1–46.

[31] Thulasidasan, S., Bhattacharya, T., Bilmes, J., Chennupati, G., and Mohd-Yusof, J. *Combating label noise in deep learning using abstention*, arXiv preprint arXiv, (2019), 1905.10964.

[32] Tian, Y. and Qi, Z. *Review on: twin support vector machines*, Ann. Data Sci. 1 (2014), 253–277.

[33] Tian,Y., Qi, Z., Ju, X., Shi, Y. and Liu, X. *Nonparallel support vector machines for pattern classification*, IEEE Trans. Cybern. 44(7) (2014), 1067–1079.

[34] Vapnik, V.N. *The Nature of Statistical Learning Theory*, Springer, New York, 1996.

[35] Vapnik, V.N. *Statistical Learning Theory*, John Wiley & Sons, New York, 1998.

[36] Witoonchart, P. and Chongstitvatana, P. *Application of structured support vector machine backpropagation to a convolutional neural network for human pose estimation*, Neural Networks 92 (2017), 39–46.

[37] Wolsey, L.A. and Nemhauser, G.L. *Integer and combinatorial optimization*, John Wiley & Sons, 55, 1999.

[38] Xiao, H., Biggio, B., Nelson, B., Xiao, H., Eckert, C. and Roli, F. *Support vector machines under adversarial label contamination*, Neurocomputing 160 (2015), 53–62.

Iranian Journal of Numerical Analysis and Optimization
Vol. 14, No. 1, 2024, pp 291–314
https://doi.org/10.22067/ijnao.2023.82958.1282
https://ijnao.um.ac.ir/

CC BY NC

Research Article

On generalized one-step derivative-free iterative family for evaluating multiple roots

H. Arora, A. Cordero* and J. R. Torregrosa

Abstract

In this study, we propose a family of iterative procedures with no derivatives for calculating multiple roots of one-variable nonlinear equations. We also present an iterative technique to approximate the multiplicity of the roots. The new class is optimal since it fits the Kung–Traub hypothesis and has second-order convergence. Derivative-free methods for calculating multiple roots are rarely found in literature, especially in the case of one-step

*Corresponding author

Received 15 June 2023; revised 6 November 2023; accepted 8 November 2023

Himani Arora
Department of Mathematics, Guru Nanak Dev University, 143005, Amritsar, India. e-mail: arorahimani362@gmail.com

Alicia Cordero
Instituto de Matemàtica Multidisciplinar, Universitat Politècnica de València, 46022, Valencia, Spain. e-mail: acordero@mat.upv.es

Juan R. Torregrosa
Instituto de Matemàtica Multidisciplinar, Universitat Politècnica de València, 46022, Valencia, Spain. e-mail: jrtorre@mat.upv.es

How to cite this article
Arora, H., Cordero, A. and Torregrosa, J. R., On generalized one-step derivative-free iterative family for evaluating multiple roots. *Iran. J. Numer. Anal. Optim.*, 2024; 14(1): 291-314. https://doi.org/10.22067/ijnao.2023.82958.1282

methods, which are the simplest ones in terms of their structure. Moreover, this new family contains almost all the existing single-step derivative-free iterative schemes as its special cases, with an additional degree of freedom. Several results are used to confirm its theoretical order of convergence. Through the complex discrete dynamics analysis, the stability of the suggested class is illustrated, and the most stable methods are found. Several test problems are included to check the performance of the proposed methods, whether the multiplicity of the roots is estimated or known, comparing the numerical results with those obtained by other methods.

AMS subject classifications (2020): Primary 65H05; Secondary 37D99.

Keywords: Nonlinear equations; Derivative–free iterative method; Multiple roots; Order of convergence; Stability.

1 Introduction

One well-known derivative-free scheme for obtaining a multiple root x^* of a nonlinear complex-valued function $f(x)$, $x \in \mathbb{C}$ is the modified Traub–Steffensen's approach [19]. The structure of this scheme is as follows:

$$x_{t+1} = x_t - m\frac{f(x_t)}{f[\eta_t, x_t]}, \quad t = 0, 1, 2, \ldots, \tag{1}$$

where m denotes the known multiplicity of x^* and $f[\eta_t, x_t] = \frac{f(\eta_t)-f(x_t)}{\eta_t - x_t}$ is the first-order divided difference at point x_t and $\eta_t = x_t + \beta f(x_t)$, $\beta \neq 0$, $\beta \in \mathbb{R}$. For $m \geq 2$, the order of convergence of this scheme is quadratic. Scheme (1) is the derivative-free version of the familiar modified Newton's method [15] as it is obtained from this method by estimating the derivative $f'(x_t)$ by $f[\eta_t, x_t]$. Centered on these two schemes, many methods [6, 13, 16, 18, 22, 11, 17, 21, 7, 10] have been derived in earlier studies. Among these, the second-order schemes presented by Kumar, Sharma, and Argyros [9] and Kansal et al. [8], are the nicest ones as they possess the following features: (i) the schemes are derivative-free, (ii) they are optimal in terms of the Kung–Traub conjecture (see [10]), and (iii) the schemes are single-step, which make them easy to implement. The general type of optimal derivative-

free procedure for multiple roots defined by Kumar, Sharma, and Argyros [9] is given by

$$x_{t+1} = x_t - H(\Phi_t), \tag{2}$$

where $\Phi_t = \frac{f(x_t)}{f[\eta_t, x_t]}$ and $H(\Phi_t)$ is a single variable weight function. Other main aspect of this scheme is that it holds Traub–Steffensen's method (1) as a special case.

The scheme from Kansal et al. [8] has the following structure:

$$x_{t+1} = x_t - m\frac{(1-a)f(\eta_t) + af(x_t)}{f[\eta_t, x_t]}, \tag{3}$$

where $a \in \mathbb{R}$ is a free parameter. For $a = 1$, this scheme reduces to Traub–Stefensen's method (1).

Motivated by these multiple root-finding schemes, we aim to design a new derivative-free iterative family that is more general than class (2). Therefore, the new class is a single-step optimal scheme with second-order convergence, and it has an additional degree of freedom. In addition, we include an iterative process that allows us to estimate the multiplicity when it is not known.

In order to choose the most stable members of the proposed class, we now introduce some dynamic terms used in this manuscript (see, for instance, [2]). To achieve this aim, the key concept is the rational function. Usually, a rational function T is found by applying the proposed class on a low-degree polynomial $p(z)$ with multiple roots. The properties of this rational function, the existence of fixed points or periodic orbits, and their repulse or attracting character, among others, allow us to detect the best elements of the class or to compare known schemes (see, for example, [4, 5]).

Let us consider a rational function $T : \widehat{\mathbb{C}} \longrightarrow \widehat{\mathbb{C}}$, being $\widehat{\mathbb{C}}$ the Riemann sphere. Then, we define the orbit of a point z as follows:

$$\{z, T(z), T^2(z), \ldots, T^n(z), \ldots\},$$

where T^k denotes the kth composition of T with itself.

We analyze T by characterizing the starting points from the asymptotical performance of their orbits. In these terms, a point z is a fixed point of T if $T(z) = z$; it is a p-periodic point, $p > 1$, if $T^p(z) = z$ and $T^k(z) \neq z$ for $k < p$.

The basin of an attractor ξ is defined as

$$\mathcal{A}(\xi) = \{z \in \widehat{\mathbb{C}} : T^n(z) \to \xi, n \to \infty\}. \tag{4}$$

The Fatou set of T, is the set of points whose orbits tend to an attractor (fixed point, periodic point, or infinity). It's complementary in $\widehat{\mathbb{C}}$ is the Julia set. So, the basin of attraction of any fixed point belongs to the Fatou set and the boundaries of these basins of attraction belong to the Julia set.

On the other hand, a point z is called a critical point of T if $T'(z) = 0$. The asymptotic behavior of the critical points is a key fact for analyzing the stability of the method: A classical result from Fatou and Julia (see, for instance, [1]). Each immediate basin of attraction holds at least one critical point; that is, in the connected component of the basin of attraction holding the attractor, there is also a critical point.

Moreover, a fixed point of T, z, is called attracting if $|T'(z)| < 1$, or super attracting if $|T'(z)| = 0$; it is repulsive if $|T'(z)| > 1$; and parabolic if $|T'(z)| = 1$.

Let us also remark that when fixed and critical points are not equivalent to the roots of $p(z)$, then they are called strange fixed and free critical points, respectively. The asymptotical performance of free critical points is a key element in this study, as when they do not belong to the basin of attraction of the searched zeros. Then, they lay in the basin of an attracting element (a periodic orbit or a strange fixed point) that should be avoided in practice.

Indeed, when T also depends on one or several parameters r_i, $i = 1, 2, \ldots, k$, and the stability of a fixed point z is analyzed, $|T'(z, r_1, \ldots, r_k)|$ is not a scalar but a function of r_i for $i = 1, 2, \ldots, k$. In this case, $|T'(z, r_1, \ldots, r_k)|$ is called a stability function of the fixed point, and it gives us the character of the fixed point in terms of the value of r_i, $i = 1, 2, \ldots, k$.

Let us summarize the rest of the paper: Section 2 includes the construction and convergence analysis of the new family. Also, in Section 2, special cases of the proposed class are found. To examine the stability of the new methods, we present the analysis of the discrete dynamical system associated with the different sub-classes defined in Section 3, selecting the most stable ones. Section 4 is devoted to the application of proposed methods on a num-

ber of numerical problems to back up the theoretical conclusions. In Section 5, we conclude with some final observations.

2 Establishment of second-order class

Let us define a new derivative-free Traub–Steffensen-type class of iterative schemes for finding multiple roots as

$$x_{t+1} = x_t - L(\Theta_t), \tag{5}$$

where $L(\Theta_t)$ is a single-variable weight function, $\Theta_t = \frac{f(x_t)}{f[\eta_t, x_t] + \gamma f(x_t)}$, $\gamma \in \mathbb{R}$, and $\eta_t = x_t + \beta f(x_t)$, $\beta \neq 0$, $\beta \in \mathbb{R}$.

We show in the following result that the proposed scheme (5) achieves maximum second-order convergence for every $\beta \neq 0$, without using any derivative evaluation. Moreover, we show that the conditions to be imposed on the weight function are directly related to the multiplicity of the root.

Theorem 1. Let $f : \mathbb{D} \subset \mathbb{C} \to \mathbb{C}$ be an analytic function in the open set $\mathbb{D}$, that holds a multiple root x^*, with multiplicity m. Then, the class of iterative schemes (5) has second-order of convergence, if

$$L(0) = 0, \quad L'(0) = m, \quad |L''(0)| < \infty, \tag{6}$$

and therefore it satisfies the error equation

$$e_{t+1} = \frac{1}{2m^2} \left(2m(\gamma + B_1) - L''(0)\right) e_t^2 + O(e_t^3),$$

where $e_t = x_t - x^*$ and $B_1 = \dfrac{1}{m+1} \dfrac{f^{(m+1)}(x^*)}{f^{(m)}(x^*)}$.

Proof. Along Taylor's series expansion of $f(x_t)$ and $f(\eta_t)$ around x^* that has multiplicity m, it is assumed that $f(x^*) = f'(x^*) = f''(x^*) = \cdots = f^{(m-1)}(x^*) = 0$ and $f^{(m)}(x^*) \neq 0$. Also, let us denote B_i by

$$B_i = \frac{m!}{(m+i)!} \frac{f^{(m+i)}(x^*)}{f^{(m)}(x^*)},$$

for $i = 1, 2, 3, \ldots$. Then,

$$f(x_t) = \frac{f^{(m)}(x^*)}{m!} e_t^m \Big(1 + B_1 e_t + B_2 e_t^2 + B_3 e_t^3 + O(e_t^4) \Big), \tag{7}$$

$$f(\eta_t) = \frac{f^{(m)}(x^*)}{m!} e_{\eta_t}^m \Big(1 + B_1 e_{\eta_t} + B_2 e_{\eta_t}^2 + B_3 e_{\eta_t}^3 + O(e_{\eta_t}^4) \Big), \tag{8}$$

where

$$e_{\eta_t} = \eta_t - x^* = e_t + \beta \frac{f^{(m)}(x^*)}{m!} e_t^m \Big(1 + B_1 e_t + B_2 e_t^2 + B_3 e_t^3 + O(e_t^4) \Big).$$

By replacing (7)–(8) in the definition of Θ_t, we obtain

$$\Theta_t = \frac{e_t}{m} - \frac{1}{m^2}(\gamma + B_1) e_t^2 + O(e_t^3). \tag{9}$$

From (9), it can be inferred that Θ_t tends to zero when t tends to infinity. Then, by using MacLaurin's theorem on the weight function $L(\Theta_t)$, we obtain

$$L(\Theta_t) = L(0) + L'(0)\Theta_t + \frac{1}{2!} L''(0) \Theta_t^2 + O(\Theta_t^3). \tag{10}$$

By imposing (9) and (10) in (5), we get

$$e_{t+1} = -L(0) + \left(1 - \frac{L^{(1)}(0)}{m} \right) e_t + \frac{1}{2m^2} \Big(2(\gamma + B_1) L^{(1)}(0) - L^{(2)}(0) \Big) e_t^2 + O(e_t^3). \tag{11}$$

From this error equation, we deduce that schemes belonging to class (5) possess, at least, second-order of convergence, provided that $L(0) = 0$ and $1 - \frac{L'(0)}{m} = 0$, that is, $L'(0) = m$.

Finally, we find the error equation

$$e_{t+1} = \frac{1}{2m^2} \Big(2m(\gamma + B_1) - L^{(2)}(0) \Big) e_t^2 + O(e_t^3), \tag{12}$$

provided $|L''(0)| < \infty$. Hence, the proposed class (5) converges to the m-multiple root x^* with second-order convergence. □

Remark 1. It is known that $E = \rho^{\frac{1}{\psi}}$ is the computational efficiency in which ρ is the order of convergence of the method under consideration and ψ is the number of functional evaluations per iteration; see [14]. The E-value corresponding to the newly proposed scheme is $E = 2^{\frac{1}{2}} \approx 1.414$, which is the same as that of modified Newton's method and modified Traub–Steffensen's method.

Remark 2. For $\gamma = 0$, the proposed class (5) reduces to scheme (2), and in addition, if we also have $L(\Theta_t) = m\Theta_t$, then we get the modified Traub–Steffensen's method (1).

2.1 Special cases

In this section, we show that the number of special cases for the new presented family is equal to the particular forms of the function $L(\Theta_t)$, fulfilling the convergence conditions of Theorem 1. A few of them are listed in Table 1.

Table 1: Some special cases of class (5)

Method	Weight function	Subcases
TM-1	$L(\Theta_t) = m\Theta_t(1 + a\Theta_t)$	$a = \frac{1}{10}$, $x_{t+1} = x_t - m\Theta_t(1 + \frac{\Theta_t}{10})$
TM-2	$L(\Theta_t) = \frac{m\Theta_t}{1+a\Theta_t}$	$a = \frac{-1}{2}$, $x_{t+1} = x_t - \frac{2m\Theta_t}{2-\Theta_t}$
TM-3	$L(\Theta_t) = \frac{m\Theta_t}{1+am\Theta_t}$	$a = \frac{-1}{2}$, $x_{t+1} = x_t - \frac{2m\Theta_t}{2-m\Theta_t}$
TM-4	$L(\Theta_t) = \frac{m\Theta_t}{(1+a\sqrt{m\Theta_t})^2}$	$a = \frac{-1}{2}$, $x_{t+1} = x_t - \frac{4m\Theta_t}{(2-\sqrt{m\Theta_t})^2}$
TM-5	$L(\Theta_t) = \frac{m(\Theta_t^2+\Theta_t)}{1+am\Theta_t}$	$a = \frac{1}{5}$, $x_{t+1} = x_t - \frac{5m(\Theta_t^2+\Theta_t)}{5-m\Theta_t}$

3 Dynamical analysis

In order to arrange this analysis, we apply our proposed sub-families TM-1 to TM-5 on the nonlinear function $p(z) = (z-1)^m(z+1)$, with $z = 1$ as a multiple root of multiplicity m and a simple root at $z = -1$. In all cases, $\gamma = 1$ and $\beta = \dfrac{1}{10}$; qualitatively, similar results are found for other values of γ and β, but the values of β close to zero make the schemes more stable, due to the estimation of the derivative provided by the first-order divided difference is better. This is the most simple nonlinear function containing two roots, one simple and one m-multiple. Although the results cannot be directly extrapolated to any nonlinear function, several analyses on different nonlinear problems confirm, in the numerical section, these results.

In what follows, we analyze the rational functions related to the particular cases of (5) presented: TM-1 to TM-5.

3.1 Fixed points and stability of rational TM cases

When the classes TM-1, TM-2, TM-3, TM-4, and TM-5 are applied on the polynomial $p(z)$ with multiplicity $m = 2$, rational functions $T_i(z,a)$, $i = 1,2,\ldots,5$ are, respectively, obtained. They all depend on the parameter of the class a, considered a free complex value. These rational operators are

$$T_1(z,a) = \frac{N_1(z,a)}{\left(z^5 - z^4 + 28z^3 + 92z^2 + 271z + 9\right)^2},$$

being $N_1(z,a) = -20000a\left(z^2-1\right)^2 + z^{11} - 2z^{10} + 57z^9 + 128z^8 + 942z^7 + 4828z^6 + 18222z^5 + 31768z^4 + 26497z^3 + 21478z^2 + 54281z + 1800$,

$$T_2(z,a) = \frac{4(25a+23)z^3 + (9-100a)z + z^6 - z^5 + 28z^4 + 71z^2 + 200}{100a\left(z^2-1\right) + z^5 - z^4 + 28z^3 + 92z^2 + 271z + 9},$$

$$T_3(z,a) = \frac{N_3(z,a)}{\left(z^5 - z^4 + 28z^3 + 92z^2 + 271z + 9\right)^2},$$

where $N_3(z,a) = -40000a\left(z^2-1\right)^2 + z^{11} - 2z^{10} + 57z^9 + 128z^8 + 942z^7 + 4828z^6 + 18222z^5 + 31768z^4 + 26497z^3 + 21478z^2 + 54281z + 1800$,

$$T_4(z,a) = \frac{N_4(z,a)}{\left(100\sqrt{2}a\left(x^2-1\right) + x^5 - x^4 + 28x^3 + 92x^2 + 271x + 9\right)^2},$$

being $N_4(z,a) = x^{11} - 2x^{10} + 57x^9 + 8(25\sqrt{2}a+16)x^8 + (942 - 200\sqrt{2}a)x^7 + (5400\sqrt{2}a+4828)x^6 + 2(10000a^2+9300\sqrt{2}a+9111)x^5 + 8(6075\sqrt{2}a+3971)x^4 + (-40000a^2 - 16600\sqrt{2}a + 26497)x^3 + (21478 - 54200\sqrt{2}a)x^2 + (20000a^2 - 1800\sqrt{2}a + 54281)x + 1800$, and

$$T_5(z,a) = \frac{N_5(z,a)}{D_5(z,a)},$$

where $N_5(z,a) = -18200 + (54281 - 1800a)z + (61478 - 54200a)z^2 + (26497 - 16600a)z^3 + 8(1471+6075a)z^4 + 6(3037+3100a)z^5 + (4828+5400a)z^6 + (942 -$

$200a)z^7+8(16+25a)z^8+57z^9-2z^{10}+z^{11}$ and $D_5(z,a)=(9+271z+92z^2+28z^3-z^4+z^5)(9+271z+92z^2+28z^3-z^4+z^5+200a(-1+z^2))$.

By solving each equation $T_i(z,a)=z$, $i=1,2,\ldots,5$, the fixed points of the ith rational function are obtained. Two of them are the roots of polynomial $p(z)$; the rest of them, if they exist, are strange fixed points. The asymptotical behavior of all the fixed points (both multiple and simple, strange or not) plays a key role in the stability of the iterative methods involved, as the convergence to fixed points different from the roots means an important drawback for an iterative method; so, we proceed below with this analysis.

In order to study the stability of the fixed points z^F of $T_i(z,a)$, we calculate its stability function, $\left|T_i'(z^F,a)\right|$. This function gives us information about the asymptotical behavior of the point in terms of the value of a. In general, the stability of other fixed points than the multiple roots of $p(z)$ depends on the value of parameter a. From these rational functions, the following result can be stated.

Theorem 2. Rational function $T_i(z,a)$ has $z=1$ and $z=-1$ as superattracting and parabolic fixed points, respectively. There can also exist strange fixed points, denoted by $r_i^j(a)$, where the index $i=1,2,\ldots,5$ corresponds to the rational operator $T_i(z,a)$ and the index $j=1,2,\ldots,5$ corresponds to the fixed point as the root of fifth-degree polynomial $q_i(t)$ described below:

$i=1$**:** Roots of polynomial $q_1(t)=9-100a+271t+(92+100a)t^2+28t^3-t^4+t^5$ are the strange fixed points of $T_1(z,a)$.

$i=2$**:** There are no strange fixed points of $T_2(z,a)$.

$i=3$**:** When $T_3(z,a)$ is analyzed, the strange fixed points are the roots of $q_3(t)=9-200a+271t+(92+200a)t^2+28t^3-t^4+t^5$.

$i=4$**:** $T_4(z,a)$ has as strange fixed points the roots of $q_4(t)=9+271t+92t^2+28t^3-t^4+t^5$.

$i=5$**:** Strange fixed points of $T_5(z,a)$ are the roots of $q_5(t)=-91+271t+192t^2+28t^3-t^4+t^5$.

Proof. By solving the equation $T_i(z,a) = z$, $i = 1,2,3,5$, it is found that, in the case when $i = 1$,

$$T_1(z,a) = z \Leftrightarrow N_1(z,a) = z\left(z^5 - z^4 + 28z^3 + 92z^2 + 271z + 9\right)^2,$$

or, equivalently,

$$(z^2-1)(9+271z+92z^2+28z^3-z^4+z^5+100a1(z^2-1)) = 0 \Leftrightarrow (z^2-1)q_1(z) = 0.$$

For $i = 2$, it can be checked that

$$T_2(z,a) = z \Leftrightarrow z^1 - 1 = 0.$$

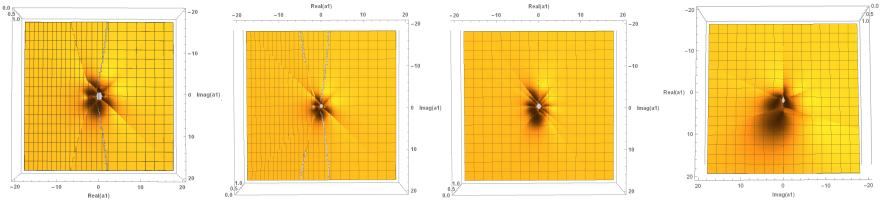
In a similar way, the statements for $i = 3$, $i = 4$, and $i = 5$ are obtained. □

So, it has been proven that the multiple root is always a superattracting fixed point, whereas the simple root is parabolic. Hence, it always lays in the Julia set and would never be reached exactly by the iterative method (although it can be very close). For $i = 1,3,4,5$, there exist strange fixed points of their respective operators, $T_i(z,a)$. Regarding the stability of these fixed points, the following remark can be checked.

Remark 3. Operators $T_i(z,a)$ for $i = 1,3,4,5$, have attracting strange fixed point at any complex value of a, except in a disk $|a| < r$ with

- $r \approx 1.12$ for $T_1(z,a)$;
- $r \approx 0.56$ for $T_3(z,a)$;
- $r \approx 0.75$ for $T_4(z,a)$;
- $r \approx 0.11$ for $T_5(z,a)$.

These small regions, where all the strange fixed points are repulsive, can be observed in Figure 1. In it, the code color is as follows: One point of the mesh is represented in orange color if there exists at least one strange fixed point that is attracting this value of the parameter. So, if one point of the mesh is not colored, it means that for this value of a, all the strange fixed points are repulsive.



(a) $|T_1'(r_1^j, a)|$ (b) $|T_3'(r_3^j, a)|$ (c) $|T_4'(r_4^j, a)|$ (d) $|T_5'(r_5^j, a)|$

Figure 1: Union of stability functions for rational operators on strange fixed points $r_i^j(a)$, $i = 1, 3, 4, 5$ and $j = 1, 2, \ldots, 5$.

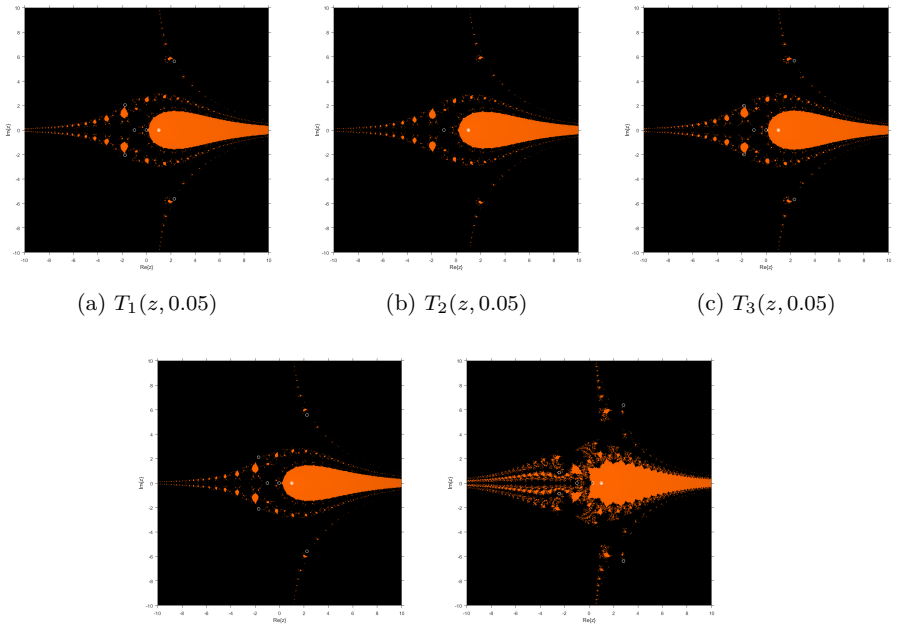
3.2 Dynamical planes

Each value of the parameter corresponds with a member of each subclass of iterative methods. When a is fixed in a rational function $T_i(z, a)$, $i = 1, 2, \ldots, 5$, the performance of the iterative process can be visualized in a dynamical plane. It is obtained by iterating the chosen element of the family under study and by using each point of the complex plane as an initial estimation. In this section, we have used a mesh of 1000×1000 points.

We represent in blue color those points whose orbit converges to infinity, in orange the points converging to the multiple root $z = 1$ (with a tolerance of 10^{-3}), and in other colors (green, red, etc.) those points whose orbit converges to one of the fixed points (all fixed points appear marked as a white star in the figures if they are attractive or by a white circle if they are repulsive or parabolic). Moreover, a point appears in black if it reaches the maximum number of 500 iterations without converging to any of the fixed points. The routines used appear in [3].

Firstly, we plot the dynamical planes related to rational functions $T_i(z, a)$, $i = 1, 2, 3, 5$ for $a = 0.05$. These planes can be visualized in Figure 2. From Theorem 2 and Remark 3, only the multiple root of $p(z)$ is an attracting fixed point.

From the dynamical planes of Figure 2, it is again deduced that the only basin of attraction is that of the multiple roots; moreover, if there exist strange fixed points, they lay in the Julia set, as per their repulsive character. Black areas correspond to divergent performance. The wideness of the basins



(a) $T_1(z, 0.05)$ (b) $T_2(z, 0.05)$ (c) $T_3(z, 0.05)$

(d) $T_4(z, 0.05)$ (e) $T_5(z, 0.05)$

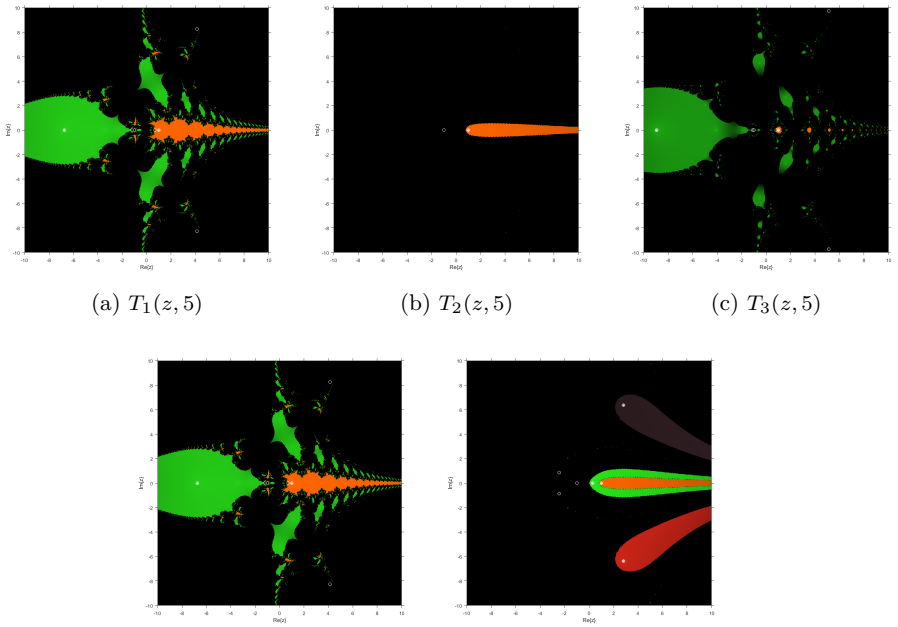
Figure 2: Dynamical planes corresponding to rational operators under study for $a = 0.05$.

of attraction composing the Fatou set is similar in all cases but slightly better in the case of $T_5(z, 0.05)$.

In Figure 3, the dynamical planes corresponding to $a = 5$ are presented. For this value of the parameter, only T_2 operator does not attract strange fixed points, from Theorem 2 and Remark 3.

We observe in Figure 3e that $T_5(z, 5)$ has three attracting strange fixed points, with their corresponding basins of attraction in green, red, and dark aubergine color, respectively. Also, $T_1(z, 5)$, $T_4(z, 5)$, and $T_3(z, 5)$ show green basins of attraction of strange fixed points. The latter case is especially inefficient, as the basin of attraction of the multiple roots is really small. The basin of attraction of the multiple roots in the case of the operator $T_2(z, 5)$ is the only one in the dynamical plane, although it is narrower than in the case of $a = 0.05$.

Figure 4 corresponds to the dynamical planes of rational functions $T_i(z, -1)$, $i = 1, 2, \ldots, 5$. Cases $i = 3$, $i = 4$, and $i = 5$ show basins of attraction of



(a) $T_1(z,5)$ (b) $T_2(z,5)$ (c) $T_3(z,5)$

(d) $T_4(z,5)$ (e) $T_5(z,5)$

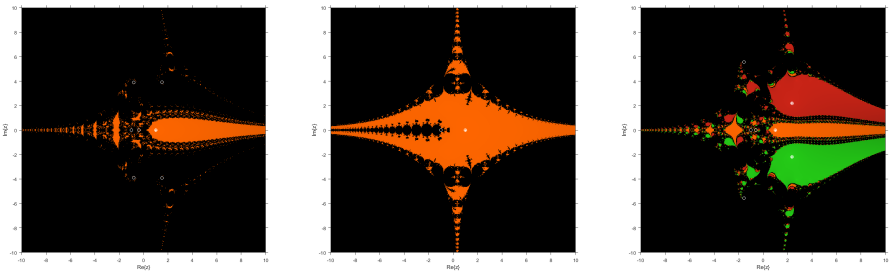
Figure 3: Dynamical planes corresponding to rational operators under study for $a = 5$.

strange fixed points. Meanwhile cases $i = 1$ and $i = 2$ only converge to the multiple roots, being the last one the widest.

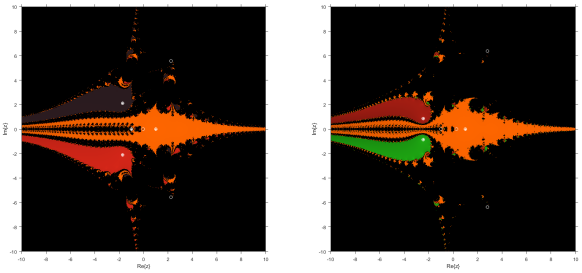
We conclude from the analysis made that the most subclass is $T_2(z, a)$, although the rest of the operators perform adequately for small absolute values of a.

4 Numerical results

In this section, we use numerical examples to confirm the performance of the newly created iterative approaches as well as to validate the theoretical results presented in the previous sections. Furthermore, we compare our new proposed methods TM-1–TM-5 with the second-order methods given by equations (1) and (3). We denote these methods by MNM and KM, respectively. For the scheme (3), we choose the parameter $a = 6/7$. All the



(a) $T_1(z,-1)$ (b) $T_2(z,-1)$ (c) $T_3(z,-1)$



(d) $T_4(z,-1)$ (e) $T_5(z,-1)$

Figure 4: Dynamical planes corresponding to rational operators under study for $a=-1$.

computations for the above-mentioned methods are performed using software Mathematica with 4096 significant digits as multiple precision arithmetic. In numerical results, we have calculated the following terms:

- The number of iterations t required to obtain the desired root fulfilling the condition that $|x_{t+1}-x_t|+|f(x_{t+1})|<10^{-50}$.
- Estimated error $e_{t+1}=|x_{t+1}-x_t|$ in the last three iterations.
- Residual error for the considered function, that is, $|f(x_{t+1})|$.
- The computational order of convergence COC [20], which is given by

$$COC=\frac{\log|\frac{x_{t+2}-x^*}{x_{t+1}-x^*}|}{\log|\frac{x_{t+1}-x^*}{x_t-x^*}|},\quad t=1,2,\ldots. \tag{13}$$

The problems used for validating the theoretical results and comparison are mentioned in Table 2. The numerical results obtained for these problems

have been presented in Tables 3–6. From them, we see that the new proposed methods are, at least, as stable as those existing methods used for comparison. Also, the presented results have verified the theoretical results proven so far. Moreover, from Tables 3–6, it is observed that methods TM-2, TM-4, and TM-5 have performed rightly well in all the problems. Furthermore, the performance of the methods TM-1 and TM-3 is either at par or better than the results obtained for the existing methods.

Table 2: Test functions considered, their corresponding roots and multiplicity

Test Functions	x_0	*Root*	m
$f_1(x) = x^3 - 5.22x^2 + 9.0825x - 5.2675$	1.8	$1.75000\ldots$	2
$f_2(x) = \left(\tan^{-1}(\frac{1}{x} - 1) + e^{x^2} - 3\right)^2$	1	$1.05653\ldots$	2
$f_3(x) = (\sin x \cos x - x^3 + 1)^9$	1.4	$1.11707\ldots$	9
$f_4(x) = ((x-1)^3 - 1)^{50}$	2.1	$2.00000\ldots$	50

Table 3: Numerical results for $f_1(x)$

Methods	t	$\lvert e_{t-1}\rvert$	$\lvert e_t\rvert$	$\lvert e_{t+1}\rvert$	$\lvert f(x_{t+1})\rvert$	*COC*
TM-1 ($\gamma = 0$)	8	5.45×10^{-15}	4.93×10^{-28}	4.04×10^{-54}	2.21×10^{-213}	2.00000
($\gamma = -1$)	8	1.45×10^{-15}	3.38×10^{-29}	1.84×10^{-56}	8.94×10^{-223}	2.00000
TM-2 ($\gamma = 0$)	8	3.25×10^{-15}	1.73×10^{-28}	4.92×10^{-55}	4.76×10^{-217}	2.00000
($\gamma = -1$)	8	8.25×10^{-16}	1.08×10^{-29}	1.87×10^{-57}	9.27×10^{-227}	2.00000
TM-3 ($\gamma = 0$)	8	1.66×10^{-15}	4.46×10^{-29}	3.22×10^{-56}	8.40×10^{-222}	2.00000
($\gamma = -1$)	8	3.97×10^{-16}	2.47×10^{-30}	9.55×10^{-59}	6.11×10^{-232}	2.00000
TM-4 ($\gamma = 0$)	8	9.41×10^{-16}	1.41×10^{-29}	3.19×10^{-57}	7.90×10^{-226}	2.00000
($\gamma = -1$)	8	2.13×10^{-16}	7.02×10^{-31}	7.63×10^{-60}	2.42×10^{-236}	2.00000
TM-5 ($\gamma = 0$)	8	9.76×10^{-16}	1.52×10^{-29}	3.69×10^{-57}	1.42×10^{-225}	2.00000
($\gamma = -1$)	8	2.22×10^{-16}	7.62×10^{-31}	8.98×10^{-60}	4.67×10^{-236}	2.00000
MNM	8	6.18×10^{-15}	6.37×10^{-28}	6.76×10^{-54}	1.74×10^{-212}	2.00000
KM	8	6.18×10^{-15}	6.37×10^{-28}	6.76×10^{-54}	1.74×10^{-212}	2.00000

Table 4: Numerical results for $f_2(x)$

Methods	t	$\lvert e_{t-1}\rvert$	$\lvert e_t\rvert$	$\lvert e_{t+1}\rvert$	$\lvert f(x_{t+1})\rvert$	COC
TM-1 ($\gamma = 0$)	7	3.54×10^{-16}	2.36×10^{-31}	1.05×10^{-61}	1.35×10^{-242}	2.00000
($\gamma = -1$)	7	2.80×10^{-18}	1.09×10^{-35}	1.63×10^{-70}	4.24×10^{-278}	2.00000
TM-2 ($\gamma = 0$)	7	6.03×10^{-17}	6.14×10^{-33}	6.36×10^{-65}	1.44×10^{-255}	2.00000
($\gamma = -1$)	7	2.48×10^{-19}	7.30×10^{-38}	6.33×10^{-75}	6.98×10^{-296}	2.00000
TM-3 ($\gamma = 0$)	7	4.87×10^{-18}	3.41×10^{-35}	1.67×10^{-69}	4.95×10^{-274}	2.00000
($\gamma = -1$)	7	6.40×10^{-21}	3.84×10^{-41}	1.38×10^{-81}	9.92×10^{-323}	2.00000
TM-4 ($\gamma = 0$)	7	4.08×10^{-19}	2.05×10^{-37}	5.15×10^{-74}	3.29×10^{-292}	2.00000
($\gamma = -1$)	7	1.27×10^{-22}	1.17×10^{-44}	1.01×10^{-88}	1.68×10^{-351}	2.00000
TM-5 ($\gamma = 0$)	7	4.03×10^{-19}	2.01×10^{-37}	4.99×10^{-74}	2.93×10^{-292}	2.00000
($\gamma = -1$)	7	1.25×10^{-22}	1.15×10^{-44}	9.74×10^{-89}	1.51×10^{-351}	2.00000
MNM	7	5.36×10^{-16}	5.56×10^{-31}	5.99×10^{-61}	1.49×10^{-239}	2.00000
KM	7	4.94×10^{-16}	4.69×10^{-31}	4.23×10^{-61}	3.66×10^{-240}	2.00000

Table 5: Numerical results for $f_3(x)$

Methods	t	$\lvert e_{t-1}\rvert$	$\lvert e_t\rvert$	$\lvert e_{t+1}\rvert$	$\lvert f(x_{t+1})\rvert$	COC
TM-1 ($\gamma = 0$)	7	8.49×10^{-14}	6.77×10^{-27}	4.30×10^{-53}	8.02×10^{-938}	2.00000
($\gamma = -1$)	7	2.05×10^{-15}	3.46×10^{-30}	9.93×10^{-60}	9.12×10^{-1058}	2.00000
TM-2 ($\gamma = 0$)	7	2.23×10^{-14}	4.45×10^{-28}	1.77×10^{-55}	6.06×10^{-981}	2.00000
($\gamma = -1$)	7	2.85×10^{-16}	6.37×10^{-32}	3.17×10^{-63}	6.68×10^{-1121}	2.00000
TM-3 ($\gamma = 0$)	7	1.91×10^{-18}	1.64×10^{-36}	1.21×10^{-72}	1.30×10^{-1292}	2.00000
($\gamma = -1$)	7	3.56×10^{-17}	4.28×10^{-34}	6.21×10^{-68}	6.26×10^{-1209}	2.00000
TM-4 ($\gamma = 0$)	7	2.23×10^{-23}	3.07×10^{-46}	5.81×10^{-92}	4.14×10^{-1639}	2.00000
($\gamma = -1$)	7	3.83×10^{-20}	7.40×10^{-40}	2.77×10^{-79}	1.10×10^{-1411}	2.00000
TM-5 ($\gamma = 0$)	7	3.45×10^{-21}	7.58×10^{-42}	3.66×10^{-83}	1.42×10^{-1480}	2.00000
($\gamma = -1$)	7	3.19×10^{-21}	5.36×10^{-42}	1.52×10^{-83}	3.18×10^{-1488}	2.00000
MNM	7	1.16×10^{-13}	1.27×10^{-26}	1.53×10^{-52}	7.46×10^{-928}	2.00000
KM	8	7.64×10^{-25}	5.54×10^{-49}	2.91×10^{-97}	7.99×10^{-1733}	2.00000

4.1 Unknown multiplicity

Now, it is clear that the proposed schemes are able to find efficiently the multiple roots of nonlinear equations with a known multiplicity. Then, the question about unknown multiplicity arises: Is it possible to modify these methods in order to simultaneously estimate the roots and their multiplicity? McNamee [12] presented different techniques due to several authors that estimated with unequal results the searched multiplicity. We propose a

Table 6: Numerical results for $f_4(x)$

Methods	t	$\lvert e_{t-1}\rvert$	$\lvert e_t\rvert$	$\lvert e_{t+1}\rvert$	$\lvert f(x_{t+1})\rvert$	COC
TM-1 ($\gamma = 0$)	7	3.38×10^{-17}	1.14×10^{-33}	1.30×10^{-66}	1.36×10^{-6565}	2.00000
($\gamma = -1$)	7	2.52×10^{-17}	6.22×10^{-34}	3.78×10^{-67}	1.25×10^{-6619}	2.00000
TM-2 ($\gamma = 0$)	7	3.01×10^{-17}	8.96×10^{-34}	7.95×10^{-67}	4.56×10^{-6587}	2.00000
($\gamma = -1$)	7	2.24×10^{-17}	4.86×10^{-34}	2.29×10^{-67}	1.42×10^{-6641}	2.00000
TM-3 ($\gamma = 0$)	7	1.20×10^{-21}	7.16×10^{-43}	2.57×10^{-85}	5.21×10^{-8451}	2.00000
($\gamma = -1$)	7	6.44×10^{-22}	1.99×10^{-43}	1.90×10^{-86}	6.21×10^{-8565}	2.00000
TM-4 ($\gamma = 0$)	7	3.80×10^{-18}	1.24×10^{-35}	1.31×10^{-70}	2.63×10^{-6968}	2.00000
($\gamma = -1$)	7	2.69×10^{-18}	6.06×10^{-36}	3.08×10^{-71}	7.54×10^{-7032}	2.00000
TM-5 ($\gamma = 0$)	7	9.27×10^{-19}	6.70×10^{-37}	3.51×10^{-73}	8.88×10^{-7228}	2.00000
($\gamma = -1$)	7	6.32×10^{-19}	3.04×10^{-37}	7.01×10^{-74}	3.04×10^{-7298}	2.00000
MNM	7	3.48×10^{-17}	1.21×10^{-33}	1.47×10^{-66}	2.98×10^{-6560}	2.00000
KM	7	3.48×10^{-17}	1.21×10^{-33}	1.47×10^{-66}	2.98×10^{-6560}	2.00000

modification of our schemes based on the idea of Traub (see [19]),

$$m = \lim_{t \to +\infty} \frac{\ln |f(x_t)|}{\ln \left| \frac{f(x_t)}{f'(x_t)} \right|},$$

for estimating the multiplicity in the Schöder scheme

$$x_{t+1} = x_t - m \frac{f(x_t)}{f'(x_t)}.$$

Expanding this idea to our class of iterative methods, we estimate the multiplicity m in expression (5) as

$$m_t \approx \frac{\ln |f(x_t)|}{\ln \left| \frac{L(\Theta)}{m_{t-1}} \right|}, \qquad t = 1, 2, \ldots,$$

starting with an initial $m_0 = \frac{\ln |f(x_t)|}{\ln |\Theta|}$. This provides different expressions of m_t for each special case of the presented family. These estimators of the multiplicity appear described in Table 7. For known methods, MNM and KM, the estimation of the multiplicity has been made by the original idea of Traub,

$$m_t = \frac{\ln |f(x_t)|}{\ln \left| \frac{f(x_t)}{f[\eta_t, x_t]} \right|}, \qquad t = 1, 2, \ldots,$$

and

$$m_t = \frac{\ln|f(x_t)|}{\ln\left|\frac{(1-a)f(\eta_t)+af(x_t)}{f[\eta_t,x_t]}\right|}, \qquad t = 1, 2, \ldots,$$

respectively.

Table 7: Estimators of the multiplicity for the special cases of class (5)

Method	m_t
TM-1	$\Theta_t(1+a\Theta_t)$
TM-2	$\frac{\Theta_t}{1+a\Theta_t}$
TM-3	$\frac{\Theta_t}{1+am_{t-1}\Theta_t}$
TM-4	$\frac{\Theta_t}{(1+a\sqrt{m_{t-1}}\Theta_t)^2}$
TM-5	$\frac{(\Theta_t^2+\Theta_t)}{1+am_{t-1}\Theta_t}$

We have made several tests on the same nonlinear equations described in Table 2, with the same initial estimations or other more distant estimates, using the same stopping criterium and number of digits. The obtained results are presented in Tables 8–11. In them, the estimated order of convergence, COC, does not appear as convergence is linear in all cases for proposed and existing methods. We consider that it is due to the natural instability provided by divided differences when the estimation of the multiplicity involves more uncertainty in data. We show the last error estimations, $|e_{t+1}|$, $|f(x_{t+1})|$, the number of iterations, iter, and the last approximation of the multiplicity provided by the iterative scheme, m_{t+1}.

Tables 8–11 show how the number of iterations has increased for all methods, possibly due to the instability generated by the estimation of the root multiplicity. All the schemes converge with a similar number of iterations, and all of them achieve, in all cases, a reasonable estimate of the multiplicity. Let us remark that the precision of the estimation of the multiplicity in the cases $f_1(x)$ to $f_3(x)$ is excellent. In the case of $f_4(x)$, the relative error of this estimation is around 4%.

Finally, we test our methods with a more challenging function $f_5(x) = 4x^2+8\sin(x)-4\pi x+\pi^2-8$, whose multiplicity is not obvious. Moreover, there exists a wide neighborhood around the multiple roots $\frac{\pi}{2}$, where the function is almost null. Due to the complexity of the problem, the stopping criterium

Table 8: Numerical results for $f_1(x)$, with $x_0 = 1.8$ and estimation of the multiplicity

Methods	*iter*	$\lvert e_{t+1}\rvert$	$\lvert f(x_{t+1})\rvert$	m_{t+1}
TM-1 ($\gamma = 0$)	32	2.48×10^{-51}	1.49×10^{-107}	2.0200
($\gamma = -1$)	32	4.89×10^{-51}	5.84×10^{-107}	2.0175
TM-2 ($\gamma = 0$)	32	3.25×10^{-51}	2.56×10^{-107}	2.0174
($\gamma = -1$)	32	5.66×10^{-51}	7.83×10^{-107}	2.0175
TM-3 ($\gamma = 0$)	32	4.49×10^{-51}	4.91×10^{-107}	2.0175
($\gamma = -1$)	32	6.71×10^{-51}	1.10×10^{-106}	2.0175
TM-4 ($\gamma = 0$)	32	5.24×10^{-51}	6.69×10^{-106}	2.0175
($\gamma = -1$)	32	7.21×10^{-51}	1.28×10^{-106}	2.0175
TM-5 ($\gamma = 0$)	32	5.12×10^{-51}	6.41×10^{-107}	2.0175
($\gamma = -1$)	32	7.13×10^{-51}	1.25×10^{-106}	2.0175
MNM	32	2.30×10^{-51}	1.28×10^{-107}	2.0174
KM	32	2.31×10^{-51}	1.29×10^{-107}	2.0174

Table 9: Numerical results for $f_2(x)$, with $x_0 = 1$ and estimation of the multiplicity

Methods	*iter*	$\lvert e_{t+1}\rvert$	$\lvert f(x_{t+1})\rvert$	m_{t+1}
TM-1 ($\gamma = 0$)	45	1.16×10^{-51}	1.79×10^{-104}	1.9605
($\gamma = -1$)	45	1.76×10^{-51}	4.17×10^{-104}	1.9603
TM-2 ($\gamma = 0$)	45	1.42×10^{-51}	2.72×10^{-104}	1.9604
($\gamma = -1$)	45	2.15×10^{-51}	6.27×10^{-104}	1.9603
TM-3 ($\gamma = 0$)	45	1.35×10^{-51}	2.47×10^{-104}	1.9604
($\gamma = -1$)	45	2.05×10^{-51}	5.71×10^{-104}	1.9603
TM-4 ($\gamma = 0$)	45	1.63×10^{-51}	3.59×10^{-104}	1.9604
($\gamma = -1$)	45	2.47×10^{-51}	8.27×10^{-104}	1.9602
TM-5 ($\gamma = 0$)	45	1.80×10^{-51}	4.38×10^{-104}	1.9603
($\gamma = -1$)	45	2.72×10^{-51}	1.01×10^{-103}	1.9602
MNM	45	1.15×10^{-51}	1.78×10^{-104}	1.9605
KM	45	1.94×10^{-51}	5.11×10^{-104}	1.9603

have changed to $|f(x_{t+1})| < 10^{-15}$. The numerical results are presented in Tables 12 and 13.

Table 10: Numerical results for $f_3(x)$, with $x_0 = 0.8$ and estimation of the multiplicity

Methods	*iter*	$\lvert e_{t+1}\rvert$	$\lvert f(x_{t+1})\rvert$	m_{t+1}
TM-1 ($\gamma = 0$)	66	1.01×10^{-51}	1.92×10^{-467}	8.7317
($\gamma = -1$)	67	4.68×10^{-52}	1.83×10^{-470}	8.7333
TM-2 ($\gamma = 0$)	66	3.77×10^{-51}	3.07×10^{-462}	8.7287
($\gamma = -1$)	67	1.77×10^{-51}	3.24×10^{-465}	8.7304
TM-3 ($\gamma = 0$)	66	1.92×10^{-51}	6.68×10^{-465}	8.7302
($\gamma = -1$)	67	8.92×10^{-52}	6.40×10^{-468}	8.7319
TM-4 ($\gamma = 0$)	66	4.75×10^{-51}	2.51×10^{-461}	8.7282
($\gamma = -1$)	67	2.22×10^{-51}	2.58×10^{-464}	8.7299
TM-5 ($\gamma = 0$)	67	6.41×10^{-52}	3.19×10^{-469}	8.7327
($\gamma = -1$)	67	9.67×10^{-51}	1.60×10^{-458}	8.7266
MNM	66	9.79×10^{-52}	1.50×10^{-467}	8.7317
KM	67	7.90×10^{-52}	2.13×10^{-468}	8.7322

Table 11: Numerical results for $f_4(x)$, with $x_0 = 1.5$ and estimation of the multiplicity

Methods	*iter*	$\lvert e_{t+1}\rvert$	$\lvert f(x_{t+1})\rvert$	m_{t+1}
TM-1 ($\gamma = 0$)	72	3.09×10^{-51}	2.07×10^{-2570}	47.9690
($\gamma = -1$)	72	4.97×10^{-51}	4.75×10^{-2560}	47.9610
TM-2 ($\gamma = 0$)	72	3.90×10^{-51}	2.48×10^{-2565}	47.9650
($\gamma = -1$)	72	6.26×10^{-51}	5.57×10^{-2555}	47.9570
TM-3 ($\gamma = 0$)	72	7.98×10^{-51}	1.15×10^{-2549}	47.9530
($\gamma = -1$)	73	5.40×10^{-52}	1.17×10^{-2608}	47.9980
TM-4 ($\gamma = 0$)	72	7.46×10^{-51}	3.98×10^{-2551}	47.9540
($\gamma = -1$)	73	5.05×10^{-52}	4.06×10^{-2610}	47.9999
TM-5 ($\gamma = 0$)	72	7.16×10^{-51}	4.96×10^{-2552}	47.9555
($\gamma = -1$)	73	4.84×10^{-52}	5.02×10^{-2611}	47.9999
MNM	72	3.08×10^{-51}	1.63×10^{-2570}	47.9699
KM	72	3.09×10^{-51}	1.96×10^{-2570}	47.9699

All the methods perform in a similar way, but schemes TM-1 to TM-5 get better results than MNM and KM methods with $x_0 = 1.8$, being approximately equal in the case where $x_0 = 1$. In all cases, we conclude that the

Table 12: Numerical results for $f_5(x)$, with $x_0 = 1.8$ and estimation of the multiplicity

Methods	*iter*	$\lvert f(x_{t+1})\rvert$	m_{t+1}
TM-1 ($\gamma = 0$)	5	5.27×10^{-16}	3.4652
($\gamma = -1$)	5	6.03×10^{-16}	3.4857
TM-2 ($\gamma = 0$)	5	5.78×10^{-16}	3.4755
($\gamma = -1$)	5	6.13×10^{-16}	3.4939
TM-3 ($\gamma = 0$)	5	6.11×10^{-16}	3.4910
($\gamma = -1$)	5	6.15×10^{-16}	3.4951
TM-4 ($\gamma = 0$)	5	6.14×10^{-16}	3.4952
($\gamma = -1$)	5	6.08×10^{-16}	3.4839
TM-5 ($\gamma = 0$)	5	6.13×10^{-16}	3.4939
($\gamma = -1$)	5	6.12×10^{-16}	3.4909
MNM	5	5.25×10^{-16}	3.4649
KM	5	5.26×10^{-16}	3.4650

Table 13: Numerical results for $f_5(x)$, with $x_0 = 1$ and estimation of the multiplicity

Methods	*iter*	$\lvert f(x_{t+1})\rvert$	m_{t+1}
TM-1 ($\gamma = 0$)	7	6.05×10^{-16}	3.4804
($\gamma = -1$)	7	5.65×10^{-16}	3.4722
TM-2 ($\gamma = 0$)	7	6.14×10^{-16}	3.4950
($\gamma = -1$)	7	2.18×10^{-16}	3.4440
TM-3 ($\gamma = 0$)	7	5.64×10^{-16}	3.4719
($\gamma = -1$)	8	5.93×10^{-16}	3.4635
TM-4 ($\gamma = 0$)	7	2.86×10^{-16}	3.4468
($\gamma = -1$)	8	6.16×10^{-16}	3.4979
TM-5 ($\gamma = 0$)	7	1.35×10^{-16}	3.4411
($\gamma = -1$)	8	6.16×10^{-16}	3.4970
MNM	7	6.04×10^{-16}	3.4788
KM	7	6.06×10^{-16}	3.4870

multiplicity of the root is 4. Regarding the precision of the estimation of multiplicity, the relative error is, in this case, between 12% and 13%.

5 Conclusions

In this paper, a novel second-order scheme free from derivatives for locating the multiple roots of univariate nonlinear functions was developed. A weight function $L(\Theta_t)$ was used during the development of the new scheme. The greatest benefit of using weight functions is that we can produce as many special cases as possible and various choices of the weight functions subject to some conditions. The error equation in the main result corroborated the second-order convergence of the proposed scheme. It has been shown that the new family of methods is a generalized one as it contains almost all the existing methods of the same type as its special cases. Not only the basins of attraction, but also a deep stability analysis in the Riemann sphere have also been performed for the new schemes, showing the good performance of all the proposed subclasses for small values of the parameter, especially the good performance of the TM-2 class. On the other hand, we have assigned to each method, parallel to the iterative process to obtain the root, an iterative scheme to approximate the multiplicity when it is not known.

The numerical results for different examples showed that our new method is efficient, especially in those cases where the derivative of the nonlinear function is not available.

Conflict of interests: The authors declare no conflict of interest.

Acknowledgements

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

References

[1] Beardon, A.F. *Iteration of rational functions: Complex analytic dynamical systems*, Springer Science & Business Media, 132, 2000.

[2] Blanchard, P. *The dynamics of Newton's method*, Proc. Symposia Appl. Math. 49 (1994), 139–154.

[3] Chicharro, F.I., Cordero, A. and Torregrosa, J.R. *Drawing dynamical and parameters planes of iterative families and methods*, Sci. World J. 2013 (2013), Article ID 780153, 11 pages.

[4] Chicharro, F.I., Cordero, A., Gutiérrez, J.M. and Torregrosa, J.R. *Complex dynamics of derivative-free methods for nonlinear equations*, Appl. Math. Comput. 219(12) (2013), 7023–7035.

[5] Cordero, A., Neta, B. and Torregrosa, J.R. *Reasons for stability in the construction of derivative-free multistep iterative methods*, Math. Meth. Appl. Sci. (2023), 1–16.

[6] Hansen, E. and Patrick, M. *A family of root finding methods*, Numer. Math. 27 (1977), 257–269.

[7] Kansal, M., Kanwar, V. and Bhatia, S. *On some optimal multiple root-finding methods and their dynamics*, Appl. Math. 10 (2015), 349–367.

[8] Kansal, M., Alshomrani, A.S., Bhalla, S., Behl, R. and Salimi, M. *One parameter optimal derivative-free family to find the multiple roots of algebraic nonlinear equations*, Mathematics, 7 (2019), 655.

[9] Kumar, D., Sharma, J.R. and Argyros, I.K. *Optimal one-point iterative function free from derivatives for multiple roots*, Mathematics, 8 (2020), 709.

[10] Kung, H.T. and Traub, J.F. *Optimal order of one-point and multipoint iteration*, Assoc. Comput. Mach. 21 (1974), 643–651.

[11] Li, S.G., Cheng, L.Z. and Neta, B. *Some fourth-order nonlinear solvers with closed formulae for multiple roots*, Comput. Math. Appl. 59 (2010), 126–135.

[12] McNamee, J.M. *A Comparison of methods for accelerating convergence of Newton's method for multiple polynomial roots*, ACM SIGNUM Newsletter, (1998), 17–22.

[13] Neta, B. *New third order nonlinear solvers for multiple roots*, Appl. Math. Comput. 202 (2008), 162–170.

[14] Ostrowski, A.M. *Solutions of equations and systems of equations*, Academic Press, 1966.

[15] Schröder, E. *Über unendlich viele Algorithmen zur Auflösung der Gleichungen*, Math. Ann. 2 (1870), 317–365.

[16] Sharifi, M., Babajee, D.K.R. and Soleymani, F. *Finding the solution of nonlinear equations by a class of optimal methods*, Comput. Math. Appl. 63 (2012), 764–774.

[17] Sharma, J.R. and Sharma, R.A. *Modified Jarratt method for computing multiple roots*, Appl. Math. Comput. 217 (2010), 878–881.

[18] Soleymani, F., Babajee, D.K.R. and Lotfi, T. *On a numerical technique for finding multiple zeros and its dynamics*, J. Egypt. Math. Soc. 21 (2013), 346–353.

[19] Traub, J.F. *Iterative methods for the solution of equations*, Prentice-Hall Series in Automatic Computation, Englewood Cliffs, NJ, USA, 1964.

[20] Weerakoon, S. and Fernando, T.G.I. *A variant of Newton's method with accelerated third-order convergence*, Appl. Math. Lett. 13 (2000), 87–93.

[21] Yun, B.I. *A non-iterative method for solving non-linear equations*, Appl. Math. Comput. 198 (2008), 691–699.

[22] Zhou, X., Chen, X. and Song, Y. *Constructing higher-order methods for obtaining the multiple roots of nonlinear equations*, J. Comput. Appl. Math. 235 (2011), 4199–4206.

Iranian Journal of Numerical Analysis and Optimization
Vol. 14, No. 1, 2024, pp 315–329
https://doi.org/10.22067/ijnao.2023.84431.1315
https://ijnao.um.ac.ir/

CC BY NC

Research Article

Wavelet approximation with Chebyshev wavelets

S. M. Jesmani*, H. Mazaheri* and S. Shojaeian

*Corresponding author

*Corresponding author

Received 13 August 2023; revised 7 December 2023; accepted 10 December 2023

S. M. Jesmani
Department of Materials and Metallurgical Engineering, Technical and Vocational University(TVU), Tehran, Iran. e-mail: smjesmani@tvu.ac.ir

H. Mazaheri
Department of Mathematics, Yazd University, 89195-741, Yazd, Iran. e-mail: hmazaheri@yazd.ac.ir

Saman Shojaeian
Department of Mathematics, Yazd University, 89195-741, Yazd, Iran. e-mail: samanshojaeian1989@yahoo.com

Abstract

In this paper, we study wavelet approximation of the Chebyshev polynomials of the first, second, third, and fourth kinds. We estimate the wavelet approximation of a function f having bounded first derivatives.

AMS subject classifications (2020): 41A65, 41A52, 46N10.

How to cite this article
Jesmani, S. M., Mazaheri, H. and Shojaeian, S., Wavelet approximation with Chebyshev wavelets. *Iran. J. Numer. Anal. Optim.*, 2024; 14(1): 315-329. https://doi.org/10.22067/ijnao.2023.84431.1315

Keywords: Wavelet approximation; Chebyshev polynomials; Shifted Chebyshev; Continuous functions.

1 Introduction

In recent years, wavelets have found their way into many different fields of science and engineering, particularly, in signal analysis, time-frequency analysis, and fast algorithms. Wavelets allow an accurate representation of several functions. The wavelet approximation technique is a new tool for finding and analyzing unexpected seismic signal processing changes. There is scarcely any area of numerical analysis where Chebyshev polynomials do not drop in like surprise visitors, and indeed there are now a number of subjects in which these polynomials take a significant position in modern developments including orthogonal polynomials, polynomial approximation, numerical integration, and spectral methods for partial differential equations (see [4, 6]).

In Table 1, we have Chebyshev polynomials on the trivial $[-1, 1]$.

Table 1: Chebyshev polynomials

Chebyshev Polynomials	Shifted Chebyshev	w (weight)	w^* (shifted weight)
$T_n(x) = \cos\ n\ \theta$	$T_n^*(x) = T_n(2x-1)$	$\frac{1}{\sqrt{1-x^2}}$	$\frac{1}{\sqrt{x-x^2}}$
$U_n(x) = \frac{\sin\ (n+1)\ \theta}{\sin\ \theta}$	$U_n^*(x) = U_n(2x-1)$	$\sqrt{1-x^2}$	$\sqrt{x-x^2}$
$V_n(x) = \frac{\cos(n+\frac{1}{2})\theta}{\cos(\frac{\theta}{2})}$	$V_n^*(x) = V_n(2x-1)$	$\sqrt{\frac{1+x}{1-x}}$	$\sqrt{\frac{x}{1-x}}$
$W_n(x) = \frac{\sin(n+\frac{1}{2})\theta}{\sin(\frac{\theta}{2})}$	$W_n^*(x) = W_n(2x-1)$	$\sqrt{\frac{1-x}{1+x}}$	$\sqrt{\frac{1-x}{x}}$

With above definition, we obtain the fundamental recurrence relation:

$$W_{n+1}(x) = 2xW_n(x) - W_{n-1}(x), \quad n \geq 2, \quad W_0(x) = 1, \quad W_1(x) = 2x + 1.$$

In Table 2, we have Chebyshev wavelets, for $k \in \mathbb{N}$, $m \geq 0$, and $n = 0, 1, 2, \ldots, 2^k$.

Table 2: Chebyshev wavelets

Sequences Chebyshev Polynomials	Wavelet Chebyshev
$T_{n,m}$	$\begin{cases} 2^{\frac{k}{2}} T_m(2^k t - n), & t \in [\frac{n}{2^k}, \frac{n+1}{2^k}], \\ 0 & \text{otherwise.} \end{cases}$
$U_{n,m}$	$\begin{cases} 2^{\frac{k}{2}} U_m(2^k t - n), & t \in [\frac{n}{2^k}, \frac{n+1}{2^k}], \\ 0 & \text{otherwise.} \end{cases}$
$V_{n,m}$	$\begin{cases} 2^{\frac{k}{2}} V_m(2^k t - n), & t \in [\frac{n}{2^k}, \frac{n+1}{2^k}], \\ 0 & \text{otherwise.} \end{cases}$
$W_{n,m}$	$\begin{cases} 2^{\frac{k}{2}} W_m(2^k t - n), & t \in [\frac{n}{2^k}, \frac{n+1}{2^k}], \\ 0 & \text{otherwise.} \end{cases}$

In the following, we suppose $\psi_{n,m} = T_{n,m}, U_{n,m}, V_{n,m}, W_{n,m}$. A function $f \in L^2[0,1)$ is expanded by wavelet series as

$$f(t) = \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} c_{n,m} \psi_{n,m}(t),$$

where

$$c_{n,m} = \int_0^1 f(t) \psi_{n,m}(t) \omega^*_{n,m}(t) dt,$$

and $\omega_{n,m}$ is the weight function of pseudo Chebyshevs. Also,

$$\int_0^1 \psi_n(x) \psi_m(x) \omega^*_{n,m}(t) dx = L,$$

where

$$L = \begin{cases} \frac{\pi}{2}, & m = n, \ \psi = T, U, \\ 0, & \text{m} \neq n, \ \psi = T, W, U, V, \\ \pi, & m = n, \ \psi = V, W, \end{cases}$$

$$\omega_{n,m}(t) = \begin{cases} \frac{1}{\sqrt{1-t^2}} & if \ \psi = T, \\ \sqrt{1-t^2} & if \ \psi = U, \\ \sqrt{\frac{1+t}{1-t}} & if \ \psi = V, \\ \sqrt{\frac{1-t}{1+t}} & if \ \psi = W. \end{cases}$$

Definition 1 (Multiresolution analysis)**.** A sequence of closed subspaces $\{V_j\}$ of $L^2(\mathbb{R})$ (inner product space), $j \in Z$, is called a multiresolution in $L^2(\mathbb{R})$ if it satisfies the following conditions:

(i) $V_j \subset V_{j+1}$;

(ii) $f(x) \in V_j \iff f(2x) \in V_{j+1}$;

(iii) $f(x) \in V_0 \iff (x+1) \in V_0$;

(iv) $\cup_{-\infty}^{\infty} V_j$ is dense in $L^2((\mathbb{R})$, and $\cap_{-\infty}^{\infty} V_j = \{0\}$;

(v) There exists a function $\phi \in V_0$ such that the collection $\{\phi(x–k) : k \in \mathbb{Z}\}$ is a Riesz basis of V_0.

Definition 2. Let $P_n(f)$ be the orthogonal projection of $L^2([0,1])$ onto V_n. Then

$$P_n f = \sum_{-\infty}^{\infty} a_{n,k}\phi_{n,k}, \quad n = 1, 2, 3, \ldots,$$

where

$$a_{n,k} = \langle f, \phi_{n,k}\rangle.$$

Thus

$$P_n(f) = \sum_{-\infty}^{\infty} \langle f, \phi_{n,k}\rangle \phi_{n,k}, \quad n = 1, 2, 3, \ldots,$$

Definition 3. The wavelet approximation of Chebyshev polynomial is defined by

$$E_n(f) = \|f–P_n(f)\|_2 = \int_a^1 (f(t)–P_n(f)(t))^2 dt = O(\phi(n)),$$

where

$$f(t) = \sum_{n=0}^{\infty}\sum_{m=0}^{\infty} c_{n,m} V_{n,m}(t)$$

or

$$f(t) = \sum_{n=0}^{\infty}\sum_{m=0}^{\infty} c_{n,m} W_{n,m}(t).$$

If $E_n(f) \to 0$ as $n \to \infty$, then $E_n(f)$ is called the best approximation of f (see [1, 2, 3]).

We conclude this section by a list of known lemmas.

Lemma 1. [5] Let $f \in L^2([0,1])$ be a continuous function such that for constant $P > 0$, $|f"(x)| \leq P$. If $f(t) = \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} c_{n,m} V_{n,m}(t)$ is expanded in terms of Chebyshev polynomial of the third kind. Then the Chebyshev wavelet approximation $E_{2^k,l}(t)$ of f by $(2^k, l)$th partial sum

$$s_{2^{k-2},M}(t) = \sum_{n=1}^{2^{k-2}} \sum_{m=0}^{M-1} c_{n,m} \psi_{n,m}(t), \tag{1}$$

of its Chebyshev wavelet series in $L^2_{w*}[0,1]$ is given by

$$\|f - s_{2^{k-2},M}\|_2^2 = O(\frac{1}{2^{2k}(M-1)^{\frac{3}{2}}}). \quad M \geq 1 \quad (M \textit{ is resolution})$$

Lemma 2. [5] Let $f \in L^2([0,1])$ be a continuous function such that for constant $P > 0$, $|f"(x)| \leq P$. If $f(t) = \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} c_{n,m} W_{n,m}(t)$ is expanded in terms of the Chebyshev polynomial of the fourth kind. Then the Chebyshev wavelet approximation $E_{2^k,l}(t)$ of f by $(2^k, l)$th partial sum

$$s_{2^{k-2},M}(t) = \sum_{n=1}^{2^{k-2}} \sum_{m=0}^{M-1} c_{n,m} \psi_{n,m}(t), \tag{2}$$

of its Chebyshev wavelet series in $L^2_{w*}[0,1]$ is given by

$$E^2_{2^k-1,l} = O(\frac{1}{M! 2^{M(k+1)}}).$$

2 Main results

In this section, we consider wavelet approximation in Chebyshev polynomials of first, second, third, and fourth kinds and Haar wavelet. We estimate the wavelet approximation of a function f having bounded first derivatives.

Theorem 1. Let $f \in L^2([0,1])$ be a continuous function and let for constant $P > 0$, $|f'(x)| \leq P$.

(i) If $f(t) = \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} c_{n,m} T_{n,m}(t)$ is expanded in terms of Chebyshev polynomial of the first kind. Then the Chebyshev wavelet approximation $E_{2^k,l}(t)$ of f by $(2^k, l)$th partial sum is

$$s_{2^k-1,l} = \sum_{n=2^{k-1}+1}^{\infty} \sum_{m=l}^{\infty} c_{n,m} T_{n,m},$$

and

$$E_{2^k,l}(f) = \|f - s_{2^k,l}\|_2^2 = O(\frac{P}{(l+1)^{\frac{3}{2}}}).$$

(ii) If $f(t) = \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} c_{n,m} U_{n,m(t)}$ is expanded in terms of Chebyshev polynomial of the second kind. Then the Chebyshev wavelet approximation $E_{2^{k-1},l}(t)$ of f by $(2^{k-1}, l)$th partial sum is

$$s_{2^k-1,l} = \sum_{n=2^{k-1}+1}^{\infty} \sum_{m=l}^{\infty} c_{n,m} U_{n,m},$$

and

$$E_{2^k-1,l}(f) = \|f - s_{2^{k-1},l}\|_2 = O(\frac{P^{\frac{1}{2}}}{2^{\frac{k}{4}} - 1}(\frac{2l+4}{(l+1)(l+3)})^{\frac{1}{2}}).$$

(iii) If $f(t) = \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} c_{n,m} V_{n,m}(t)$ is expanded in terms of Chebyshev polynomial of the third kind. Then the Chebyshev wavelet approximation $E_{2^k,l}(t)$ of f by $(2^k, l)$ th partial sum is

$$s_{2^k-1,l} = \sum_{n=2^{k-1}+1}^{\infty} \sum_{m=0} c_{n,m} V_{n,m},$$

and

$$E_{2^k,l}^2(f) = \|f - s_{2^k,l}\|_2^2 = O(2^{-\frac{k}{2}-1}(\ln \frac{l}{(l+1)(l+3)} - \frac{1}{l+1} - \frac{27}{8}\frac{1}{l+2} - \frac{1}{l})).$$

(iv) If $f(t) = \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} c_{n,m} W_{n,m}(t)$ is expanded in terms of Chebyshev polynomial of the fourth kind. Then the Chebyshev wavelet approximation $E_{2^k,l}(t)$ of f by $(2^k, l)$th partial sum is

$$s_{2^k-1,l} = \sum_{n=2^{k-1}+1}^{\infty} \sum_{m=0} c_{n,m} W_{n,m},$$

and

$$E_{2^k-1,l}^2 = O(2^{-\frac{3k}{2}-3}(\frac{1}{4(l+1)} + \frac{1}{l} + \frac{1}{l+3} + \frac{5}{3}\ln \frac{l}{l+3})).$$

Proof. (i) We have

$$\begin{aligned}
c_{n,m} &= < f(t), T_{n,m} >_{w_1} \\
&= 2^{\frac{k}{2}} \int_0^1 f(t) T_m(2^k t - n) w_1(kt^k - n) dt \\
&= 2^{\frac{k}{2}} \int_0^\pi f(\frac{\cos\theta + n}{2^k}) T_m(\cos\theta) \frac{1}{\sqrt{1-\cos^2\theta}} \sin\theta d\theta \\
&= 2^{\frac{k}{2}} \int_0^\pi f(\frac{\cos\theta + n}{2^k}) \cos(m\theta) d\theta \\
&= 2^{\frac{k}{2}} [(f(\frac{\cos\theta + n}{2^k})(-\frac{1}{m}\sin m\theta))|_0^\pi \\
&\quad + \frac{1}{2^k m} \int_0^\pi f'(\frac{\cos\theta + n}{2^k}) \sin\theta \sin(m\theta) d\theta] \\
&= 2^{\frac{k}{2}} [\frac{1}{2^k m} \int_0^\pi f'(\frac{\cos\theta + n}{2^k}) \sin\theta \sin(m\theta) d\theta] \\
&\leq 2^{\frac{k}{2}} [\frac{P}{2^k m} \int_0^\pi \sin\theta \sin(m\theta) d\theta] \\
&\leq 2^{\frac{k}{2}} (-\frac{P}{2^k m})[-\frac{1}{m}\sin\theta\cos(m\theta)|_0^\pi - \frac{1}{m}\int_0^\pi \cos(m\theta)\cos(\theta) d\theta] \\
&\leq \frac{P}{2^{\frac{k}{2}} m^2} (\int_0^\pi \cos(m\theta)\cos(\theta) d\theta).
\end{aligned}$$

Therefore

$$\begin{aligned}
|c_{n,m}| &= \frac{P}{2^{\frac{k}{2}} m^2} |\int_0^\pi \cos(m\theta)\cos(\theta) d\theta| \\
&\leq \frac{P}{2^{\frac{k}{2}} m^2} \int_0^\pi |\cos(m\theta)||\cos(\theta)| d\theta \\
&\leq \frac{P\pi}{2^{\frac{k}{2}} m^2} \\
|c_{n,m}|^2 &\leq \frac{P^2\pi^2}{2^k m^4}.
\end{aligned}$$

For $T_{n,m}$

$$\begin{aligned}
\|T_{n,m}\|_2^2 &= 2^k \int_0^1 |T_m(2^k t - n)|^2 w_1(2^k t - n) dt \\
&\leq 2^k \int_{-1}^1 T_m(2^k t - n)^2 w_1(2^k t - n) dt \\
&= \frac{\pi}{2} 2^k \\
&= \pi 2^{k-1},
\end{aligned}$$

and

$$\begin{aligned}
E^2_{2^k-1,l} &= \|f - s_{2^k-1,l}\|_2^2 \\
&= \|\sum_{n=0}^{\infty}\sum_{m=0}^{\infty} c_{n,m}T_{n,m} - \sum_{n=2^{k-1}+1}^{\infty}\sum_{m=0}^{l} c_{n,m}T_{n,m}\|_2^2 \\
&= \|\sum_{n=0}^{2^k-1}\sum_{m=l+1}^{\infty} c_{n,m}T_{n,m}\|_2^2 \\
&= \sum_{n=0}^{2^k-1}\sum_{m=l+1}^{\infty} |c_{n,m}|^2\|T_{n,m}\|_2^2 \\
&\leq \frac{\pi}{2}\sum_{n=0}^{2^k-1}\sum_{m=l+1}^{\infty}\frac{P^2\pi^2}{m^4} \\
&= \frac{P^2\pi^3}{2}\sum_{n=0}^{2^k-1}\sum_{m=l+1}^{\infty}\frac{1}{m^4}.
\end{aligned}$$

It is follows that

$$E_{2^k-1,l} = O(\frac{P}{(l+1)^{\frac{3}{2}}}).$$

(ii) We have

$$\begin{aligned}
\|U_{n,m}\|_2^2 &= 2^k\int_a^1 |U_m(2^kt-n)|^2 w_2(2^kt-n)dt \\
&= 2^k\int_a^1 U_m(2^kt-n)^2 w_2(2^kt-n)dt \\
&= \frac{\pi}{2}2^k \\
&= \pi 2^{k-1}.
\end{aligned}$$

Also,

$$\begin{aligned}
c_{n,m} &= < f(t), U_{n,m} >_w \\
&= 2^{\frac{k}{2}}\int_0^1 f(t)U_m(2^kt-n)w_1(2^kt-n)dt \\
&= 2^{\frac{k}{2}}\int_0^{\pi} f(\frac{\cos\theta+n}{2^k})U_m(\cos\theta)\sin\theta\sin\theta d\theta \\
&= 2^{\frac{k}{2}}\int_0^{\pi} f(\frac{\cos\theta+n}{2^k})\frac{\sin(m+1)\theta}{\sin\theta}\sin^2\theta d\theta
\end{aligned}$$

$$
\begin{aligned}
&= 2^{\frac{k}{2}} \int_0^{\pi} f(\frac{\cos\theta + n}{2^k}) \sin(m+1)\theta \sin\theta d\theta \\
&= 2^{\frac{k}{2}} \int_0^{\pi} f(\frac{\cos\theta + n}{2^k}) \frac{1}{2}(\cos(m\theta) - \cos(m+2)\theta) d\theta \\
&= 2^{\frac{k}{2}-1} \int_0^{\pi} f(\frac{\cos\theta + n}{2^k})(\cos(m\theta) - \cos(m+2)\theta) d\theta \\
&= 2^{\frac{k}{2}-1} [f(\frac{\cos\theta + n}{2^k})(\frac{1}{m}\sin(m\theta) - \frac{1}{m+2}\sin(m+2)]_0^{\pi} \\
&\quad + \frac{1}{2^{\frac{k}{2}+1}} \int_0^{\pi} f'(\frac{\cos\theta + n}{2^k}) \sin\theta (\frac{1}{m}\sin(m\theta) - \frac{1}{m+2}\sin(m+2)) d\theta \\
&\leq \frac{P}{2^{\frac{k}{2}+1}} \int_0^{\pi} \sin\theta (\frac{1}{m}\sin(m\theta) - \frac{1}{m+2}\sin(m+2)) d\theta \\
&= \frac{P}{2^{\frac{k}{2}+1}} [\sin\theta (\frac{1}{m^2}\cos m\theta + \frac{1}{(m+2)^2}\cos(m+2)\theta]_0^{\pi} \\
&\quad + \frac{P}{2^{\frac{k}{2}+1}} \int_0^{\pi} \cos\theta (\frac{1}{m^2}\cos m\theta + \frac{1}{(m+2)^2}\cos(m+2)\theta) d\theta.
\end{aligned}
$$

Therefore

$$
\begin{aligned}
|c_{n,m}| &\leq \frac{P}{2^{\frac{k}{2}+1}} | \int_0^{\pi} \cos\theta (\frac{1}{m^2}\cos m\theta + \frac{1}{(m+2)^2}\cos(m+2)\theta) d\theta | \\
&\leq \frac{P}{2^{\frac{k}{2}+1}} \int_0^{\pi} |\cos\theta| (\frac{1}{m^2}|\cos m\theta| + \frac{1}{(m+2)^2}|\cos(m+2)|(\theta) d\theta \\
&\leq \frac{P}{2^{\frac{k}{2}+1}} \int_0^{\pi} (\frac{1}{m^2} + \frac{1}{(m+2)^2}) d\theta
\end{aligned}
$$

and

$$
\begin{aligned}
E^2_{2^k-1,l} &= \|f - s_{2^k-1,l}\|_2^2 \\
&= \| \sum_{n=0}^{\infty} \sum_{m=0}^{\infty} c_{n,m} U_{n,m} - \sum_{n=2^{k-1}+1}^{\infty} \sum_{m=0}^{l} c_{n,m} T_{n,m} \|_2^2 \\
&= \| \sum_{n=0}^{2^{k-1}} \sum_{m=l+1}^{\infty} c_{n,m} U_{n,m} \|_2^2 \\
&= \sum_{n=0}^{2^{k-1}} \sum_{m=l+1}^{\infty} |c_{n,m}|^2 \|U_{n,m}\|_2^2 \\
&\leq \frac{P}{2^{\frac{k}{2}+1}} \sum_{n=0}^{2^{k-1}} \sum_{m=l+1}^{\infty} (\frac{1}{m^2} + \frac{1}{(m+2)^2})
\end{aligned}
$$

$$\leq \frac{P}{2^{\frac{k}{2}+1}} \sum_{n=0}^{2^{k-1}} (\frac{2l+4}{(l+1)(l+3)})$$
$$\leq \frac{P}{2^{\frac{k}{2}+1}} \sum_{n=0}^{2^{k-1}} (\frac{2l+4}{(l+1)(l+3)}).$$

Therefore

$$E_{2^k-1,l} = \|f - s_{2^k-1,l}\|_2 = O(\frac{P^{\frac{1}{2}}}{2^{\frac{k}{4}}-1}(\frac{2l+4}{(l+1)(l+3)})^{\frac{1}{2}}).$$

(iii) We have

$$\begin{aligned}
c_{n,m} &= < f(t), V_{n,m} >_w \\
&= 2^{\frac{k}{2}} \int_0^1 f(t) V_m(2^k t - n) w_1(2^k t - n) dt \\
&= -2^{\frac{k}{2}} \int_0^\pi f(\frac{\cos\theta + n}{2^k}) \frac{\cos(m+\frac{1}{2})\theta}{\cos(\frac{\theta}{2})} \sqrt{\frac{1+\cos\theta}{1-\cos\theta}} \sin\theta d\theta \\
&= -2^{\frac{k}{2}} \int_0^\pi f(\frac{\cos\theta + n}{2^k}) \frac{\cos(m+\frac{1}{2})\theta}{\cos(\frac{\theta}{2})} \sqrt{\frac{2\cos^2(\frac{\theta}{2})}{2\sin^2(\frac{\theta}{2})}} 2\sin\frac{\theta}{2}\cos\frac{\theta}{2} d\theta \\
&= -2^{\frac{k}{2}} \int_0^\pi f(\frac{\cos\theta + n}{2^k}) 2\cos(m+\frac{1}{2})\theta \cos\frac{\theta}{2} d\theta \\
&= -2^{\frac{k}{2}} \int_0^\pi f(\frac{\cos\theta + n}{2^k})(\cos(m+1)\theta + \cos m\theta) d\theta \\
&= -2^{\frac{k}{2}} [f(\frac{\cos\theta + n}{2^k})(\frac{1}{m+1}\sin(m+1)\theta + \frac{1}{m}\sin m\theta)]_0^\pi \\
&\quad +2^{\frac{k}{2}} \int_0^\pi f'(\frac{\cos\theta + n}{2^k}) \frac{\sin\theta}{2^k} (\frac{1}{m+1}\sin(m+1)\theta + \frac{1}{m}\sin m\theta) d\theta \\
&= 2^{\frac{k}{2}} \int_0^\pi f'(\frac{\cos\theta + n}{2^k}) \frac{\sin\theta}{2^k} (\frac{1}{m+1}\sin(m+1)\theta + \frac{1}{m}\sin m\theta) d\theta \\
&\leq 2^{-\frac{k}{2}} P \int_0^\pi \sin\theta (\frac{1}{m+1}\sin(m+1)\theta + \frac{1}{m}\sin m\theta)) d\theta \\
&= 2^{-\frac{k}{2}} P [(\frac{1}{m+1}\sin\theta\sin(m+1)\theta + \frac{1}{m}\sin\theta\sin m\theta)]_0^\pi \\
&\quad -2^{-\frac{k}{2}-1} P \int_0^\pi \cos\theta (\frac{1}{m+1}\sin(m+1)\theta + \frac{1}{m}(\sin m\theta) d\theta \\
&= 2^{-\frac{k}{2}-1} P \int_0^\pi (\frac{1}{m+1}\cos\theta\sin(m+1)\theta - \frac{1}{m}\cos\theta\sin m\theta) \\
&= 2^{-\frac{k}{2}-1} P \int_0^\pi (\frac{1}{2(m+1)}(\sin m\theta + \sin(m+2)\theta)
\end{aligned}$$

$$
\begin{aligned}
&-\frac{1}{2m}(\sin(m-1)\theta+\sin(m+1)\theta))d\theta\\
&=2^{-\frac{k}{2}-1}P[-\frac{1}{2m(m+1)}\cos m\theta-\frac{1}{(m+1)(m+2)}\cos(m+2)\theta\\
&+\frac{1}{2m(m-1)}\cos(m-1)\theta+\frac{1}{m(m+1)}\cos(m+1)\theta]_0^{\pi}\\
&=2^{-\frac{k}{2}-1}P[-\frac{1}{2m(m+1)}(\cos m\pi-1)\\
&-\frac{1}{(m+1)(m+2)}(\cos(m+2)\pi-1)\\
&+\frac{1}{2m(m-1)}(\cos(m-1)\pi-1)+\frac{1}{m(m+1)}(\cos(m+1)\pi-1)],
\end{aligned}
$$

and

$$
\begin{aligned}
|c_{n,m}|&=2^{-\frac{k}{2}-1}P|-\frac{1}{2m(m+1)}\cos(m\pi-1)\\
&-\frac{1}{(m+1)(m+2)}(\cos(m+2)\pi-1)\\
&+\frac{1}{2m(m-1)}(\cos(m-1)\pi-1)+\frac{1}{m(m+1)}(\cos(m+1)\pi-1)|\\
&\leq 2^{-\frac{k}{2}-1}P(\frac{1}{m(m+1)}+\frac{2}{(m+1)(m+2)}+\frac{1}{m(m-1)}+\frac{2}{m(m+1)})\\
|c_{n,m}|^2&=2^{-k-2}P^2(\frac{1}{(m^2(m+1)^2}+\frac{4}{(m+1)^2(m+2)^2}+\frac{1}{m^2(m-1)^2}\\
&+\frac{4}{m^2(m+1)^2}+\frac{2}{m(m+2)(m+1)^2}+\frac{1}{m^2(m+1)(m-1)}\\
&+\frac{2}{m^2(m+1)^2}+\frac{2}{m(m+1)(m-1)(m+2)}\\
&+\frac{4}{m(m+2)(m+1)^2}+\frac{2}{m^2(m-1)(m+1)})\\
&=2^{-k-2}P^2(\frac{4}{(m+1)^2(m+2)^2}+\frac{1}{m^2(m-1)^2}\\
&+\frac{6}{m^2(m+1)^2}+\frac{6}{m(m+2)(m+1)^2}\\
&+\frac{3}{m^2(m+1)(m-1)}+\frac{2}{m(m+1)(m-1)(m+2)})\\
&=-10\frac{1}{m}+4\frac{1}{m^2}+10\frac{1}{m+1}+8\frac{1}{(m+1)^2}-\frac{1}{6}\frac{1}{m-1}\\
&+\frac{1}{(m-1)^2}-\frac{23}{2}\frac{1}{m+2}.
\end{aligned}
$$

For $V_{n,m}$, we have

$$\begin{aligned}\|V_{n,m}\|_2^2 &= 2^{-\frac{k}{2}}\int_0^1 |V_m(2^k t-n)|^2 w(t)dt\\ &\le 2^{-\frac{k}{2}}\int_{-1}^1 V_m(2^k t-n)^2\sqrt{\frac{1+t}{1-t}}dt\\ &= 2^{-\frac{k}{2}}\frac{\pi}{2}\\ &= 2^{-\frac{k}{2}-1}\pi.\end{aligned}$$

Therefore

$$\begin{aligned}E^2_{2^k-1,l} &= \|f-s_{2^k-1,l}\|_2^2\\ &= \|\sum_{n=0}^{\infty}\sum_{m=0}^{\infty}c_{n,m}V_{n,m}-\sum_{n=2^{k-1}+1}^{\infty}\sum_{m=0}^{l}c_{n,m}V_{n,m}\|_2^2\\ &= \|\sum_{n=0}^{2^k-1}\sum_{m=l+1}^{\infty}c_{n,m}V_{n,m}\|_2^2\\ &= \sum_{n=0}^{2^k-1}\sum_{m=l+1}^{\infty}|c_{n,m}|^2\|V_{n,m}\|_2^2\\ &\le 2^{-\frac{k}{2}-1}\pi\sum_{m=l+1}^{\infty}(-10\frac{1}{m}+4\frac{1}{m^2}+10\frac{1}{m+1}\\ &\quad +8\frac{1}{(m+1)^2}-\frac{1}{m-1}+\frac{1}{(m-1)^2}-10\frac{1}{m+2}.\\ &= 2^{-\frac{k}{2}-1}\pi(-10\ln(l+1)-\frac{1}{2(l+1)}+10\ln(l+2)\\ &\quad -\frac{27}{8}\frac{1}{l+2}+\ln l-\frac{1}{l}-\ln(l+3)).\\ &= 2^{-\frac{k}{2}-1}\pi(10\ln\frac{l}{(l+1)(l+3)}-\frac{1}{l+1}-\frac{27}{8}\frac{1}{l+2}-\frac{1}{l}).\end{aligned}$$

Therefore

$$E^2_{2^k-1,l} = O(2^{-\frac{k}{2}-1}(\ln\frac{l}{(l+1)(l+3)}-\frac{1}{l+1}-\frac{27}{8}\frac{1}{l+2}-\frac{1}{l}).$$

(iv) We have

$$\begin{aligned}c_{n,m} &= <f(t),W_{n,m}>_w\\ &= 2^{\frac{k}{2}}\int_0^1 f(t)W_m(2^k t-n)\sqrt{\frac{1-t}{1+t}}dt\end{aligned}$$

$$
\begin{aligned}
&= -2^{\frac{k}{2}} \int_0^{\pi} f(\frac{\cos\theta + n}{2^k}) \frac{\sin(m+\frac{1}{2})\theta}{\sin(\frac{\theta}{2})} \sqrt{\frac{1-\cos\theta}{1+\cos\theta}} \sin\theta d\theta \\
&= -2^{\frac{k}{2}} \int_0^{\pi} f(\frac{\cos\theta + n}{2^k}) \frac{\sin(m+\frac{1}{2})\theta}{\sin(\frac{\theta}{2})} \sqrt{\frac{2\sin^2(\frac{\theta}{2})}{2\cos^2(\frac{\theta}{2})}} 2\sin\frac{\theta}{2}\cos\frac{\theta}{2} d\theta \\
&= -2^{\frac{k}{2}} \int_0^{\pi} f(\frac{\cos\theta + n}{2^k}) 2\sin(m+\frac{1}{2})\theta \sin\frac{\theta}{2} d\theta \\
&= -2^{\frac{k}{2}} \int_0^{\pi} f(\frac{\cos\theta + n}{2^k}) (-\cos(m+1)\theta + \cos m\theta) d\theta \\
&= -2^{\frac{k}{2}} [f(\frac{\cos\theta + n}{2^k})(-\frac{1}{m+1}\sin(m+1)\theta + \frac{1}{m}\sin m\theta)]_0^{\pi} \\
&\quad +2^{\frac{k}{2}} \int_0^{\pi} f'(\frac{\cos\theta + n}{2^k}) \frac{\sin\theta}{2^k} (-\frac{1}{m+1}\sin(m+1)\theta + \frac{1}{m}\sin m\theta) d\theta \\
&= 2^{\frac{k}{2}} \int_0^{\pi} f'(\frac{\cos\theta + n}{2^k}) \frac{\sin\theta}{2^k} (-\frac{1}{m+1}\sin(m+1)\theta + \frac{1}{m}\sin m\theta) d\theta \\
&\leq 2^{-\frac{k}{2}} P \int_0^{\pi} \sin\theta (\frac{1}{m+1}\sin(m+1)\theta + \frac{1}{m}\sin m\theta)) d\theta \\
&= 2^{-\frac{k}{2}} P [(\frac{1}{m+1}\sin\theta\sin(m+1)\theta + \frac{1}{m}\sin\theta\sin m\theta)]_0^{\pi} \\
&\quad -2^{-\frac{k}{2}-1} P \int_0^{\pi} \cos\theta (\frac{1}{m+1}\sin(m+1)\theta + \frac{1}{m}(\sin m\theta) d\theta \\
&= 2^{-\frac{k}{2}-1} P \int_0^{\pi} (\frac{1}{m+1}\cos\theta\sin(m+1)\theta - \frac{1}{m}\cos\theta\sin m\theta) \\
&= 2^{-\frac{k}{2}-1} P \int_0^{\pi} (\frac{1}{2(m+1)}(\sin m\theta + \sin(m+2)\theta) \\
&\quad -\frac{1}{2m}(\sin(m-1)\theta + \sin(m+1)\theta)) d\theta \\
&= 2^{-\frac{k}{2}-1} P [-\frac{1}{2m(m+1)}\cos m\theta - \frac{1}{(m+1)(m+2)}\cos(m+2)\theta \\
&\quad +\frac{1}{2m(m-1)}\cos(m-1)\theta + \frac{1}{m(m+1)}\cos(m+1)\theta]_0^{\pi} \\
&= 2^{-\frac{k}{2}-1} P [-\frac{1}{2m(m+1)}\cos m\pi - \frac{1}{(m+1)(m+2)}\cos(m+2)\pi \\
&\quad +\frac{1}{2m(m-1)}\cos(m-1)\pi + \frac{1}{m(m+1)}\cos(m+1)\pi].
\end{aligned}
$$

Therefore

$$
\begin{aligned}
|c_{n,m}| = 2^{-\frac{k}{2}-1} P | &-\frac{1}{2m(m+1)}\cos m\pi - \frac{1}{(m+1)(m+2)}\cos(m+2)\pi \\
&+\frac{1}{2m(m-1)}\cos(m-1)\pi + \frac{1}{m(m+1)}\cos(m+1)\pi|
\end{aligned}
$$

$$\leq 2^{-\frac{k}{2}-1}P|\frac{3}{2m(m+1)}+\frac{1}{(m+1)(m+2)}+\frac{1}{2m(m-1)}]$$
$$\leq 2^{-\frac{k}{2}-1}P(\frac{1}{2m}+\frac{1}{2(m+1)}+\frac{1}{m-1}-\frac{1}{m+2})$$
$$|c_{n,m}|^2 = 2^{-k-2}P^2(\frac{1}{4m^2}+\frac{1}{4(m+1)^2}+\frac{1}{(m-1)^2}+\frac{1}{(m+2)^2}$$
$$+\frac{1}{2m(m+1)}+\frac{1}{m(m-1)}+\frac{1}{m(m+2)}+\frac{1}{(m+1)(m-1)}$$
$$+\frac{1}{(m+1)(m+2)}+\frac{2}{(m-1)(m+2)})$$
$$= 2^{-k-2}P^2(\frac{1}{4m^2}+\frac{1}{4(m+1)^2}+\frac{1}{(m-1)^2}+\frac{1}{(m+2)^2}$$
$$+\frac{5}{3}\frac{1}{m-1}-\frac{5}{3}\frac{1}{m+2}),$$

$$E^2_{2^k-1,l} = \|f-s_{2^k-1,l}\|_2^2$$
$$= \|\sum_{n=0}^{\infty}\sum_{m=0}^{\infty}c_{n,m}V_{n,m}-\sum_{n=2^{k-1}+1}^{\infty}\sum_{m=0}^{l}c_{n,m}W_{n,m}\|_2^2$$
$$= \|\sum_{n=0}^{2^k-1}\sum_{m=l+1}^{\infty}c_{n,m}W_{n,m}\|_2^2$$
$$= \sum_{n=0}^{2^k-1}\sum_{m=l+1}^{\infty}|c_{n,m}|^2\|W_{n,m}\|_2^2$$
$$\leq 2^{-\frac{3k}{2}-3}\pi P^2\sum_{m=l+1}^{\infty}(\frac{1}{4m^2}+\frac{1}{4(m+1)^2}+\frac{1}{(m-1)^2}+\frac{1}{(m+2)^2}$$
$$+\frac{5}{3}\frac{1}{m-1}-\frac{5}{3}\frac{1}{m+2})$$
$$= 2^{-\frac{3k}{2}-3}\pi P^2(\frac{1}{4(l+1)}+\frac{1}{l}+\frac{1}{l+3}+\frac{5}{3}\ln\frac{l}{l+3}).$$

Therefore

$$E^2_{2^k-1,l} = O(2^{-\frac{3k}{2}-3}(\frac{1}{4(l+1)}+\frac{1}{l}+\frac{1}{l+3}+\frac{5}{3}\ln\frac{l}{l+3})).$$

□

References

[1] Çakmak, M. and Uslu, K.A. *generalization of Chebyshev polynomials with well-known kinds and transition relations*, Acta Univ. Apulensis Math. Inform. 57 (2019), 19–30.

[2] Dehghan, M. and Eslahchi, M.R. *Best uniform polynomial approximation of some rational functions*, Comput. Math. Appl. 59(1) (2010), 382–390.

[3] Eslahchi, M.R. and Dehghan, M. *The best uniform polynomial approximation to the class of the form* $\frac{1}{(a^2 \pm x^2)}$, Nonlinear Anal. 71(3-4) (2009), 740–750.

[4] Mason, J.C. and Handscomb, D.C. *Chebyshev polynomials*, Chapman and Hall/CRC, Boca Raton, FL, 2003.

[5] Nigam, H.K., Mohapatra, R.N. and Murari, K. *Wavelet approximation of a function using Chebyshev wavelets*, Thai J. Math. (2020), 197–208.

[6] Rivlin, T.J. *An introduction to the approximation of functions*, Corrected reprint of the original. Dover Books on Advanced Mathematics. Dover Publications, Inc., New York, 1981.

Aims and scope

Iranian Journal of Numerical Analysis and Optimization (IJNAO) is published twice a year by the Department of Applied Mathematics, Faculty of Mathematical Sciences, Ferdowsi University of Mashhad. Papers dealing with different aspects of numerical analysis and optimization, theories and their applications in engineering and industry are considered for publication.

Journal Policy

All submissions to IJNAO are first evaluated by the journal's Editor-in-Chief or one of the journal's Associate Editors for their appropriateness to the scope and objectives of IJNAO. If deemed appropriate, the paper is sent out for review using a single blind process. Manuscripts are reviewed simultaneously by reviewers who are experts in their respective fields. The first review of every manuscript is performed by at least two anonymous referees. Upon the receipt of the referee's reports, the paper is accepted, rejected, or sent back to the author(s) for revision. Revised papers are assigned to an Associate Editor who makes an evaluation of the acceptability of the revision. Based upon the Associate Editor's evaluation, the paper is accepted, rejected, or returned to the author(s) for another revision. The second revision is then evaluated by the Editor-in-Chief, possibly in consultation with the Associate Editor who handled the original paper and the first revision, for a usually final resolution.

The authors can track their submissions and the process of peer review via: http://ijnao.um.ac.ir

All manuscripts submitted to IJNAO are tracked by using "iThenticate" for possible plagiarism before acceptance.

Instruction for Authors

The Journal publishes all papers in the fields of numerical analysis and opti-

mization. Articles must be written in English.

All submitted papers will be refereed and the authors may be asked to revise their manuscripts according to the referee's reports. The Editorial Board of the Journal keeps the right to accept or reject the papers for publication.

The papers with more than one authors, should determine the corresponding author. The e-mail address of the corresponding author must appear at the end of the manuscript or as a footnote of the first page.

It is strongly recommended to set up the manuscript by Latex or Tex, using the template provided in the web site of the Journal. Manuscripts should be typed double-spaced with wide margins to provide enough room for editorial remarks.

References should be arranged in alphabetical order by the surname of the first author as examples below:

[1] Brunner, H. *A survey of recent advances in the numerical treatment of Volterra integral and integro-differential equations*, J. Comput. Appl. Math. 8 (1982), 213-229.

[2] Stoer, J. and Bulirsch, R. *Introduction to Numerical Analysis*, Springer-verlag, New York, 2002.

Iranian Journal of Numerical Analysis and Optimization

CONTENTS Vol. 14, No. 1, pp 1-329

Chebyshev wavelet-based method for solving various stochastic optimal control problems and its application in finance 1
M. Yarahmadi and S. Yaghobipour

A new numerical approach to the solution of the nonlinear Kawahara equation by using combined Taylor–Dickson approximation 20
D. Priyadarsini, M. Routaray and P.K. Sahu

A new approach for ranking decision-making units in data envelopment analysis by using communication game theory . 44
M. Amiri and A. Ashrafi

Efficient numerical methods on modified graded mesh for singularly perturbed parabolic problem with time delay . . . 77
A. Kumar and S. Gowrisankar

Approximate proper solutions in vector optimization with variable ordering structure 107
S. Shahbeyk

A study of global dynamics and sensitivity analysis of a discrete-time model of the COVID-19 epidemic 136
M. E.-B. Keddar and O. Belhamiti

Modal spectral Tchebyshev Petrov–Galerkin stratagem for the time-fractional nonlinear Burgers' equation 172
Y.H. Youssri and A.G. Atta

Minimization of sub-topical functions over a simplex 200
M.H. Daryaei and M.A. Yaghoobi

Richardson extrapolation technique on a modified graded mesh for singularly perturbed parabolic convection-diffusion problems 219
K. K. Sah and S. Gowrisankar

Noisy label relabeling by nonparallel support vector machine 265
A. Sahleh and M. Salahi

On generalized one-step derivative-free iterative family for evaluating multiple roots 291
H. Arora, A. Cordero and J. R. Torregrosa

Wavelet approximation with Chebyshev wavelets 315
S. M. Jesmani, H. Mazaheri and S. Shojaeian

web site: https://ijnao.um.ac.ir
Email: ijnao@um.ac.ir
ISSN-Print: 2423-6977
ISSN-Online: 2423-6969