Ferdowsi University of Mashhad

Iranian Journal of Numerical Analysis and Optimization

Volume 13, Number 2

June 2023

Serial Number: 25

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 13, Number 2, June 2023

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 13, Number 2, June 2023

Ferdowsi University of Mashhad - Iran

©2023 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

A three-free-parameter class of power series based iterative method for approximation of nonlinear equations solution . . 157
O. Ogbereyivwe and O. Izevbizua

Stability and Hopf Bifurcation in Leech Heart Interneuron Model 170
F. Parvizi, M. Razvan and Y. Alipour Fakhri

Approximate solutions to the Allen–Cahn equation using rational radial basis functions method 187
M. Shiralizadeh, A. Alipanah and M. Mohammadi

Regularization of the generalized auto-convolution Volterra integral equation of the first kind 205
S. Pishbin and A. Ebadi

The interpolated variational iteration method for solving a class of nonlinear optimal control problems 224
M. Shirazian

A robust method for optimal control problems governed by system of Fredholm integral equations in mechanics 243
A. Ebrahimzadeh

Fitted scheme for singularly perturbed time delay reaction-diffusion problems 262
M. Amsalu Ayele, A. Andargie Tiruneh and G. Adamu Derese

Modified Runge–Kutta method with convergence analysis for nonlinear stochastic differential equations with Hölder continuous diffusion coefficient 285
A. Haghighi

Two-layer shallow water formula with slope and uneven bottom solved by finite volume method 317
U. Habibah, I.R. Lina and W.M. Kusumawinahyu

An operational matrix method for solving a class of nonlinear Volterra integral equations arising in steady activation of a skeletal muscle 336
S. Torkaman, M. Heydari and G.B. Loghmani

Iranian Journal of Numerical Analysis and Optimization
Vol. 13, No. 2, 2023, pp 157–169
https://doi.org/10.22067/ijnao.2022.74901.1095
https://ijnao.um.ac.ir/

How to cite this article
Research Article

A three-free-parameter class of power series based iterative method for approximation of nonlinear equations solution

O. Ogbereyivwe*, and O. Izevbizua

Abstract

In this manuscript, for approximation of solutions to equations that are nonlinear, a new class of two-point iterative structure that is based on a weight function involving two converging power series, is developed. For any method constructed from the developed class of methods, it requires three separate functions evaluation in a complete iteration cycle that is of order four convergence. Also, some well-known existing methods are typical members of the new class of methods. The numerical test on some concrete methods derived from the class of methods indicates that they are effective and competitive when employed in solving a nonlinear equation.

AMS subject classifications (2020): Primary 65H05; Secondary 41A25.

Keywords: Newton method, Iterative method, Power series, Weight function.

1 Introduction

One famous iterative method (IM) that is effective when used for the approximation of the solution s^* of the nonlinear equation (NE) $f(s) = 0$ is the

*Corresponding author

Received 23 January 2022; revised 12 March 2022; accepted 2 July 2022

Oghovese Ogbereyivwe
Department of Mathematics and Statistics, Delta State University of Science and Technology, Delta State, Nigeria. e-mail: ogho2015@gmail.com, oghovese.ogbereyivwe@dspz.edu.ng

Osamede Izevbizua
Department of Mathematics, University of Benin, Benin City, Nigeria. e-mail: izevbizua@uniben.edu

Newton Iterative Method (NIM) [18], expressed as

$$w_{k+1} = s_k - u\left(s_k\right), \tag{1}$$

where $u\left(s_k\right) = \frac{f(s_k)}{f'(s_k)}$, $k = 0, 1, 2, \ldots$, and $\left|f'\left(s_k\right)\right| \neq 0$. The convergence order (CO) of the NIM is 2 with efficiency index (EI) 1.4142. Since the emergence of the NIM, the plethora of techniques have been employed to modify it with the motivation of enhancing its CO or EI when utilized to solve an NE. One of such techniques that have been used is the weight function, which is a power series based on the form:

$$w\left(\theta(s)\right) = 1 + \sum_{i=1}^{m} \alpha_i (\theta(s))^i \tag{2}$$

or

$$w\left(\theta(s)\right) = 1 + \sum_{i=1}^{m} \alpha_i (\mu(s) - 1)^i, \tag{3}$$

where $\alpha_i \in R$, $1 \leq m \leq \infty$ and $\theta\left(s\right), \mu\left(s\right)$ are often quotients of two real functions involving the evaluations of $f\left(\cdot\right)$ or $f'\left(\cdot\right)$.

In order to acknowledge some good literature in this direction, we note that, in the work of Khattri and Abbabandy [9], the NIM was adjusted and used as a predictor and a corrector iterative function that involves the series (2) to construct a one-parameter family of IM with CO four presented as

$$\begin{cases} y(s_k) = s_k - \frac{2}{3} u(s_k), \\ s_{k+1} = s_k - u(s_k) \left[1 + \sum_{i=1}^{3} \tau_i (v(s))^i\right], \\ \tau_1 = \frac{21}{8} - \alpha_4, \quad \tau_2 = \left(-\frac{9}{2} - 3\alpha_4\right), \quad \tau_3 = \left(\frac{15}{8} - 3\alpha_4\right), \end{cases} \tag{4}$$

where $v(s_k) = \frac{f'(y_k)}{f'(s_k)}$ and $\alpha_4$is a real parameter that is free.

In a similar manner, the power series (2) was utilized in Chun [4] to suggest the two-point IM

$$\begin{aligned} w_k &= s_k - u(s_k), \\ s_{k+1} &= s_k - \frac{f(w_k)}{f'(s_k)} \left[1 + 2\frac{f(w_k)}{f(s_k)} + \left(\frac{f(w_k)}{f(s_k)}\right)^2\right]. \end{aligned} \tag{5}$$

Ghanbari [6] used a power series like the one of (2) given as

$$s_{k+1} = w_k - \left[\frac{Af(s_k)^2 + Bf(w_k)f(s_k) + Cf(w_k)^2}{Df(s_k)^2 + Ef(w_k)f(s_k) + Ff(w_k)^2}\right] \frac{f(w_k)}{f'(s_k)}. \tag{6}$$

To develop a class of three-free-parameter CO three, the IM is obtained as

$$s_{k+1} = w_k - \left[\frac{1 + (2+\alpha)\, t_k + \sigma t_k^2}{1 + \alpha t_k + \beta t_k^2}\right] \frac{f\,(w_k)}{f'\,(s_k)}, \tag{7}$$

where $t = \frac{(w_k)}{f(s_k)}$ and $\alpha, \beta, \sigma \in R$. It was also shown in [6] that the class of methods (7) has the methods presented by Ostrowski [14], Chun [4, 3], Kou, Li, and Wang [10], and Chun and Ham [5] as concrete members.

In the work of Sharma and Bahl [16], the combination and power series of the form (2) were utilized to construct a CO four method as

$$s_{k+1} = s_k - \left[-\frac{1}{2} + \frac{9f'(s_k)}{8f'(y_k)} + \frac{3f'(y_k)}{8f'(s_k)}\right] \frac{f(s_k)}{f'(s_k)}. \tag{8}$$

Ogbereyivwe and Ojo-Orobosa [13] employed a kind of the power series (2) and an additional other weight function to put forward an iterative structure of the form:

$$s_{k+1} = w_k - \left[1 + \sum_{i=1}^{m} \alpha_i \Phi\,(t,u)^i\right] G\,(u)\,, \tag{9}$$

where $t = \frac{f(w_k)}{f(x_k)}$ and $\Phi\,(t,u) = \frac{t}{1+u}$. This iterative structure was utilized in developing a three-parameter family of IM of CO four.

In the separate work of Ahmad [1] and Babajee [2], the power series (3) was used as a weight function to develop IMs with CO four presented as

$$s_{k+1} = s_k - u(s_k)\left[1 - \frac{3\,(\mu(s_k) - 1)}{4} + \frac{9\,(\mu(s_k) - 1)^2}{8}\right], \tag{10}$$

where $\mu(s_k) = \frac{f'(y_k)}{f'(s_k)}$.

Also, Mahdu [11] used (3) to develop methods that utilize the same function evaluations to achieve order four presented as

$$s_{k+1} = s_k - u(s_k)\frac{1}{2}\,(3 - \mu(s_k))\left[1 - \frac{(\mu(s_k) - 1)}{4} + (\mu(s_k) - 1)^2\right]. \tag{11}$$

Motivated by the above literature that uses the power series of the kinds (2) or (3) as weight functions in developing some good IMs for obtaining s^* of NE, we put forward a new class of power series-based IM that has some existing class of IMs that are also power series based, as concrete members. The method uses the quotient of two power series in the form of (3) as weight functions as against one power series used in Ahmad [1] and Babajee [2]. This technique led to the development of a class of three real and free parameter IM for solving NE with CO four. The manuscript is organized in the following sequence. Section 1 discusses some literature background, while Section 2 puts forward the methodology. In Section 3, the convergence test on the method is carried out. Section 4 presents the results of a numerical

experiment on the developed method. The final section of the work includes the conclusion and suggested areas for further research.

2 The method

Consider the IM (iterative method) that is based on the terminal and convergent power series put forward as:

$$s_{k+1} = s_k - u(s_k)\left[\left(1+\sum_{i=1}^{m}\alpha_i(\mu(s_k)-1)^i\right)/\left(1+\sum_{i=1}^{m}\beta_i(\mu(s_k)-1)^i\right)\right], \tag{12}$$

where $u(s_k) = \frac{f(s_k)}{f'(s_k)}$, $\mu(s_k) = \frac{f'(y_k)}{f'(s_k)}$, α_i and $\beta_i, i = 1, 2, 3, \ldots$, are free and real parameters to be established. To illustrate the contribution of the IM (12), the case of $m = 3$ is considered for investigation and its convergence conditions. First, the following definitions are acknowledged.

Definition 1 (Asymptotic error, Asymptotic constant, and CO)**.** Let $A_j = s_j - s^*$ be the IM error at the jth iteration, and suppose that the equation $A_{j+1} = \eta A_k^{\rho} + O(A_j^{\rho+1})$ can be obtained from an IM by the use of the Taylor expansions on the functions $f(\cdot)$. Then A_{j+1} is called the Asymptotic error equation, η is Asymptotic constant, and ρ is CO.

Definition 2 (Efficiency)**.** If the equation $A_{j+1} = \eta A_j^{\rho} + O(A_j^{\rho+1})$ is derived from an IM as described in Definition 1, then the value $E_{eff} = \rho^{\frac{1}{\tau}}$ (where τ is total number of distinct functions $f(\cdot)$ in one IM cycle) is called the efficiency index of the IM.

3 Method convergence investigation

The investigation on the convergence of the method put forward in (12) for the case $m = 3$ is presented in this section. Firstly, the theorem that enables us the establishment of the method's convergence is stated and then followed by its proof.

Theorem 1. Define $f : D \subset R \to R$ as a function that is real valued such that it is differentiable for at least three times in the domain D. Suppose that $s^* \subset D$ and that $|f'(\cdot)| \neq 0$ in D. Then for an initial guess s_0 close to s^* in D, the sequence $\{s_j\}_{j\geq 0}, (s_j \in D)$ of approximations, generated by the class of IM in (12), where $m = 3$, converges to s^* with CO four whenever the free parameters α_i and β_i satisfy the following conditions: $\beta_1 - \alpha_1 + \frac{3}{4} = 0$, $\beta_2 - \frac{1}{16}(16\alpha_2 + 12\alpha_1 - 19) = 0$ and $\beta_3 - \frac{1}{64}(64\alpha_3 + 48\alpha_2 - 36\alpha_1 + 54) = 0$.

Proof. Suppose in the Taylor expansion of $f(s)$ and $f'(s)$ about s^* that we set $s{=}s_k$. Then the following expression can be obtained:

$$f(s_j) = f'(s^*)\,(A_j + \sum_{n=2}^{4} c_n A_j^n + O\left(A_j^5\right)), \tag{13}$$

and

$$f'(s_j) = f'(s^*)\,(1 + \sum_{n=2}^{5} n c_n A_j^{n-1} + O\left(A_j^5\right)), \qquad j{=}0,1,2,\ldots, \tag{14}$$

where $c_n = \frac{1}{n!}\frac{f^n(s^*)}{f'(s^*)}$, $n \geq 2$.

Using (13) and (14), the series expansions of $u(s_k)$ and $\mu(s_k)$ are, respectively,

$$u(s_k) = A_k - c_2 A_k^2 + \left(2c_2^2 - 2c_3\right) A_k^3 + \left(-4c_2^3 + 7c_2c_3 - 3c_4\right) A_k^4 + O(A_k^5), \tag{15}$$

and

$$\begin{aligned}\mu(s_k) =& 1 + \frac{4}{3}c_2 A_k + \left(4c_2^2 - \frac{8}{3}c_3\right) A_k^2 + \left(36c_2^3 - 45c_2c_3 + 13c_4\right) A_k^3 \\ &+ \left(\frac{80}{3}c_2^4 - \frac{148}{3}c_2^2c_3 + \frac{48}{27}c_4 - 5c_5\right) A_k^4 + O(A_k^5).\end{aligned} \tag{16}$$

From (15), we have

$$\begin{aligned}y_k &= A_k - \frac{2}{3}u(s_k) \\ &= \frac{1}{3}A_k + \frac{2}{3}c_2 A_k^2 - \frac{4}{3}(c_2^2 - c_3)A_k^3 + \frac{2}{3}\left(4c_2^3 - 7c_2c_3 + 3c_4\right) A_k^4 + O(A_k^5).\end{aligned} \tag{17}$$

Using (17), the expansion of the function $f(y_k)$ is obtained as

$$\begin{aligned}f(y_k) = &\frac{1}{3}A_k + \frac{7}{9}c_2 A_k^2 \\ &+ \left(\frac{4c_2^2}{9} - \frac{4(c_2^2 - c_3)}{3} + \frac{c_3}{27}\right) A_k^3 \\ &+ \left(c_2\left(\frac{4}{9}c_2^2 - \frac{8}{9}(c_2^2 - c_3)\right) + \frac{2}{9}c_2c_3 + \frac{2}{3}\left(4c_2^3 - 7c_2c_3 + 3c_4\right)\right) A_k^4 \\ &+ O(A_k^5),\end{aligned} \tag{18}$$

and using (18), its derivative $f'(y_k)$ is

$$
\begin{aligned}
f'(y_k) =& 1+\frac{2}{3}c_2A_k \\
&+\left(\frac{4}{3}c_2^2+\frac{1}{3}c_3\right)A_k^2+\left(-\frac{8}{3}c_2\left(c_2^2-c_3\right)+\frac{4}{3}c_2c_3+\frac{4}{27}c_4\right)A_k^3 \\
&+\left(\left(\frac{4}{3}c_2^2-\frac{8}{3}\left(c_2^2-c_3\right)\right)c_3+\frac{8}{9}c_2c_4+\frac{4}{3}c_2\left(4c_2^3-7c_2c_3+3c_4\right)\right)A_k^4 \\
&+O(A_k^5).
\end{aligned}
\tag{19}
$$

From the expressions in (14) and (19), the Taylor expansion of the quotient $\frac{f'(y_k)}{f'(s_k)}$ is

$$
\begin{aligned}
\frac{f'(y_k)}{f'(s_k)} =& 1+\frac{4}{3}c_2A_k \\
&+\left(4c_2^2+\frac{8}{3}c_3\right)A_k^2-\frac{8}{27}\left(36c_2^3-45c_2c_3+13c_4\right)A_k^3 \\
&+\left(\frac{80}{3}c_2^4-\frac{148}{3}c_2^2c_3+\frac{32}{3}c_3^2+\frac{484}{27}c_2c_4-5c_5\right)A_k^4 \\
&+O(A_k^5).
\end{aligned}
\tag{20}
$$

Using (16), the expansion of the quotient of the two power series in (12) is obtained as

$$
\begin{aligned}
&\left(1+\sum_{i=1}^{3}\alpha_i(\mu(s_k)-1)^i\right)/\left(1+\sum_{i=1}^{3}\beta_i(\mu(s_k)-1)^i\right) \\
&=1-\frac{4}{3}\left(\alpha_1-\beta_1\right)c_2A_k \\
&\quad+\frac{4}{9}\left(4\alpha_2c_2^2-9\beta_1c_2^2+\beta_1^2c_2^2-4\beta_2c_2^2+\alpha_1\left(\left(9-4\beta_1\right)c_2^2-6c_3\right)+6\beta_1c_3\right)A_k^2 \\
&\quad+\frac{8}{27}(8\alpha_3c_2^3-36\beta_1c_2^3+36\beta_1^2c_2^3-8\beta_1^3c_2^3+\cdots \\
&\quad+\alpha_1\left(4\left(9-9\beta_1+2\beta_1^2-2\beta_2\right)c_2^3+3\left(-15+8\beta_1\right)c_2c_3+13c_4\right))A_k^3 \\
&\quad+\frac{1}{81}(1728\alpha_3c_2^4+256\alpha_4c_2^4-2160\beta_1c_2^4-256\alpha_3\beta_1c_2^4+3600\beta_1^2c_2^4+\cdots \\
&\quad+4\left(-363+208\beta_1\right)c_2c_4+9\left(32\left(-3+2\beta_1\right)c_3^2+45c_3\right))A_k^4+O\left(A_k^5\right).
\end{aligned}
\tag{21}
$$

Now, substituting (15) and (21) into the iterative structure in (12), the following error equation is obtained:

$$\begin{aligned} s_{k+1} =& s^* + \left(c_2 + \frac{4}{3}\left(\alpha_1 - \beta_1\right) c_2\right) A_k^2 \\ &+ \frac{2}{9}((-9 - 8\alpha_2 + 8\alpha_1\left(\beta_1 - 3\right) + 24\beta_1 - 8\left(\beta_1^2 - \beta_2\right)\right) c_2^2 \\ &+ 3\left(3 + 4\alpha_1 - 4\beta_1\right) c_3) A_k^3 \\ &+ \frac{1}{27}((4(27 + 84\alpha_2 + 16\alpha_3 - 117\beta_3 - 16\alpha_2\beta_1 + 84\beta_1^2 - 16\beta_1^3 \\ &+ \alpha_1\left(117 - 84\beta_1 + 16\beta_1^2 - 16\beta_2\right) - 84\beta_2 + 32\beta_1\beta_2 - 16\beta_3) c_2^3 \\ &+ 3\left(-63 - 64\alpha_2 + 168\beta_1 - 64\beta_1^2 + 8\alpha_1\left(8\beta_1 - 21\right) + 64\beta_1\right) c_2 c_3 \\ &+ \left(81 + 104\alpha_1 - 104\beta_1\right) c_4)) A_k^4 + O\left(A_k^5\right). \end{aligned} \tag{22}$$

For the method (12) to attain order four convergence, the second and third terms of (22) must be annihilated. This can be achieved when the following set of equations are satisfied:

$$\begin{aligned} \beta_1 &= \alpha_1 + \frac{3}{4} \\ \beta_2 &= \frac{1}{16}\left(16\alpha_2 + 12\alpha_1 - 9\right) \\ \beta_3 &= \frac{1}{64}\left(64\alpha_3 + 48\alpha_2 - 36\alpha_1 + 54\right). \end{aligned} \tag{23}$$

When the conditions in (23) are substituted in (22), the error equation is obtained as

$$A_{k+1} = s^* - \left(\frac{-9c_2c_3 + c_4}{9}\right) A_k^4 + O\left(A_k^5\right). \tag{24}$$

Following Definition 1, the asymptotic error equation for the method (12) is the expression on (24), and consequently, the CO is four. □

Remark 1. By substituting the conditions (23) into (12), the class of three-free-parameter IM for solving NE is obtained as

$$\begin{cases} s_{k+1} = s_k - u\left(s_k\right)\left[\frac{1+\sum_{i=1}^{3}\alpha_i(\mu(s_k)-1)^i}{1+\sum_{i=1}^{3}\beta_i(\mu(s_k)-1)^i}\right], \\ \beta_1 = \alpha_1 + \frac{3}{4}, \quad \beta_2 = \frac{16\alpha_2+12\alpha_1-9}{16}, \\ \beta_3 = \frac{64\alpha_3+48\alpha_2-36\alpha_1+54}{64}. \end{cases} \tag{25}$$

We emphasis here that, for any free choice of $\alpha_i \in R$ in (25), a CO four method with $E_{eff} = 1.5867$ can be constructed.

Remark 2. Obtaining suitable replacement function(s) for the weight function(s) used in [17, 13, 12, 15], and many other existing methods that were developed via the weight function(s) technique, requires subjecting the replacement function(s) to satisfy certain conditions that involve its evaluation and many times differentiation at some fixed points. These procedures are tedious and cumbersome compared to the proposed three-parameter-based

class of methods (25) that requires the arbitrary substitution of values for the parameters. In fact, for any real values for the parameters, the method works.

Remark 3. It is important to note that for $m = 2$ in (12), a subclass of method (25) is obtained as

$$s_{k+1} = s_k - u(s_k)\left(\frac{1+\alpha_1(\mu(s_k)-1)+\alpha_2(\mu(s_k)-1)^2}{1+\left(\alpha_1+\frac{3}{4}\right)(\mu(s_k)-1)+\left(\frac{16\alpha_2+12\alpha_1-9}{16}\right)(\mu(s_k)-1)^2}\right). \tag{26}$$

Some existing methods and their variants are particular members of the class of methods in (26).

For instance, the method developed by Chun [4] is a concrete method of (26).

Furthermore, when $\alpha_1 = -\frac{3}{4}$ and $\alpha_2 = \frac{9}{8}$ in (26), the CO four methods by Ahmad [1] and Babajee [2] are rediscovered as

$$s_{k+1} = s_k - u(s_k)\left(1-\frac{3}{4}(\mu(s_k)-1)+\frac{9}{8}(\mu(s_k)-1)^2\right). \tag{27}$$

Again, for $\alpha_1 = \frac{3}{4}$ and $\alpha_2 = 0$ in (26), the well-known Jarratt method in [8] is rediscovered as

$$s_{k+1} = s_k - u(s_k)\left(\frac{1+\frac{3}{4}(\mu(s_k)-1)}{1+\frac{3}{2}(\mu(s_k)-1)}\right). \tag{28}$$

For some new IMs of the kind (26), the following cases are presented:

Case 1

If $\alpha_1 = -\frac{3}{4}$ and $\alpha_2 = 0$, then the following iterative structure is discovered:

$$M_1: \quad s_{k+1} = s_k - u(s_k)\left(\frac{1-\frac{3}{4}(\mu(s_k)-1)}{1-\frac{9}{8}(\mu(s_k)-1)^2}\right).$$

Case 2

The substitution of $\alpha_1 = \frac{11}{6}$ and $\alpha_2 = \frac{1}{4}$ in (26) yields the method

$$M_2: \quad s_{k+1} = s_k - u(s_k)\left(\frac{1+\frac{11}{6}(\mu(s_k)-1)+\frac{1}{4}(\mu(s_k)-1)^2}{1+\frac{31}{12}(\mu(s_k)-1)+\frac{17}{16}(\mu(s_k)-1)^2}\right).$$

Case 3

Consider $\alpha_1 = \frac{4}{3}$ and $\alpha_2 = 0$. Then a new method is obtained as

$$M_3: \quad s_{k+1} = s_k - u(s_k)\left(\frac{1+\frac{9}{16}(\mu(s_k)-1)^2}{1+\frac{3}{4}(\mu(s_k)-1)}\right).$$

Remark 4. In order to illustrate some new methods of kind (25), the following IMs are constructed for some real values of α_i:

Case 4

When $\alpha_1 = \alpha_2 = \alpha_3 = 0$, then the method M_4 is constructed as

$$s_{k+1} = s_k - u(s_k)\left(\frac{1}{1 + \frac{3}{4}(\mu(s_k) - 1) - \frac{9}{16}(\mu(s_k) - 1)^2 + \frac{27}{32}(\mu(s_k) - 1)^3}\right).$$

Case 5

Consider the case $\alpha_1 = \alpha_2 = \alpha_3 = \frac{3}{2}$. Then the IM M_5 is obtained as

$$s_{k+1} = s_k - u(s_k)\left[\frac{1 + \frac{3}{2}\left((\mu(s_k) - 1) + (\mu(s_k) - 1)^2 + (\mu(s_k) - 1)^3\right)}{1 + \frac{9}{4}(\mu(s_k) - 1) + \frac{33}{16}(\mu(s_k) - 1)^2 + \frac{21}{8}(\mu(s_k) - 1)^3}\right].$$

4 Numerical test

In this section, the methods derived from the developed class of methods are subjected to numerical tests in order to verify their effectiveness when used to obtain a solution to NE. The numerical results of the new methods are compared with the results of some existing methods that are of CO four. These include the method developed by Sharma and Bahl [16] (presented in (8)) denoted by SBM, method developed by Ogbereyivwe and Ojo-Orobosa in [13] (presented in (9)) with the parameters $\beta = 2, \eta = \sigma = 0$, denoted by OOM, the method of Hafiz and Khirallah [7], denoted as HKM, the bi-parameter method in Shams et al. [15] (given in (28)) with $\beta = 1$ and $\alpha = 2$ and denoted as SEM, the methods in Ghanbari [6] (presented in (7)) obtained by taking $\alpha = -2, \beta = 1, \theta = 4$ and $\alpha = -2, \beta = 1, \sigma = -1$, and denoted as GM1 and GM2, respectively. The metric used for comparison includes the computational CO ρ_{coc} is given by

$$\rho_{coc} = \frac{\log_{10}|f(s_{k+1})| / |f(s_k)|}{\log_{10}|f(s_k)| / |f(s_{k-1})|},$$

and $|f(s_k)|$. All computation programs for the methods were written and executed in the MAPLE 2017 software environment using Intel Celeron(R) computer with a 2GB RAM processor. The error tolerance $|f(s_k)| \leq 200^{-200}$ was used in terminating program execution. The computational outputs are maintained at 2000 digits precision in all computations. The NE $f_i(s) = 0$ used for the test and their respective solutions s^* presented in 15 digits precision are as given in Examples 1–5. For comparison, computation results are presented in Tables 1 and 2.

Consider the following examples:

Example 1. [6] $f_1(s) = s^2 - e^s - 3s + 2 = 0, \qquad s^* = 0.257530285439860\ldots.$

Example 2. [6] $f_2(s) = \sin^2(s) - s^2 + 1 = 0, \qquad s^* = 1.404491648215341\ldots.$

Example 3. [13] $f_3(s) = \arctan(s) = 0, \qquad s^* = 0.$

Example 4. [6] $f_4(s) = (s-1)^3 - 2 = 0, \qquad s^* = 2.259921049894873\ldots.$

Example 5. [13] $f_4(s) = \sqrt{\frac{1}{s}} + 2\log_{10}\left(\frac{\epsilon/D}{3.7} + \frac{2.51}{R\sqrt{s}}\right) = 0$, where $\frac{\epsilon}{D} = 10^{-4}$ and $R = 10^5$, $s^* = 0.004129319146099\ldots.$

Table 1: Methods results comparison for Examples 1– 2

Methods	s_0	$\lvert f(s_1)\rvert$	$\lvert f(s_2)\rvert$	$\lvert f(s_3)\rvert$	$\lvert f(s_4)\rvert$	$\lvert f(s_5)\rvert$	ρ_{coc}
SBM		$5.3e-4$	$7.1e-18$	$2.3e-73$	$2.7e-295$	–	4.0
HKM		$7.3e-3$	$3.4e-16$	$1.6e-66$	$7.1e-268$	-	4.0
SEM		$5.9e-3$	$1.7e-13$	$1.1e-55$	$1.7e-224$	-	4.0
OOM		$1.7e-2$	$2.9e-11$	$2.5e-46$	$1.4e-186$	-	4.0
GM1		$2.8e-4$	$3.4e-19$	$7.5e-79$	$1.7e-317$	-	4.0
GM2	-0.3	$3.0e-3$	$1.1e-14$	$1.6e-60$	$8.4e-244$	-	4.0
M1		$1.6e-4$	$4.8e-20$	$4.8e-82$	$4.3e-330$	-	4.0
M2		$2.3e-3$	$3.9e-15$	$3.1e-62$	$1.2e-250$	-	4.0
M3		$7.3e-5$	$2.3e-21$	$2.5e-87$	$3.1e-351$	-	4.0
M4		$2.1e-3$	$2.6e-15$	$5.1e-63$	$1.5e-253$	-	4.0
M5		$2.4e-3$	$4.0e-15$	$3.3e-62$	$1.6e-250$	-	4.0
SBM		$5.3e-2$	$4.5e-7$	$2.8e-27$	$4.0e-108$	$1.8e-431$	4.0
HKM		$3.7e-2$	$6.2e-8$	$5.0e-31$	$2.1e-123$	$6.3e-493$	4.0
SEM		$1.2e-2$	$1.5e-9$	$4.1e-37$	$2.2e-147$	$1.7s-588$	4.0
OOM		$1.0e-2$	$1.6e-5$	$1.7e-20$	$2.2e-80$	$3.6e-320$	4.0
GM1		$6.1e-2$	$1.0e-6$	$1.1e-25$	$1.1e-101$	$1.4e-405$	4.0
GM2	2	$1.4e-2$	$2.8e-9$	$3.9e-36$	$1.6e-143$	$4.7e-573$	4.0
M1		$5.6e-2$	$7.5e-7$	$3.0e-26$	$6.7e-104$	$1.8e-414$	4.0
M2		$6.1e-2$	$1.2e-6$	$1.9e-25$	$1.3e-100$	$2.9e-401$	4.0
M3		$5.9e-2$	$9.3e-7$	$6.8e-26$	$1.9e-102$	$1.1e-408$	4.0
M4		$2.7e-2$	$2.1e-9$	$9.8e-38$	$5.1e-151$	$3.6e-604$	4.0
M5		$1.5e-2$	$2.4e-10$	$1.9e-41$	$6.8e-166$	$1.1e-663$	4.0

From the numerical results in Tables 1 and 2, we observe that all the particular methods constructed from the developed class of methods (25) solved the test problems very well. Furthermore, the new methods' computational order of convergence displayed in the last column of Tables 1 and 2 agree with the theoretical CO established in Section 3. The methods computation results outputs are also highly competitive with the methods compared.

5 Conclusion

The new three-parameter class of IM put forward in this manuscript was shown to have a capacity of approximating the solution of NE with CO four requiring three evaluations of distinct functions in an iteration cycle.

Table 2: Methods results comparison for Examples 3– 5

Methods	s_0	$\|f(s_1)\|$	$\|f(s_2)\|$	$\|f(s_3)\|$	$\|f(s_4)\|$	$\|f(s_5)\|$	ρ_{coc}
SBM		$6.8e-3$	$2.3e-12$	$1.1e-59$	$2.2e-296$	-	4.0
HKM		$4.7e-3$	$3.7e-13$	$1.1e-63$	$2.3e-316$	-	4.0
SEM		$4.7e-3$	$5.0e-13$	$6.8e-62$	$3.2e-312$	-	4.0
OOM		$6.1e-2$	$1.5e-7$	$1.6e-35$	$2.6e-175$	-	4.0
GM1		$9.3e-3$	$1.6e-11$	$2.2e-55$	$1.1e-274$	-	4.0
GM2	-0.5	$1.3e-3$	$7.3e-16$	$4.6e-77$	$4.8e-383$	-	4.0
M1		$9.2e-3$	$1.1e-11$	$2.4e-56$	$1.2e-279$	4.0	
M2		$2.9e-3$	$3.1e-14$	$4.8e-69$	$4.4e-343$	-	4.0
M3		$3.1e-3$	$4.4e-14$	$2.7e-68$	$2.2e-339$	-	4.0
M4		$1.5e-3$	$1.4e-15$	$9.6e-76$	$1.3e-376$	-	4.0
M5		$3.1e-3$	$4.5e-14$	$3.2e-68$	$5.6e-339$	-	4.0
SBM		$2.5e-1$	$2.5e-5$	$3.8e-21$	$2.0e-84$	$1.3e-337$	4.0
HKM		$1.5e-1$	$1.8e-6$	$4.8e-26$	$2.2e-104$	$1.0e-417$	4.0
SEM		$8.2e-1$	$7.2e-2$	$4.0e-7$	$3.0e-28$	$1.0e-112$	4.0
OOM		$4.6e-1$	$7.5e-4$	$1.0e-14$	$4.0e-58$	$8.5e-232$	4.0
GM1		$2.9e-1$	$5.7e-5$	$1.3e-19$	$3.3e-78$	$1.4e-312$	4.0
GM2	3	$3.2e-2$	$1.1e-8$	$1.5e-34$	$5.9e-138$	$1.3e-551$	4.0
M1		$2.7e-1$	$4.2e-5$	$3.8e-20$	$2.6e-80$	$5.8e-321$	4.0
M2		$3.5e-2$	$2.3e-9$	$4.4e-38$	$5.8e-153$	$1.7e-612$	4.0
M3		$3.4e-2$	$6.6e-10$	$1.0e-40$	$5.2e-164$	$3.8e-657$	4.0
M4		$9.6e-2$	$9.4e-8$	$1.2e-31$	$3.5e-127$	$2.2e-509$	4.0
M5		$1.2e-2$	$3.3e-11$	$1.9e-45$	$2.1e-182$	$3.2e-730$	4.0
SBM		$3.7e-1$	$1.4e-5$	$2.6e-23$	$3.7e-94$	$1.5e-377$	4.0
HKM		$1.6e-3$	$8.5e-16$	$7.6e-65$	$4.3e-261$	-	4.0
SEM				Failed to converge			
OOM		$6.1e-1$	$3.4e-4$	$2.8e-17$	$1.4e-69$	$8.7e-279$	4.0
GM1		$4.5e-1$	$3.7e-5$	$2.0e-21$	$1.6e-86$	$6.8e-347$	4.0
GM2	0.002	$2.9e-1$	$1.5e-5$	$8.0e-23$	$7.3e-92$	$5.2e-368$	4.0
M1		$4.1e-1$	$2.9e-5$	$7.8e-22$	$4.3e-88$	$4.0e-353$	4.0
M2		1.0509	$8.1e-4$	$2.4e-16$	$1.9e-66$	$7.4e-267$	4.0
M3		$5.1e-1$	$3.3e-5$	$4.9e-22$	$2.6e-89$	$2.0e-358$	4.0
M4		$1.2e-1$	$1.4e-7$	$1.9e-31$	$8.1e-127$	$2.4e-508$	4.0
M5		$7.3e-1$	$1.7e-4$	$5.0e-19$	$3.6e-77$	$9.7e-310$	4.0

Several famous existing methods, such as Chun [4], Ahmad [1], Babajee [2], Jarrat [8], and others, are typical members of the class of developed methods. For future research, the implementation of the method developed herein in Banach spaces may be considered.

Conflicts of Interest

There is no existence of any form of conflict between the authors as per the publication of this manuscript.

Acknowledgment

The authors are thankful to the Editor and anonymous reviewers for their constructive comments that led to the improvement of the manuscript.

Funding

The first author is supported by TETFund Nigeria via the Institution Based Research (IBR) Fund.

References

[1] Ahmad, F., *Comment on: On the Kung-Traub conjecture for iterative methods for solving quadratic equations,* Algorithms 2016, 9, 1. Algorithms (Basel) 9(2) (2016), Paper No. 30, 11 pp.

[2] Babajee, D.K.R., *On the Kung-Traub conjecture for iterative methods for solving quadratic equations*, Algorithms (Basel) 9(1) (2016), Paper No. 1, 16 pp.

[3] Chun, C., *Some variants of King's fourth-order family of methods for nonlinear equations*, Appl. Math. Comput. 190, (2007), 57–62.

[4] Chun, C., *Some fourth-order iterative methods for solving nonlinear equations*, Appl. Math. Comput. 195(2) (2008) 454–459.

[5] Chun, C. and Ham, Y.M., *Some fourth-order modifications of Newtons method*, Appl. Math. Comput. 197, (2008), 654–658.

[6] Ghanbari, B., *A new general fourth-order family of methods for finding simple roots of nonlinear equations*, J. King Saud Univ. Sci. 23(4), (2011) 395–398.

[7] Hafiz, M.A. and Khirallah, M.Q., *An optimal fourth order method for solving nonlinear equations*, J. Math. and Computer Sc., 23,(2021), 86–97.

[8] Jarratt, P., *A review of methods for solving nonlinear algebraic equations in one variable,* Numerical methods for nonlinear algebraic equations (Proc. Conf., Univ. Essex, Colchester, 1969), pp. 1–26. Gordon and Breach, London, 1970.

[9] Khattri, S.K. and Abbasbandy, S., *Optimal fourth order family of iterative methods,* Mat. Vesnik 63(1) (2011) 67–72.

[10] Kou, J., Li, Y. and Wang, X., *A composite fourth-order iterative method for solving non-linear equations*, Appl. Math. Comput. 184(2) (2007) 471–475.

[11] Mahdu, K., *Two-point iterative methods for solving quadratic equations and its applications,* Math. Sci. Appl. E-Notes 6(2) (2018) 66–80.

[12] Ogbereyivwe, O. *High order quadrature based iterative method for approximating the solution of nonlinear equations*, Caspian Journal of Mathematical Sciences, 9 (2), (2020), 243–255.

[13] Ogbereyivwe, O. and Ojo-Orobosa, V., *Family of optimal two-step fourth order iterative method and its extension for solving nonlinear equations*, Journal of Interdisciplinary Mathematics, 24 (5), (2021), 1347–1365.

[14] Ostrowski, A.M., *Solution of equations in Euclidean and Banach spaces,* Third edition of Solution of equations and systems of equations. Pure and Applied Mathematics, Vol. 9. Academic Press [Harcourt Brace Jovanovich, Publishers], New York-London, 1973.

[15] Shams, M., Mir, N. A., Rafiq, N., Almatroud, A. O., and Akram, S., *On dynamics of iterative techniques for nonlinear equation with applications in engineering,* Math. Probl. Eng. 2020, Art. ID 5853296, 17 pp.

[16] Sharma, R., and Bahl, A., *An optimal fourth order iterative method for solving nonlinear equations and its dynamics,* J. Complex Anal. 2015, Art. ID 259167, 9 pp.

[17] Soleymani F., Khattri S.K. and Karimi V. S., *Two new classes of optimal Jarrat type fourth-order methods*, Appl. Math. Lett. **25**, (2012), 847–853.

[18] Traub, J.F., *Iterative methods for the solution of equations,* Prentice Hall, New York, 1964.

How to cite this article
Ogbereyivwe, O. and Izevbizua, O., A three-free-parameter class of power series based iterative method for approximation of nonlinear equations solution. *Iran. j. numer. anal. optim.*, 2023; 13(2): 157-169. https://doi.org/10.22067/ijnao.2022.74901.1095

Iranian Journal of Numerical Analysis and Optimization
Vol. 13, No. 2, 2023, pp 170–186
https://doi.org/10.22067/ijnao.2022.75087.1101
https://ijnao.um.ac.ir/

How to cite this article
Research Article

Stability and Hopf bifurcation in leech heart interneuron model

F. Parvizi*, M. Razvan and Y. Alipour Fakhri

Abstract

This article investigates the activity regimes of a realistic neuron model (as a slow-fast system). The authors study this model using the dynamical systems theory, for example, qualitative theory methods of slow-fast systems. The authors obtain the stability conditions of equilibria in leech heart interneurons under defined pharmacological conditions and following Hodgkin–Huxley formalism. Although in neuronal models, the membrane is usually considered capacitance as a fixed parameter, the membrane capacitance parameter is assumed as a control parameter to guarantee the existence of Hopf bifurcation using the Routh–Hurwitz criteria. The authors investigate the transition mechanism between the silent phase and tonic spiking mode. Furthermore, some simulations are provided using XPPAUT software for analytical results.

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

Keywords: Stability; Hopf bifurcation; Routh–Hurwitz criteria.

*Corresponding author

Received 2 February 2022; revised 28 May 2022; accepted 20 July 2022

Fatemeh Parvizi
Department of Mathematics, Payame Noor University(PNU), Tehran, Iran. e-mail: Parvizifa@gmail.com

Mohammadreza Razvan
Department of Mathematics, Faculty of Mathematical Sciences, University of Sharif, Tehran, Iran. e-mail: razvan@sharif.ir

Yousef Alipour Fakhri
Department of Mathematics,, Payame Noor University(PNU), Tehran, Iran. e-mail: $y_$alipour@pnu.ac.ir

1 Introduction

Periodic orbits play an important role in nonlinear dynamical models, especially in computational neuroscience from the point of view of their topology. Now one of the models that pay attention to researchers, is neuron models whose topology is associated with specific and regular activities such as tonic spiking, bursting, and resting.

Most neurons show potential membrane oscillations as endogenous or in terms of external perturbations. Many kinds of activities on neuronal models based on models following the formalism of Hodgkin–Huxley were described in terms of the qualitative theory of fast-slow systems [1, 11]. Fast-slow systems are a special class of dynamical systems with at least two distinct time scales. The solutions of fast-slow systems possess features that are constrained to stay near the slow-motion manifolds, composed of equilibria and periodic orbits of the fast subsystem. If both manifolds are transient for the solutions of the corresponding neuron model, then it exhibits a bursting behavior, a repetitive alternation of tonic spiking and quiescent periods. Otherwise, the model indicates the tonic spiking activity if there is a stable periodic orbit on the tonic spiking manifold, or it shows no oscillations when solutions are attracted to a stable equilibrium state on the quiescent manifold [12].

Classifying transition mechanisms between various regimes is a fundamental problem of the theory of dynamical systems and neuroscience. Real neurons can show various types of firing patterns, such as tonic spiking, bursting oscillations, and silent states, that are frequently observed in neuronal electrophysiological experiments. We investigate the stability of the silent phase and the existence of periodic solutions in the Leech heart interneuron model.

Malashchenko, Shilnikov, and Cymbalyuk [14] described that the leech heartbeat is one of the best-studied invertebrate neuronal networks with an identified function, a set of identified participating neurons, and well-developed biophysically accurate models. It consists of a small number of interneurons distributed over several ganglia. Those located in ganglia 3 and 4 are responsible for generating basic rhythm [2]. Here we focus on the dynamics of a single interneuron from either ganglion 3 or 4. The canonical model has proved itself as a powerful tool for predicting phenomena.
This model has already been studied by Shilnikov and Cymbaluk [19] and Wanga et al. [21]. They consider different parameters, for example, V_{K2}^{shift} or E_l, g_l, as a control parameter with different initial conditions and investigated the existence of different codimension-one or two bifurcations. Researchers found the parameter regimes in which bursting, tonic spiking, and silence are stable. They located several kinds of multi-stability in which more than one activity is stable, such as bi- stability in which bursting coexistent with tonic spiking, bursting coexistent with silence, and tonic spiking with silence.

Cymbalyuk et al. [4] have examined the single heart interneuron model. Using extracellular recording methods, they demonstrated that oscillator and

premotor cardiac interneurons continue to burst when pharmacologically isolated with bicuculline, albeit in certain preparations, the bursting is not strong. They performed a bifurcation analysis of their model using the leak conductance g_l and reversal potential E_l as parameters. The bifurcation diagram shows a narrow strip of parameter values, where bursting occurs, separating large zones of tonic spiking and silence. Moreover, the authors found multi-stability zones, which are regions where the major activities coexist. They propounded how robustness is achieved in this system, how critical variables such as period and duty cycle may be controlled, and how modulation may affect this control.

Shilnikov and Cymbalyuk [19] showed how two different codimension-one bifurcations of a saddle-node periodic orbit with homoclinic orbits could explain a scenario of transitions between tonic spiking and bursting activities in neuron models following Hodgkin–Huxley formalism. They also found the co-existence of tonic spiking and bursting modes, which are separated by a saddle periodic orbit. This mechanism is based on a Lukyanov–Shilnikov bifurcation for a saddle-node periodic orbit with noncentral homoclinic orbits. They showed that as a bifurcation parameter is varied, the unstable periodic orbit separates the basins of attraction of two co-existing modes: the tonic spiking periodic orbit and bursting regime.

Cymbalyuk and Shilinkov [5] presented conditions under which this model demonstrates another kind of bi-stability with two coexistent tonic spiking modes. They developed a geometrical framework for Pontryagin's averaging method of singularly perturbed systems and located periodic orbits and studied their bifurcations. Pontryagin's averaging method gives a clear geometrical interpretation of this mechanism. They showed that as a bifurcation parameter is varied, the unstable periodic orbit separates the basins of attraction of two co-existing modes: tonic spiking regime coexisting with another spiking regime.

Furthermore, by synchronization of Wanga et al. [21] in gap-junction, coupled neurons with co-existing attractors of spiking and bursting firings are investigated as the coupling strength gets increased for this model. Kolomiets and Shilnikov [12] considered this model as an example to investigate generic mechanisms of transition between distinct patterns of activity in it. They considered neuronal bursting as a modular activity consisting of different limiting branches corresponding to oscillatory and equilibrium regimes from the fast subsystem, based on the Poincaré return map to analyze nonlocal bifurcation to understand better and determine what makes bursting and spiking attractors fluctuate their shape and stability.

The lipid bilayer of the cell membrane is a dielectric that creates a capacitance by separating charge inside and outside the cell. Such capacitance behavior is represented in electrical equivalent circuit models of the cell membrane with an electrical capacitor element [16]. The membrane-specific capacitance is considered constant and ubiquitous [6]. Moreover, myriad models of excitable cells and axons incorporate a constant membrane capacitance

[16]. However, the dielectric properties of living tissues exhibit frequency-dependent behavior [7].
Howell et al. [10] checked out the effect of a frequency-dependent capacitance on neuronal excitability. They found that the effect of dielectric dispersion on the sinusoidal stimulation thresholds depended on the electrical properties of the neuronal membrane. Moreover, dielectric dispersion increased action potential conduction velocity. When analyzing the response qualities of brain elements at various frequencies, it is necessary to account for the influence of frequency-dependent capacitance . Furthermore, in this paper, we consider the membrane capacity as a control parameter in the neuronal model. To the best of our knowledge, this is the first work that shows how to occur Hopf bifurcation in this model when the membrane capacitance parameter varies.

The concept of Hopf bifurcation, which is known as the Poincaré–Andronov–Hopf bifurcation, seems complicated and abstract. However, this theory applies to many real problems. Furthermore, the application of Hopf bifurcation theory is quite wide. This theory is applied in tests of various oscillations caused by wind gusts (which is very important in the field of construction), LCR oscillations in electrical circuits, transitions among neuronal activity regimes in neuronal models, periodic generation of nerve impulses in the nervous system, as well as, for instance in epidemiological models that describe fluctuations in the number of patients with an infectious disease. Through experimental measurements, the researchers found that the membrane capacity decreases with increasing frequency.

In many neuronal models, Hopf bifurcation occurs by changing various parameters. Using computer software, the Hopf bifurcation can be determined in multidimensional dynamical systems that depend on certain parameters. The presence of Hopf bifurcation is proved, and the bifurcation parameter value is estimated using the Routh–Hurwitz criterion, which determines the difference of this research.

This manuscript considers a model of a pharmacologically isolated heartbeat interneuron when the membrane capacitance is the control parameter. Then, the transition mechanism between the silent phase and tonic spiking mode is investigated. Hopf bifurcation occurs when a periodic solution or boundary cycle, which surrounds the equilibrium point, appears or disappears with a change in the parameter value. This type of Hopf bifurcation occurs where a pair of complex conjugate eigenvalues of the Jacobian matrix passes through the imaginary axis while all other eigenvalues have negative real parts. Furthermore, the existence of this type of Hopf bifurcation leads limit cycle losing stability and disappearing when changing a control parameter. First, we find that such a region for the existence of periodic solutions and Hopf bifurcation is being investigated. The type of bifurcation can be determined analytically, but it is very complicated. However, using computers (numerical experiments), we determine the type of bifurcation much faster and easier. The trajectories of this model were obtained using Dormand–Prince solvers, which are related to Runge–Kutta and are adaptive method

of solving ODE. The integration and bifurcation analysis was performed utilizing Mathematica and XPPAUT. These simulations present our theoretical results.

The manuscript is organized as follows: first, we provide some necessary preliminaries. Then we introduce the neuronal model and specify the parameters and the initial condition. In Section 3, we investigate a general geometrical framework for the analysis of (equilibrium points of a system) solutions to slow-fast dynamical systems. In the last section, we present some simulations to illustrate analytical results.

2 Preliminaries

2.1 Routh–Hurwitz criterion for the existence of Hopf bifurcation

In this section, an important criterion, which is known as the Routh–Hurwitz criterion, is presented that gives necessary and sufficient conditions for all roots of the characteristic equation that real parts of eigenvalues are negative. This criterion is stated in the next theorem; see [20, 17].

Theorem 1 (Routh–Hurwitz criterion)**.** Consider a polynomial as follows:

$$\alpha_k x^k + \alpha_{k-1} x^{k-1} + \alpha_{k-2} x^{k-2} + \cdots + \alpha_1 x + \alpha_0,$$

when $\alpha > 0$, the polynomial has all roots with negative real parts if and only if all the leading principal minors of the $k \times k$ matrix

$$A = \begin{pmatrix} \alpha_1 & \alpha_0 & \cdots & 0 \\ \alpha_2 & \alpha_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{2k-1} & \alpha_{2k-2} & \cdots & \alpha_k \end{pmatrix}$$

are positive and $\alpha_k > 0$. We assume that $\alpha_j = 0$ if $j < 0$ or $j > k$.

Let us consider a system

$$\dot{x} = f(x, \mu), \qquad x \in \mathbb{R}^n, \qquad \mu \in \mathbb{R}^1, \tag{1}$$

with an equilibrium (x_0, μ_0) and $f \in C^\infty$.
There is another theorem, which shows the existence of Hopf bifurcation in system (1).

Theorem 2. Assume that system (1) has an equilibrium $(x_0; \mu_0)$ at which the following properties are satisfied:

(SH1) $D_x f_{\mu_0}(x_0)$ has a simple pair of pure imaginary eigenvalues and other eigenvalues have negative real parts.

Thus, there exists a smooth curve of equilibria $(x(\mu);\mu)$ with $x(\mu_0) = x_0$. The eigenvalues $\lambda(\mu)$; $\bar{\lambda}(\mu)$ of $D_x f_{\mu_0}(x(\mu))$ which are imaginary at $\mu = \mu_0$ vary smoothly with μ. In addition, if

(SH2) $d(Re(\lambda(\mu))/d\mu \neq 0$,

then a simple Hopf bifurcation will occur.

Proof. See [8]. □

Moreover, researchers can apply n principal sub-determinants to obtain a parameter value for Hopf bifurcation Theorem. It should be noted that the characteristic polynomial of the Jacobian matrix $J(\mu)$ of system (1) is denoted as follows:

$$P(\lambda;\mu) = det(\lambda I_n - J(\mu)) = p_0(\mu) + p_1(\mu)\lambda + \cdots + p_n(\mu)\lambda^n,$$

where every $p_i(\mu)$ is a smooth function of μ, and $p_n(\mu) = 1$. We consider the case $p_0(\mu) > 0$, because there is not any nonnegative real root. Let

$$L_n(\mu) = \begin{pmatrix} p_1(\mu) & p_0(\mu) & \cdots & 0 \\ p_2(\mu) & p_1(\mu) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ p_{2n-1}(\mu) & p_{2n-2}(\mu) & \cdots & p_n(\mu) \end{pmatrix}$$

where $p_i(\mu) = 0$ if $i < 0$ or $i > n$. Moreover, when $p_0(\mu) > 0$, according to the Routh–Hurwitz criterion, the polynomial $P(\lambda;\mu)$ of λ has all roots with negative real parts if and only if the following n principal subdeterminants of $L_n(\mu)$ are positive:

- $D_1(\mu) = det(L_1(\mu)) = p_1(\mu) > 0$,
- $D_2(\mu) = det(L_2(\mu)) = det \begin{pmatrix} p_1(\mu) & p_0(\mu) \\ p_3(\mu) & p_2(\mu) \end{pmatrix} > 0$,

  $\vdots$
- $D_n(\mu) = det(L_n(\mu)) > 0$.

Since $D_n(\mu) = p_n(\mu)D_{n-1}(\mu)$ and for characteristic equation $p_n(\mu) = 1$, the Routh–Hurwitz criterion conditions can be formulated as

$$p_0(\mu) > 0, D_1 > 0, D_2 > 0, \ldots, D_{n-1} > 0.$$

Theorem 3. Suppose that there is a smooth curve of equilibria $(x(\mu), \mu)$ with $x(\mu_0) = x_0$ for (1). Then conditions $(SH1)$ and $(SH2)$ for a simple Hopf bifurcation are equivalent to the following conditions on the coefficients of the characteristic polynomial $P(\lambda; \mu)$:

1. $p_0(\mu_0) > 0; D_1(\mu_0) > 0; \ldots; D_{n-2}(\mu_0) > 0; D_{n-1}(\mu_0) = 0,$
2. $\frac{dD_{n-1}(\mu_0)}{d\mu} \neq 0.$

Proof. see [13]. □

2.2 Neuron model

Because a complete study of a neuronal model including all the currents identified experimentally is very complex, we study the dynamics of a reduction of the canonical leech heart interneuron model. Depending on the parameter values, the model may display several regimes, including tonic spiking, bursting, and quiet phase. These regimes represent the intricacy of the dynamics of diverse membrane ionic currents acting on distinct time scales. These neurons are particularly attractive for analysis in terms of the theory of dynamical systems since their membrane ionic currents were measured via voltage-clamp experiments and well described by a canonical model using the Hodgkin–Huxley formalism. We focus on the dynamics of a single interneuron when isolated pharmacologically from the rest of the network. In these neurons, eight voltage-dependent ionic currents have been well identified and characterized [18, 9]. These currents are separated into four groups and classified by their ionic specificity in Table 1. Moreover, "C.N." is the abbreviation of current names. The model equations for I_{Na} current were

Table 1: Ionic currents in leech heart interneuron model

Ion	C.N.	Description	Method of quantify
Na	I_{Na}	fast sodium current	Hodgkin–Huxley
	I_{NaP}	persistent sodium current	Voltage clamp
K	I_{K1}	delayed rectifier-like potassium current	Voltage clamp
	I_{K2}	persistent potassium current	
	I_{Ka}	fast transient potassium	
Ca	I_{CaF}	rapidly inactivating current	Voltage clamp
	I_{CaS}	slowly inactivating current	
Na, K	I_h	hyperpolarization activated current	Voltage clamp

borrowed from the original work by Hodgkin and Huxley, which adjusted for leech kinetics. These currents are independent of the intracellular concentration of any particular ion. A canonical model of a single neuron was

described by a system of 14 differential equations running at multiple time scales which vary from a few milliseconds through seconds. As pointed out before, a complete investigation of this concept would be exceedingly tough and difficult. By blocking clusters of currents in live cardiac interneurons and exploiting their distinctive characteristics, neuronal dynamics are simplified. From the standpoint of the theory of dynamical systems, these distinctive behaviors provide intriguing phenomena for investigation [19].
Shilnikov and Cymbalyuk [19] removed from the 14-D canonical model developed in [9] the equations and terms describing blocked currents: I_{CaF}, I_{CaS}, and I_h. For simplicity, they assumed that the partial block of outward currents completely removes I_{K1}, as well as I_{KA}, and the current I_{NaP} is ignored, whereas it reduces I_{K2}. Now, the neuron can be described by the model based on just two currents activation of I_{K2} as m_{K2} and inactivation of I_{Na} as h_{Na} and V as the membrane potential. Therefore the resulting model described in [3] is as follows:

$$\begin{cases} CV' = -(\bar{g}_{K2}m_{K2}^2(V-E_K) + g_l(V-E_l) \\ \qquad +\bar{g}_{Na}f(-150,0.0305,V)^3 h_{Na}(V-E_{Na})), \\ m'_{K2} = \frac{f(-83,0.018+V_{K2}^{shift},V)-m_{K2}}{\tau_{K2}}, \\ h'_{Na} = \frac{f(500,0.03391,V)-h_{Na}}{\tau_{Na}}, \end{cases} \tag{2}$$

where the variables V, m_{K2}, and h_{Na} display as membrane potential, activation of I_{K2}, and an inactivation of I_{Na}, respectively.

Moreover, f is a Boltzman function: $f(x,y,z) = \frac{1}{(1+e^{x(y+z)})}$ and other

Table 2: Electrical properties of the leech heart interneuron Model

Parameter	Description	Unit(s)	Value
C	Membrane capacitance	nF	B.P.
$\bar{g}_{Na}$	Maximum conductance of I_{Na}	nS	200
$\bar{g}_{K2}$	Maximum conductance of I_{K2}	nS	30
g_l	Leakage conductance of current	nS	8
E_{Na}	Sodium reversal voltage	mV	45
E_{K2}	Potassium reversal voltage	mV	-70
E_l	Leakage reversal voltage	mV	-46
V_{K2}^{shift}	Shift of the membrane potential of half-inactivation of I_{K2} from its canonical value	mV	-25.98
τ_{Na}	Time constants of inactivation of I_{Na}	S	0.0405
τ_{K2}	Time constants of activation of I_{K2}	S	0.9

parameters are described in Table 2. As we noted before, we use C as a bifurcation parameter or (B.P.) in system (2).

3 Main results

Now, we replace the parameter values are given in Table 2 in system (2) and compute its equilibrium points. These points are as follows:

$$q_1 = (0.004327, 0.424776, 0), \qquad q_2 = (-0.02793, 0.04831, 0.04788).$$

To investigate the dynamical behavior of equilibrium points, we need to get the Jacobian matrix of system (2) at these points. For this aim, we obtain the Jacobian matrix of system (2) as follows:

$$\begin{pmatrix} -\frac{200z}{(a+1)^3} - \frac{90000a(x-0.045)z}{(a+1)^4} - 30y^2 - 8 & -60(x+0.07)y & -\frac{200(x-0.045)}{(a+1)^3} \\ \frac{92.2222Ce^{-83(x-0.00798)}}{\left(e^{-83(x-0.00798)}+1\right)^2} & -1.11111C & 0 \\ -\frac{12345.7Ce^{500(x+0.03391)}}{\left(e^{500(x+0.03391)}+1\right)^2} & 0 & -24.6914C \end{pmatrix},$$

where $(V, m_{K2}, h_{Na}) = (x, y, z)$ and $a = e^{-150(x+0.0305)}$. The Jacobian matrix and the characteristic equation of system 2 at q_1 and q_2 are computed below

- at q_1

$$A_{q_1}(C) = \begin{pmatrix} -13.4131 & -1.89443 & 8.00404 \\ 22.5345C & -1.11111C & 0 \\ -0.0000613486C & 0 & -24.6914C \end{pmatrix},$$
$$P_{q_1}(\lambda, C) = 1422.06C^2 + (27.4349C^2 + 388.782C)\lambda + (13.4131 + 25.8025C)\lambda^2 + \lambda^3,$$

- at q_2

$$A_{q_2}(c) = \begin{pmatrix} 16.7353 & -0.121944 & 3.07555 \\ 4.24029C & -1.11111C & 0 \\ -562.805C & 0 & -24.6914C \end{pmatrix},$$
$$P_{q_2}(\lambda, c) = 1476.9C^2 + (1299.64C + 27.4349C^2)\lambda + (25.8025C - 16.7353)\lambda^2 + \lambda^3.$$

Now, we investigate the stability of equilibria by the Routh–Hurwitz criterion. Therefore we need to compute $L_i(C), i = 1, 2$ and find where the coefficients of the characteristic polynomials are positive. In the following, we denote the coefficients of $P_{q_i}(\lambda, C)$ by P_{ij}, where j shows the power of λ parameter. Hence the Routh–Hurwitz criterion holds for $P_{q_1}(\lambda, C)$ if

$$P_{10} = 1422.06C^2 > 0;$$
$$P_{11} = 27.4349C^2 + 388.782C > 0;$$
$$P_{12} = 25.8025C + 13.4654 > 0;$$
$$P_{13} = 1 > 0;$$

- $D_1(C) = det(L_1(C)) = p_1(C) = 27.4349C^2 + 388.782C > 0;$

- $D_2(C) = det(L_2(C))$
$$= det \begin{pmatrix} 27.4349C^2 + 388.782C & 1422.06C^2 \\ 1 & 25.8025C + 13.4654 \end{pmatrix} > 0.$$

We have similar conditions for $P_{q_2}(\lambda, C)$ as follows:

$$P_{20} = 1476.9C^2 > 0;$$
$$P_{21} = 27.4349C^2 + 1299.64C > 0;$$
$$P_{22} = 25.8025C - 16.7353 > 0;$$
$$P_{23} = 1 > 0;$$

- $D_1(C) = det(L_1(C)) = p_1(C) = 27.4349C^2 + 1299.64C > 0;$

- $D_2(C) = det(L_2(C))$
$$= det \begin{pmatrix} 27.4349C^2 + 1299.64C & 1476.9C^2 \\ 1 & 25.8025C - 16.7353 \end{pmatrix} > 0.$$

By the above computations, we have the following theorem.

Theorem 4. Let $q_1 = (0.00433, 0.42478, 0)$ and $q_2 = (-0.02793, 0.04831, 0.04788)$ be the equilibrium points of system (2). Then

1. q_1 is unstable for $0 < C \leq 1$,
2. q_2 is stable for $C > 0.678033$ and unstable for $C < 0.678033$.

Proof. It is trivial. □

If we put $C_0 = 0.678033$ at $q_2 = (-0.02793, 0.04831, 0.04788)$, then we get the following relations:

$$P_{20}(0.678033) = 1476.9(0.678033^2) > 0;$$
$$P_{21}(0.678033) = 27.4349(0.678033^2) + 1299.64(0.678033) > 0;$$
$$D_2(0.678033) = det \begin{pmatrix} 27.4349C_0^2 + 1299.64C_0 & 1476.9C_0^2 \\ 1 & 25.8025C_0 - 16.7353 \end{pmatrix}$$
$$= 707.889C_0^2 + 31597.9C_0 - 21749.9 \simeq 0;$$

$$\frac{dD_2(C)}{dC}\Big|_{C=C_0} = \frac{d(707.889C^2 + 31597.9C - 21749.9)}{dC}\Big|_{C=0.678033}$$
$$= 1415.778(0.678033) + 31597.9 \neq 0.$$

Therefore, according to these results we can state the following theorem.

Theorem 5. Suppose that for $C_0 = 678033$ and equilibrium point $q = q_2$ of system (2) the following relations satisfy:

1. $P_{20}(C_0) = 1476.9C_0^2 > 0$;
2. $P_{21} = 27.4349C^2 + 1299.64C_0 > 0$;
3. $D_2(C_0) \simeq 0$;
4. $\frac{dD_2(C)}{dC}\Big|_{C=C_0} \neq 0$.

Then C_0 is a simple Hopf bifurcation value for system (2) at the equilibrium point $q = q_2$.

4 Numerical simulation

In this section, several numerical simulations are presented to verify the efficiency of our analytical results. In addition, we draw the bifurcation diagram that shows periodic orbits crossing the plane $V = -0.025$. The simulation parameter values are shown in Table 3. We study the stability of the system's equilibrium points for particular C parameter values satisfying Theorems 4 and 5.

If we consider the starting point $p_1 = (0.0043, 0.42, 0.0001)$ near equilibrium point q_1, for $C = 0.64$, then it is observed that the equilibrium point is unstable; see Figure 1.

Table 3: C parameters values and initial values

C	Initial value	Figure
0.64	(0.004,0.413,0.001)	Figure 1
0.685	(-0.0266,0.0478,0.0475)	Figure 2
0.6	(-0.0266,0.0478,0.0475)	Figure 3

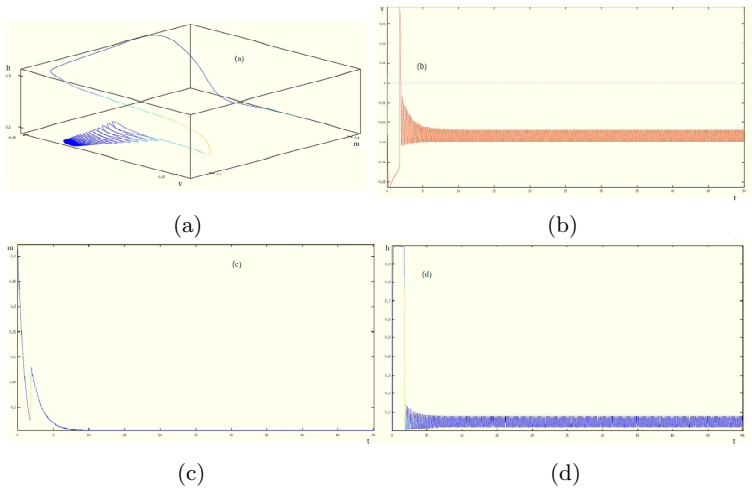


Figure 1: (a). Phase portrait in the space (V, m_{K2}, h_{Na}) at $(0.0043, 0.42, 0.0001)$ and $C = 0.64$. (b), (c), and (d). Time history of the neuronal model in terms of V, m_{K2} and h_{Na}, respectively.

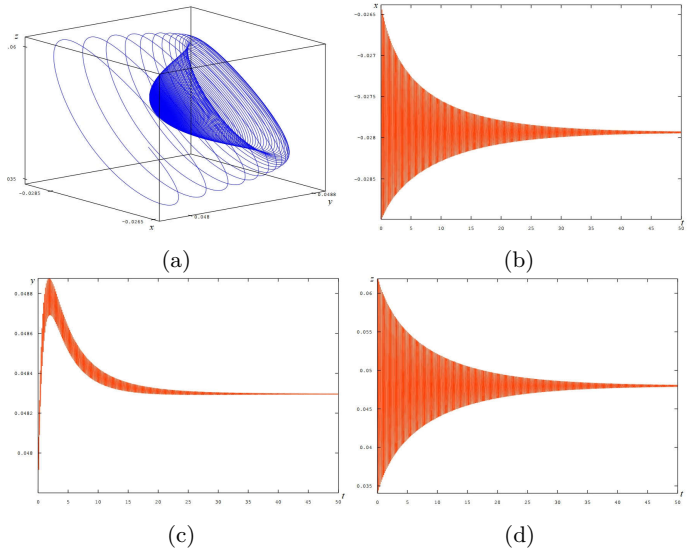


Figure 2: (a). Phase portrait in the space (V, m_{K2}, h_{Na}) at $(-0.0266, 0.0478, 0.0475)$ and $C = 0.685$. (b), (c), and (d). Time history of the neuronal model in terms of V, m_{K2} and h_{Na}, respectively.

Now suppose that the starting point $p_2 = (-0.0266, 0.0478, 0.0475)$ near equilibrium point q_2, for parameter $C = 0.64$. Then it is observed that the equilibrium point is unstable; see Figure 2.

Therefore, by Theorem 5, $C_0 = 0.678033$ is a Hopf bifurcation value

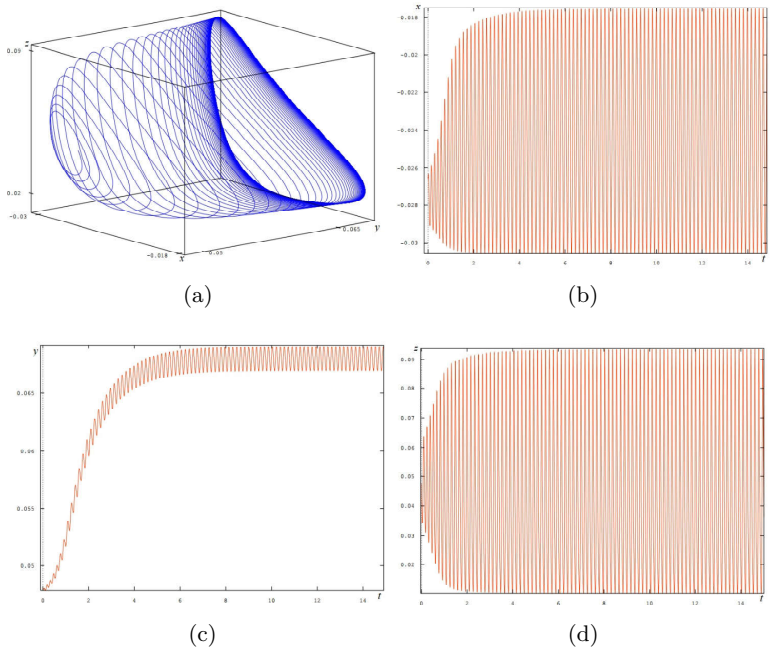


Figure 3: (a). Phase portrait in the space (V, m_{K2}, h_{Na}) at $(-0.0266, 0.0478, 0.0475)$ and $C = 0.6$. (b), (c), and (d). Oscillatory waveforms generated by the neuron model (2) in terms of V, m_{K2} and h_{Na}, respectively.

for equilibrium q_2 of leech heart interneuron model. Furthermore, for C_0, the Jacobian matrix $A_{q_2}(C_0)$ has a pair of pure imaginary eigenvalues $\lambda_1; \lambda_2 \simeq \pm 29.842209i$, and a negative real eigenvalues $\lambda_3 = -0.762921$.
We solved equation (2) by means of the fourth and fifth-order Runge–Kutta (RK-45) method. To show the occurrence of Hopf bifurcation, we present some numerical simulations. The related trajectories in the phase space (V, m_{K2}, h_{Na}) and the oscillatory waveforms generated by the neuronal model (2) for different values of parameter C are presented in Figures 3 and 4.
In Figure 5, we show that by considering a fixed point and changing C, the periodic orbit at the bifurcation value disappeared. As it is illustrated in Figure 4, the equilibrium point

$$q_2 = (-0.02793, 0.04831, 0.04788)$$

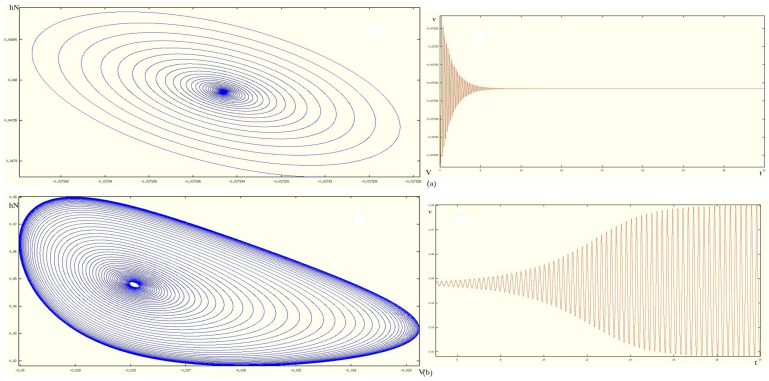


Figure 4: Comparison between phase portraits and oscillatory waveforms in (V, h_{Na})-space for $C = 0.645$ and $C = 0.72$.

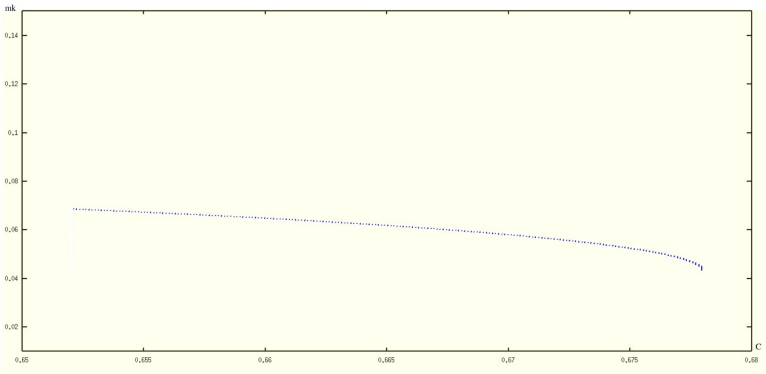


Figure 5: Bifurcation diagram of system (2)

is a stable focus when the membrane capacitance is bigger than C_0. In this case, through time, the amplitude of oscillatory waveforms will vanish. While for $C < C_0$, the equilibrium point q_2 turns out to be an unstable focus surrounded by a stable limit cycle.

This subsystem exhibits either tonic spiking, which corresponds to a stable limit cycle or quiescence which corresponds to a stable equilibrium point. Stability loss of the stable limit cycle in the fast subsystem gives rise to a stable equilibrium through a supercritical Andronov–Hopf bifurcation. When a stable limit cycle shrinks into an unstable equilibrium, then the equilibrium becomes stable. That is, the equilibrium undergoes supercritical Andronov–Hopf bifurcation.

5 Conclusion

Since finding bifurcations, such as Hopf bifurcations in dynamical systems, is not an easy task, mathematical researchers are looking for computer tools and methods to observe Hopf bifurcations and estimate the number of bifurcation parameters.

In this paper, a leech heart interneurons model has been considered when membrane capacitance is the control parameter. Then by using the Routh–Hurwitz criteria, the sufficient conditions have been examined such that under which the system undergoes supercritical Andronov–Hopf bifurcation. Hence, in this model, the transition between the silent phase and tonic spiking mode was through Hopf bifurcation. Moreover, the authors introduced an interval in the parameters space in which the system has a stable limit cycle in which the neuron is in a tonic spiking mode. Numerical simulations were carried out using the fourth and fifth-order Runge–Kutta method to support our analytical results. In fact, by simulations and Hopf bifurcation theory, we showed that the stable limit cycle could be disappeared.

Acknowledgements

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

References

[1] Bertram, R., Butte, M.J., Kiemel, T. and Sherman, A. *Topological and phenomenologial classification of bursting oscillations*, Bull. Math. Biol. 57(3) (1995) 413–439.

[2] Calabrese, R.L., Nadim, F. and Olsen, Ø.H. *Heartbeat control in the medicinal leech: a model system for understanding the origin, coordination, and modulation of rhythmic motor patterns,* J. Neurobiol. 27(3) (1995) 390–402.

[3] Cymbalyuk, G.S. and Calabrese, R.L. *A model of slow plateau-like oscillations based upon the fast Na+ current in a window mode*, Neurocomputing 38 (2001) 159–166.

[4] Cymbalyuk, G.S., Gaudry, Q., Masino, M.A., and Calabrese, R.J. *Bursting in leech heart interneurons: cell-autonomous and network-based mechanisms,* J. Neurosci. 22 (24) (2002) 10580–10592. Neurosci. 22, 10580 (2002).

[5] Cymbalyuk, G. and Shilnikov, A. *Coexistence of tonic spiking oscillations in a leech neuron model,* J. Comput. Neurosci. 18(3) (2005) 255–263.

[6] Gentet, L.J., Stuart G.J., and Clements J.D. *Direct measurement of specific membrane capacitance in neurons,* Biophys. J. 79 (2000) 314–320.

[7] Grimnes, S. and Martinsen, Ø.G. *Alpha-dispersion in human tissue*, J. Phys.: Conf. Ser. 224(1) (2010) 012073.

[8] Guckenheimer, J. and Holmes, P.J. *Nonlinear oscillations, dynamical systems and bifurcations of vector fields*, Vol. 42. Springer Science & Business Media, 2013.

[9] Hill, J., Lu, M., Masino, O. Olsen, R. and Calabrese, L. *A model of a segmental oscillator in the leech heartbeat neuronal network*, J. Comput. Neuroscience. 10 (2001) 281–302.

[10] Howell, B., Medina, L.E. and Grill, W.M. *Effects of frequency-dependent membrane capacitance on neural excitability*, J. Neural Eng. 12 (2015) 056015.

[11] Izhikevich, E. *Neural excitability, spiking and bursting*, Int. J. Bifurc. Chaos 10 (6) (2000) 1171–1266.

[12] Kolomiets, M. L. and Shilnikov, A. L. *Poincarè return maps in neural dynamics: three examples*, International Conference on Difference Equations and Applications, pp. 45–57. Springer, Cham, 2019.

[13] Liu, W. *Criterion of Hopf bifurcation without using eigenvalues*, J. Math. Anal. Appl. 182(1) (1994) 250–256.

[14] Malashchenko, T., Shilnikov, A. and Cymbalyuk, G. *Bistability of bursting and silence regimes in a model of a leech heart interneuron*, Phys. Rev. E 84(4) (2011) 041910.

[15] McIntyre, C.C., Richardson A.G., and Grill, W.M. *Modeling the excitability of mammalian nerve fibers: influence of after potentials on the recovery cycle* J. Neurophysiol. 87 (2002) 995–1006.

[16] McNeal, D. R. *Analysis of a model for excitation of myelinated nerve IEEE Trans.* Biomed. Eng. 23 (1976) 329–337 .

[17] Monfared, Z. and Dadi, Z. *Analysing panel flutter in supersonic flow by Hopf bifurcation*, Iranian Journal of Numerical Analysis and Optimization 4(2) (2014) 1–14.

[18] Opdyke, C.A. and Calabrese, R.L. *A persistent sodium current contributes to oscillatory activity in heart interneurons of the medicinal leech*, J. Comp. Physiol. 175. (1994) 781–789.

[19] Shilnikov, A.L. and Cymbalyuk G. *Transition between tonic spiking and bursting in a neuron model via the blue-sky catastrophe*, Phys. Rev. Lett. 94 (4) (2005) 048101.

[20] Süli E. *Numerical solution of ordinary differential equations*, Mathematical Institute, University of Oxford, 2010.

[21] Wanga Q., Duana, Z., Fengc, Z., Chena, G. and Lu, Q. *Synchronization transition in gap-junction-coupled leech neurons*, Phys. A: Stat. Mech. Appl. 387 (2008) 4404–4410.

How to cite this article
Parvizi, F., Razvan, M. and Alipour Fakhri, Y., Stability and Hopf Bifurcation in Leech Heart Interneuron Model. *Iran. j. numer. anal. optim.*, 2023; 13(2): 170-186. https://doi.org/10.22067/ijnao.2022.75087.1101

Iranian Journal of Numerical Analysis and Optimization
Vol. 13, No. 2, 2023, pp 187–204
https://doi.org/10.22067/ijnao.2022.76010.1123
https://ijnao.um.ac.ir/



How to cite this article
Research Article



Approximate solutions to the Allen–Cahn equation using rational radial basis functions method

M. Shiralizadeh*, A. Alipanah and M. Mohammadi

Abstract

We apply the rational radial basis functions (RRBFs) method to solve the Allen–Cahn (A.C) equation, particularly when the equation has a solution with steep front or sharp gradients. We approximate the spatial derivatives by the RRBFs method. Then we apply an explicit, fourth-order Runge–Kutta method to advance the resulting semi-discrete system in time. It is well known that the A.C equation has a nonlinear stability feature, meaning that the free-energy functional is reduced by time. The presented method maintains the total energy reduction property of the A.C equation. In the end, five examples to confirm the efficiency and accuracy of the proposed method are provided.

AMS subject classifications (2020): 65N99, 74G15, 97N40.

Keywords: Allen–Cahn equation, RBFs, Rational RBFs method, Runge–Kutta method.

---

*Corresponding author

Received 4 April 2022; revised 26 July 2022; accepted 26 August 2022

Mansour Shiralizadeh
Department of Applied Mathematics, University of Kurdistan, Sanandaj, Iran.
Department of Mathematics, Payame Noor University (PNU), Tehran, Iran. e-mail: m.shiralizadeh@pnu.ac.ir

Amjad Alipanah
Department of Applied Mathematics, University of Kurdistan, Sanandaj, Iran. e-mail: a.alipanah@uok.ac.ir

Maryam Mohammadi
Faculty of Mathematical Sciences and Computer, Kharazmi University, Tehran, Iran. e-mail: m.mohammadi@khu.ac.ir

1 Introduction

In this paper, we examine the Allen–Cahn (A.C) equation as follows:

$$w_t = \nu w_{zz} - h(w), \qquad z \in [a,b],\ t \in (0,T], \tag{1}$$

with the initial condition

$$w(z,0) = g(z), \qquad z \in [a,b], \tag{2}$$

and Dirichlet boundary conditions as

$$w(a,t) = l(t), \qquad w(b,t) = k(t), \qquad t \in [0,T], \tag{3}$$

where the parameter ν is a positive constant and usually represents the interfacial width. The term $h(w) = H^{'}(w)$ with $H(w) = \frac{1}{4}(w^2-1)^2$ is a given energy potential.

The A.C equation was initially given by Allen and Cahn [1] to illustrate the motion of anti-phase boundaries in crystal solids. The A.C equation is a specific type of nonlinear partial differential equation used extensively to model various phenomena, such as crystal growth [3, 12], phase transitions [6, 28], image analysis [2, 8, 20, 26], texture growth [9, 23, 24, 27], and surface dynamics in material science [13, 21].

We know that the solution $w(z,t)$ to the A.C equation has the feature that the total energy $E_T(w)$ is reduced over time, and it is given as follows:

$$E_T(w) = \int_a^b \left(\tfrac{1}{2}\nu|w_z|^2 + H(w)\right) dz, \tag{4}$$

and may be used to assess the accuracy of the rational radial basis functions (RRBFs) method presented here.

In general, the analytical solution of the A.C equation is not available. So many researchers tend to use numerical methods, such as finite difference [22], finite element [11], Fourier spectral with periodic boundary condition [25], and radial basis functions (RBFs) methods [29]. In recent years, researchers have been interested in developing numerical methods that can preserve the energy dissipation property of the A.C equations.

Among these methods, the meshless technique avoids mesh generation. It uses scattered points instead of meshing the domain of the problem. For more description, see [14, 15, 16, 30, 36, 38]. The RBFs methods are substantial instruments for scattered points interpolation and solving partial differential equations (PDEs). When the underlying function or the solution of PDE is sufficiently smooth, RBFs methods can produce exponential accuracy, but if the underlying function or the solution of PDE has steep gradients or discontinuities, then the RBFs method may produce solutions with oscillation. In such situations, the RRBFs method can be used to approximate derivatives of functions and to solve PDEs [35]. RRBFs method approximating solutions in

nonlinear spaces generated by RBFs is more computationally expensive than the standard RBFs methods. For smooth problems, the accuracy obtained from these methods may not be worth the additional computations. However, for problems with discontinuities and steep fronts, these methods will be considerably more accurate, and the additional computations will be justified.

In 2009, the RRBFs method was used to interpolate functions with steep gradients by Jakobsson, Andersson, and Edelvik [17]. Sarra and Bai [35] extended the method of Jakobsson et al. to interpolate discontinuous functions and solve Burger's equation. The RRBFs-based partition of unity interpolation was used by De Marchi, Martinez, and Perracchione [4], and in [37], the authors used the RRBFs method for solving the Sine-Gordon equation.

In this paper, we apply the RRBFs method for solving (1) with conditions given in (2)–(3), particularly when the equation has a solution with steep front or sharp gradients. We approximate the spatial derivatives by the RRBFs method. Then we apply an explicit, fourth-order Runge–Kutta (RK4) method to advance the resulting semi-discrete system in time.

This paper is organized as follows: In Section 2, we introduce the RBFs and RRBFs methods. The method implementation is given in Section 3. We analyze stability issues in Section 4. In Section 5, we present the results of numerical examples. Section 6 is dedicated to the conclusion.

2 RBFs and Rational RBFs interpolation

2.1 RBFs interpolation

Let $\Gamma \subseteq \mathbb{R}^m$ be a bounded set, let $Z = \{\boldsymbol{z}_1^c, \ldots, \boldsymbol{z}_N^c\} \subseteq \Gamma$ be a set of N distinct points, hereinafter referred to as centers, and let $F = \{f(\boldsymbol{z}_1^c), \ldots, f(\boldsymbol{z}_N^c)\}$ be a set of function values. An RBF

$$\psi(\boldsymbol{z}) = \psi\left(\|\boldsymbol{z} - \boldsymbol{z}^c\|_2, \epsilon\right), \qquad \boldsymbol{z}, \boldsymbol{z}^c \in \mathbb{R}^m,$$

is a function of one variable $r = \|\boldsymbol{z} - \boldsymbol{z}^c\|_2$ that is centered at $\boldsymbol{z}^c$, which ϵ is a free parameter and called the shape parameter [10, 35]. Also, in this paper, $m = 1$.

The inverse quadratic RBF

$$\psi(r) = \frac{1}{1 + (\epsilon r)^2}$$

is a strictly positive definite RBF that we use it in all examples.

The RBF interpolant takes the form

$$p(\boldsymbol{z}) = \sum_{j=1}^{N} c_j \ \psi\left(\|\boldsymbol{z} - \boldsymbol{z}_j^c\|_2 \, , \, \epsilon\right), \tag{5}$$

which the coefficients c_j are obtained by solving the linear system

$$A\mathbf{c} = \mathbf{f},$$

based on the interpolation conditions $p\left(\boldsymbol{z}_i^c\right) = f_i$, where $\mathbf{f} = [f(\boldsymbol{z}_1^c), \ldots, f(\boldsymbol{z}_N^c)]^T$. The elements of the matrix A are of the form

$$a_{ij} = \psi\left(\|\boldsymbol{z}_i^c - \boldsymbol{z}_j^c\|_2 \, , \, \epsilon\right), \qquad i, j = 1, \ldots, N. \tag{6}$$

Moreover, A is a symmetric positive definite matrix and thus invertible. The evaluation of the interpolant (5) at M points $\boldsymbol{z}_j$ is done by multiplying $\mathbf{c}$ by E, where the elements of the matrix E are of the form

$$e_{ij} = \psi\left(\|\boldsymbol{z}_i - \boldsymbol{z}_j^c\|_2 \, , \, \epsilon\right), \qquad i = 1, \ldots, M, \quad j = 1, 2, \ldots, N. \tag{7}$$

The first and second derivatives of RBF interpolant are of the form

$$D(p(\boldsymbol{z})) = \sum_{j=1}^{N} c_j \ D\left(\psi\left(\|\boldsymbol{z} - \boldsymbol{z}_j^c\|_2 \, , \, \epsilon\right)\right).$$

Thus

$$D(p(\boldsymbol{z}_i^c)) = \sum_{j=1}^{N} c_j \ D\psi\left(\|\boldsymbol{z}_i^c - \boldsymbol{z}_j^c\|_2 \, , \, \epsilon\right),$$

that is,

$$D\mathbf{f} \simeq E_D\mathbf{c},$$

where the elements of E_D are of the form

$$(E_D)_{ij} = D\psi\left(\|\boldsymbol{z}_i^c - \boldsymbol{z}_j^c\|_2 \, , \, \epsilon\right), \qquad i, j = 1, \ldots, N,$$

and

$$D(D(p(\boldsymbol{z}))) = \sum_{j=1}^{N} c_j \ D\left(D\psi\left(\|\boldsymbol{z} - \boldsymbol{z}_j^c\|_2 \, , \, \epsilon\right)\right).$$

Thus

$$D(D(p(\boldsymbol{z}_i^c))) = \sum_{j=1}^{N} c_j \ D\left(D\psi\left(\|\boldsymbol{z}_i^c - \boldsymbol{z}_j^c\|_2 \, , \, \epsilon\right)\right),$$

that is,

$$D(D\mathbf{f}) \simeq E_{DD}\mathbf{c},$$

where the elements of E_{DD} are of the form

$$(E_{DD})_{ij} = D\left(D\psi\left(\|\boldsymbol{z}_i^c - \boldsymbol{z}_j^c\|_2\,,\,\epsilon\right)\right), \qquad i,j = 1,\ldots,N.$$

The following theorem [4] gives an error bound in terms of the power function $P_{\psi,Z}$, where $\mathcal{N}_\psi(\Gamma)$ is the native reproducing kernel Hilbert space corresponding to the symmetric positive definite RBF [10].

Theorem 1. Let $\psi \in C(\Gamma \times \Gamma)$ be a strictly positive definite RBF, and suppose that $Z = \{\boldsymbol{z}_1^c,\ldots,\boldsymbol{z}_N^c\} \subseteq \Gamma$ is a set of distinct points and that p is the interpolant to $f \in \mathcal{N}_\psi(\Gamma)$. Then for all $\boldsymbol{z} \in \Gamma$, we have

$$|f(\boldsymbol{z}) - p(\boldsymbol{z})| \leq P_{\psi,Z}(\boldsymbol{z})\|f\|_{\mathcal{N}_{\psi(\Gamma)}}.$$

2.2 Rational RBFs interpolation

The RBFs interpolants have problems, such as ill-conditioning, especially when the shape parameter tends to zero. These problems might lead to inaccurate solutions, particularly when the functions with steep front or sharp gradients are considered [33]. Moreover, when the functions have steep fronts or sharp gradients, the rational RBFs method approximates them better than the standard RBFs method [35]. These are the main reasons that we use rational RBFs interpolants.

The RRBF interpolant of function f is given by

$$\mathcal{Q}(\boldsymbol{z}) = \frac{u(\boldsymbol{z})}{v(\boldsymbol{z})},$$

that satisfies in the interpolation conditions

$$\mathcal{Q}(\boldsymbol{z}_k^c) = f(\boldsymbol{z}_k^c), \qquad k = 1,2,\ldots,N,$$

and $u(\boldsymbol{z})$ and $v(\boldsymbol{z})$ are the RBF interpolants

$$u(\boldsymbol{z}) = \sum_{j=1}^{N} c_j^u\ \psi\left(\|\boldsymbol{z} - \boldsymbol{z}_j^c\|_2\,,\,\epsilon\right),$$

and

$$v(\boldsymbol{z}) = \sum_{j=1}^{N} c_j^v\ \psi\left(\|\boldsymbol{z} - \boldsymbol{z}_j^c\|_2\,,\,\epsilon\right).$$

Let $\mathbf{u} = [u(\boldsymbol{z}_1^c),\ldots,u(\boldsymbol{z}_N^c)]^T$ and $\mathbf{v} = [v(\boldsymbol{z}_1^c),\ldots,v(\boldsymbol{z}_N^c)]^T$. By applying the interpolation conditions, we have a system of equations that is underdetermined. Thus in order for the rational interpolant to be defined uniquely, we add an additional condition (for more descriptions, see [17, 35]), which leads the native space semi-norms [10] of the RBF interpolants $u(\boldsymbol{z})$ and $v(\boldsymbol{z})$ to be

minimized. By adding the condition, we will have a minimization problem with the solution $\mathbf{v}$, which is the eigenvector corresponding to the smallest eigenvalue problem

$$R\mathbf{v} = \lambda\mathbf{v},$$

where

$$R = \operatorname{diag}\left(1/\left(\frac{\mathbf{f}^2}{\|\mathbf{f}\|_2^2}+1\right)\right)\left(\frac{DA^{-1}D}{\|\mathbf{f}\|_2^2}+A^{-1}\right), \tag{8}$$

and A is the RBF system matrix given in (6), $\mathbf{f} = [f(\boldsymbol{z}_1^c),\ldots,f(\boldsymbol{z}_N^c)]^T$, and $D = \operatorname{diag}(f(\boldsymbol{z}_1^c),\ldots,f(\boldsymbol{z}_N^c))$. Moreover, division is elementwise and $\mathbf{f}^2$ is an elementwise squaring of the elements of the vector $\mathbf{f}$. When $\mathbf{v}$ is found, $\mathbf{u}$ is obtained by $\mathbf{u} = D\mathbf{v}$. When $\mathbf{u}$ and $\mathbf{v}$ are found, we obtain the coefficients of the RBF interpolants by solving two linear systems,

$$Ac^u = \mathbf{u} \quad \text{and} \quad Ac^v = \mathbf{v}. \tag{9}$$

Now we evaluate the rational interpolant at M points $\boldsymbol{z}_j$ by

$$\mathbf{Q} = \frac{Ec^u}{Ec^v}, \tag{10}$$

where division is elementwise, E is the RBF evaluation matrix (7), and $\mathbf{Q} = [\mathcal{Q}(\boldsymbol{z}_1),\ldots,\mathcal{Q}(\boldsymbol{z}_M)]^T$.

Now, we calculate the derivatives of the rational interpolant at N points $\boldsymbol{z}_i^c$ by applying quotient rule as below:

$$\mathbf{Q}' = \frac{(Ac^v).(E_Dc^u)-(Ac^u).(E_Dc^v)}{(Ac^v)^2}, \tag{11}$$

$$\mathbf{Q}'' = \frac{2(Ac^u).(E_Dc^v)^2+(Ac^v)^2.(E_{DD}c^u)}{(Ac^v)^3} - \frac{(Ac^v).\Big(2\,(E_Dc^u).(E_Dc^v)+(Ac^u).(E_{DD}c^v)\Big)}{(Ac^v)^3}, \tag{12}$$

where $\mathbf{Q}' = \left[\mathcal{Q}'(\boldsymbol{z}_1^c),\ldots,\mathcal{Q}'(\boldsymbol{z}_N^c)\right]^T$, $\mathbf{Q}'' = \left[\mathcal{Q}''(\boldsymbol{z}_1^c),\ldots,\mathcal{Q}''(\boldsymbol{z}_N^c)\right]^T$, A is the RBF system matrix, and E_D and E_{DD} are the first and second derivatives of evaluation matrix at N points $\boldsymbol{z}_i^c$.

It is concluded from (9) and (10) that the RRBFs interpolant is made by the RBFs interpolation matrix A. Therefore, we are able to provide error bounds for the RRBFs interpolation [4]. For this purpose, we must consider $\mathbf{u}$ and $\mathbf{v}$ as the values obtained by sampling some functions r and $s \in \mathcal{N}_\psi(\Gamma)$, respectively.

Proposition 1. Suppose that $\Gamma \subseteq \mathbb{R}^m$, $\psi \in C(\Gamma\times\Gamma)$ is a strictly positive definite RBF, and let $Z = \{\boldsymbol{z}_1^c, \boldsymbol{z}_2^c, \ldots, \boldsymbol{z}_N^c\} \subseteq \Gamma$ be a set of distinct points

and let $\mathcal{Q}$ be the RRBF interpolant to $f \in \mathcal{N}_\psi(\Gamma)$. Moreover, let us suppose $r, s \in \mathcal{N}_\psi(\Gamma)$. Then for all $\boldsymbol{z} \in \Gamma$, we have

$$|f(\boldsymbol{z}) - \mathcal{Q}(\boldsymbol{z})| \leq \frac{1}{|v(\boldsymbol{z})|} \left(|f(\boldsymbol{z})| \|s\|_{\mathcal{N}_\psi(\Gamma)} + \|r\|_{\mathcal{N}_\psi(\Gamma)} \right) P_{\psi,Z}(\boldsymbol{z}).$$

So Proposition 1 gives the error bound for the RRBFs interpolant in terms of data values and power function, as for the standard RBFs interpolant.

3 RRBFs method for solving A.C equation

Now, we use the RRBFs method to solve the A.C equation. In fact, we develop RRBFs to study the A.C equation with Dirichlet boundary conditions. In our method, we employ the derivatives of the RRBFs to approximate the spatial derivatives. Then we apply an explicit, RK4 method to advance the resulting semi-discrete system in time. In fact, we first approximate the spatial derivatives at N centers $\boldsymbol{z}_i^c$ at time t^n, that is,

$$\mathbf{W}_{zz}(z, t^n) = \left[W_{zz}(\boldsymbol{z}_1^c, t^n), W_{zz}(\boldsymbol{z}_2^c, t^n), \ldots, W_{zz}(\boldsymbol{z}_N^c, t^n) \right]^T,$$

of the A.C equation

$$w_t = \nu w_{zz} - h(w)$$

by the RRBFs method, which is given by computing (12) when

$$\mathbf{W}^n = \left[W(\boldsymbol{z}_1^c, t^n), W(\boldsymbol{z}_2^c, t^n), \ldots, W(\boldsymbol{z}_N^c, t^n) \right]^T,$$

is considered instead of $\mathbf{f}$. This leads to a system of ordinary differential equations of the form

$$\mathbf{W}_t = G(\mathbf{W}, t).$$

Then we apply RK4 method to solve the above system of equations as follows:

$$\begin{aligned}
\mathbf{K}_1 &= \Delta t G(\mathbf{W}^n, t^n), \\
\mathbf{K}_2 &= \Delta t G(\mathbf{W}^n + 0.5k_1, t^n + 0.5\Delta t), \\
\mathbf{K}_3 &= \Delta t G(\mathbf{W}^n + 0.5k_2, t^n + 0.5\Delta t), \\
\mathbf{K}_4 &= \Delta t G(\mathbf{W}^n + k_3, t^n + \Delta t), \\
\mathbf{W}^{n+1} &= \mathbf{W}^n + \tfrac{1}{6}(\mathbf{K}_1 + 2\mathbf{K}_2 + 2\mathbf{K}_3 + \mathbf{K}_4),
\end{aligned} \tag{13}$$

where $t^n = t^{n-1} + \Delta t$ and Δt is the time step size. Clearly, to obtain $\mathbf{W}^1$ initial value, $\mathbf{W}^0$ is required in the iterative computation. Using the initial condition in (2), we have $\mathbf{W}^0 = \left[g(\boldsymbol{z}_1^c), g(\boldsymbol{z}_2^c), \ldots, g(\boldsymbol{z}_N^c) \right]^T$.

To apply the Dirichlet boundary conditions (3), we directly replace $W(\boldsymbol{z}_1^c, t^n)$ with $w(a, t^n) = l(t^n)$ and $W(\boldsymbol{z}_N^c, t^n)$ with $w(b, t^n) = k(t^n)$.

4 Stability issue

Since the RRBFs interpolant is constructed by using the standard RBFs interpolation matrix, the RRBFs method can also suffer from instability problems, especially for $\epsilon \to 0$. In fact, serious problems of ill-conditioning may occur by choosing the wrong values of the shape parameter, particularly for infinitely smooth RBFs. In order for the system matrix B to be well-conditioned, the shape parameter ϵ must not be too small, but small shape parameters are required to obtain better accuracy for the interpolation with RBFs. This is known as the uncertainty principle. We choose the shape parameter ϵ so that the collocation matrix B has a condition number, $\kappa(B)$, in the range $10^{13} \leq \kappa(B) \leq 10^{15}$, in order to determine the proper value for ϵ [34]. These bounds for the condition number are valid when using a computer that implements double precision floating point arithmetic, but the bounds will be different when using other floating point number systems.

5 Numerical results

In this section, we consider five numerical examples of the A.C equation to validate the presented scheme. The accuracy and efficiency of the method are confirmed by calculating L_∞, L_2, and RMS error norms as follows. We also investigate energy dissipation property in our numerical experiments to illustrate the efficiency of the presented method.

$$L_\infty = \|\mathbf{W} - \mathbf{w}\|_\infty,$$

$$L_2 = \|\mathbf{W} - \mathbf{w}\|_2,$$

$$RMS = \sqrt{\tfrac{1}{N}\|\mathbf{W} - \mathbf{w}\|_2},$$

where $\mathbf{w}$ and $\mathbf{W}$ are the exact and numerical solutions, respectively. Moreover, in all numerical examples, we have used uniformly spaced centers

$$\boldsymbol{z}_i^c = a + \frac{i-1}{N-1}(b-a), \qquad i = 1, 2, \ldots, N.$$

Also in the examples, we evaluate the total energy E_T given in (4) using the composite trapezoidal rule for integration.

Example 1. In this example, we examine the A.C equation with three large gradients

$$w_t = \nu w_{zz} + w - w^3, \qquad z \in [-1, 1],$$

and with the initial and Dirichlet boundary conditions [18]

$$w(z, 0) = 0.5z - 0.5\sin(1.5\pi z), \quad w(1, t) = -w(-1, t) = 1.$$

The results are obtained using $\nu = 10^{-5}$, $k = \Delta t = 0.01$, shape parameter $\epsilon = 6$, and $N = 100$ centers. In Figure 1, we present the approximate solution at time $t = 10$ and the space-time graph of the approximate solution. Table 1 shows the values of the energy at various times. It can be seen from Table 1 that energy reduces over time and that the RRBFs method resolves the problem accurately.

Table 1: Energy of the A.C equation for Example 1 at several times.

$Time\,(t)$	$t = 0$	$t = 1$	$t = 2$	$t = 5$	$t = 10$	$t = 20$
$Energy\,(E_T)$	0.3403	0.1941	0.0738	0.0017	7.6908e-04	7.6900e-04

Example 2. In this example, we examine the A.C equation

$$w_t = \nu w_{zz} + w - w^3,$$

in the computational domain $[-1, 1] \times [0, 10]$ with the parameter $\nu = 10^{-6}$. The initial condition is given by [31, 5]

$$w(z, 0) = 0.6z + 0.4\sin\left(\frac{\pi}{2}(z^2 - 3z - 1)\right),$$

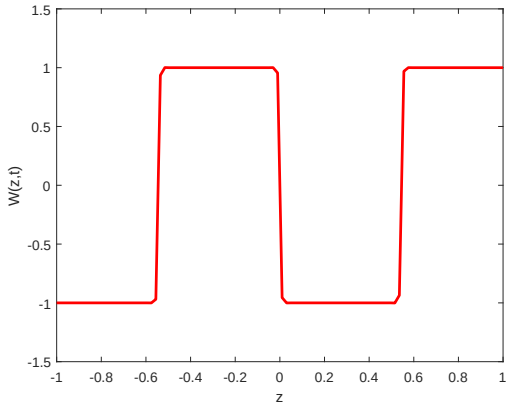
and the Dirichlet boundary conditions are

$$w(1, t) = -w(-1, t) = 1.$$

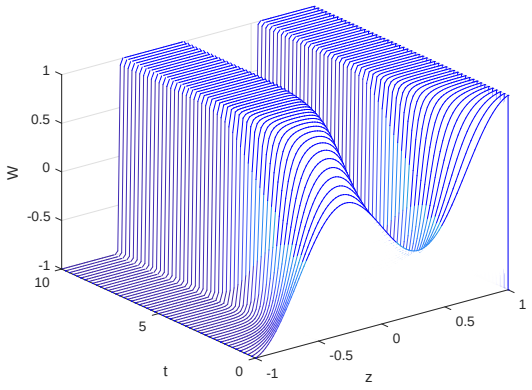
We solve the problem with the RRBFs method with $k = \Delta t = 0.01$, $N = 100$ centers, and the shape parameter $\epsilon = 6$. We have obtained the solution up to time $t = 10$. The graph of approximate solutions at times $t = 0, 2, 4, 6, 10$ and space-time graph for $t \in [4, 10]$ are shown in Figure 2. As it is seen, the solution has a steep front at different times. Table 2 shows the energy at several times. It can be seen from Table 2 that energy reduces over time as theory guarantees and that the RRBFs method resolves the problem accurately.

Example 3. Consider the A.C equation

$$w_t = \nu w_{zz} + w - w^3, \qquad z \in [0, 1], \quad t \geq 0$$

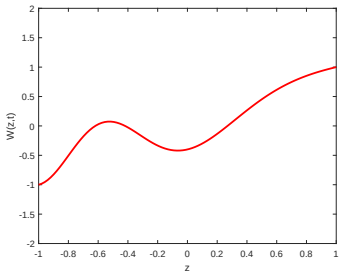


(a) RRBFs solution at time $t = 10$.

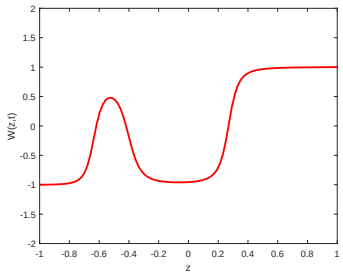


(b) Space-time graph.

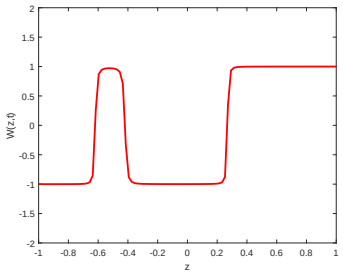
Figure 1: Numerical results for the solution of Example 1.



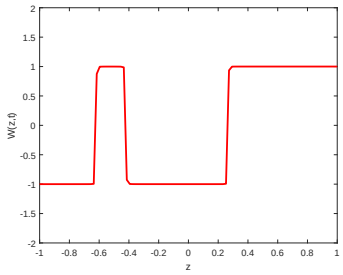
(a) RRBFs solution at time $t = 0$.



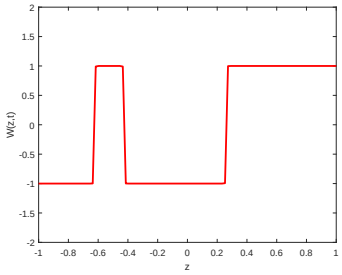
(b) RRBFs solution at time $t = 2$.



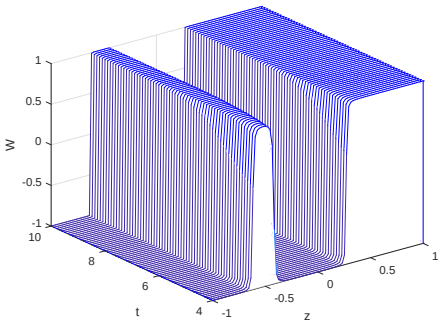
(c) RRBFs solution at time $t = 4$.



(d) RRBFs solution at time $t = 6$.



(e) RRBFs solution at time $t = 10$.



(f) Space-time graph.

Figure 2: Numerical results for the solution of Example 2.

Table 2: Energy of the A.C equation for Example 2 at several times.

$Time\,(t)$	$t=0$	$t=2$	$t=4$	$t=6$	$t=10$
$Energy\,(E_T)$	0.3161	0.0945	0.0154	5.3755e-04	7.1318e-05

with the initial and Dirichlet boundary conditions [7]

$$w(z,0)=\sin(7\pi z),\quad w(0,t)=w(1,t)=0.$$

For the numerical solution, we use $N=100$ centers, $\nu=0.0025$, $k=\Delta t=0.001$, and the shape parameter $\epsilon=8$. The plots of the numerical solution and the energy are shown in Figure 3. It can be seen that energy reduces over time, as the theory guarantees. Compared with [7], better accuracy is obtained.

Example 4. Consider the A.C equation

$$w_t=w_{zz}+\frac{1}{\nu^2}(w-w^3),$$

in the region $0\le z\le 15$, where the parameter $\nu=0.08$, and the initial condition is given by

$$w(z,0)=0.5-0.5\tanh\left(\frac{z-2}{2\sqrt{2}\nu}\right).$$

The analytical solution is given in [19] as

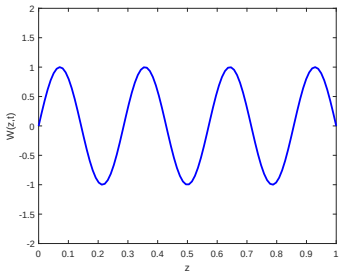
$$w\,(z,t)=0.5-0.5\tanh\left(\frac{z-2-rt}{2\sqrt{2}\nu}\right),$$

where $r=\frac{3}{\sqrt{2}\nu}$, we obtain the Dirichlet boundary conditions from the exact solution and solve this problem with the RRBFs method with $k=\Delta t=0.00001$, $N=200$ centers and the shape parameter $\epsilon=2.25$.

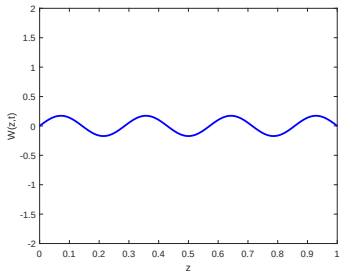
The graphs of exact and approximate solutions at time $t=0.04$ are shown in Figure 4. As it is seen, the solution has a steep front. Table 3 shows the L_∞, L_2, RMS errors, and energy

$$E_T(w)=\int_a^b\left(\frac{1}{2}|w_z|^2+\frac{1}{4\nu^2}(w^2-1)^2\right)dz$$

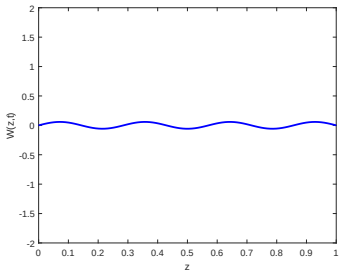
at several times. It can be seen from Table 3 that the energy reduces over time and that the RRBFs method resolves the problem accurately. Also, the results of the method are well matched with analytical solutions.



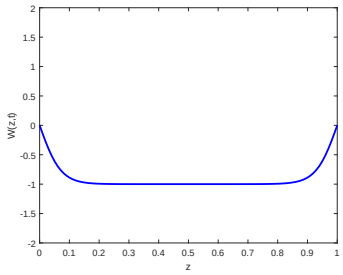
(a) RRBFs solution at time $t = 0$.



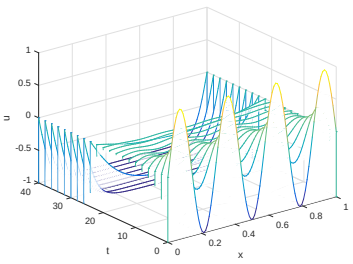
(b) RRBFs solution at time $t = 5$.



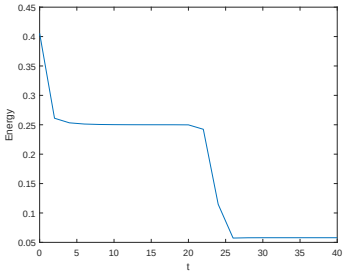
(c) RRBFs solution at time $t = 10$.



(d) RRBFs solution at time $t = 40$.



(e) Space-time graph.

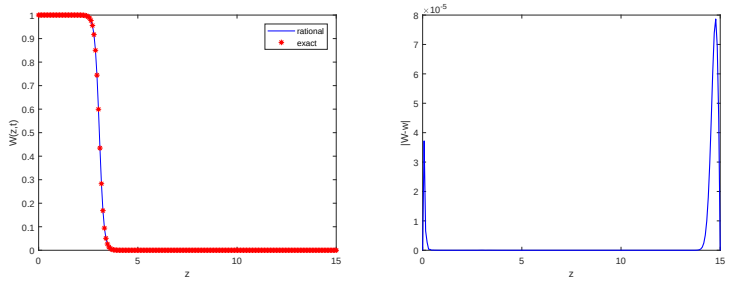


(f) energy evolution.

Figure 3: Numerical results for the solution of Example 3.

Table 3: Errors and energy of Example 4.

t	L_∞	L_2	RMS	$Energy\,(E_T)$
0	0	0	0	509.2856
0.001	2.3112e-05	2.3259e-05	1.6447e-06	508.2498
0.005	5.8036e-04	1.2000e-03	8.2960e-05	504.0963
0.01	5.8573e-04	1.2000e-03	8.2963e-05	498.9173
0.04	7.8661e-05	1.6396e-04	1.1594e-05	467.8536



(a) RRBFs and analytical solutions at $t = 0.04$.

(b) Pointwise error at $t = 0.04$.

Figure 4: Numerical results for the solution of Example 4.

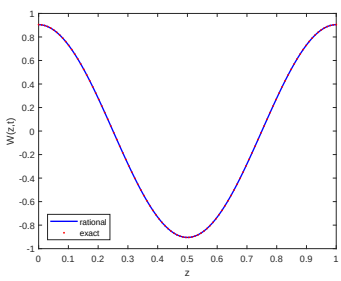
Example 5. Consider the A.C equation

$$w_t = w_{zz} + \frac{1}{\nu^2}(w - w^3),$$

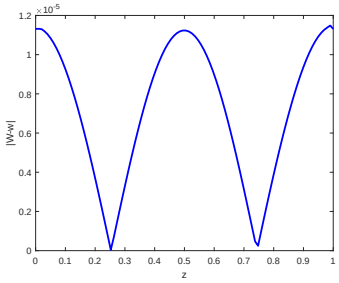
with a forcing term h added by imposing analytical solution $w(z,t) = e^{-t}\cos(2\pi z)$ in the domain $0 \le z \le 1$, where the parameter $\nu = 0.1$ [32]. We obtain the Dirichlet boundary conditions from the analytical solution and solve this problem with the RRBFs method with $k = \Delta t = 10^{-4}/8$, $N = 100$ centers and the shape parameter $\epsilon = 8$. The graphs of exact and approximate solutions at time $t = 0.1$ are shown in Figure 5. Table 4 shows the L_∞, L_2, RMS errors, and energy at several times. It can be seen from Table 4 that the energy reduces over time and that the RRBFs method resolves the problem accurately. Also, the results of the method are well matched with analytical solutions and somehow better than the earlier works.

Table 4: Errors and energy of Example 5.

	RRBF				RBF	[32]
t	L_∞	L_2	RMS	E_T	L_∞	L_∞
0	0	0	0	19.2446	0	-
0.001	1.6671e-07	2.6333e-07	2.6333e-08	19.2374	2.1880e-06	-
0.005	1.7437e-07	3.7249e-07	3.7249e-08	19.2095	2.3599e-06	-
0.01	1.7140e-07	4.2040e-07	4.2040e-08	19.1766	2.3762e-06	8.3500e-06
0.1	1.1485e-05	8.0532e-05	8.0532e-06	18.8965	1.2587e-05	6.6300e-05
0.3	4.2165e-05	2.5234e-04	2.5234e-05	19.5199	6.5511e-04	-



(a) RRBFs and analytical solutions at $t = 0.1$.



(b) Pointwise error at $t = 0.1$.

Figure 5: Numerical results for the solution of Example 5.

6 Conclusion

This paper investigated the application of the RRBFs method to find the numerical solution of the A.C equation with Dirichlets's boundary conditions, especially when the equation has a solution with steep front or sharp gradients. The combination of RRBFs in space and RK4 in time gave accurate and reliable solutions to the A.C equation. To illustrate the accuracy and efficiency of the presented method, we carried out five examples and presented the corresponding graphs and tables, and compared the results with some earlier works. Moreover, it can be observed that in the case where the analytical solution exists, the numerical solutions described the good accuracy of the method. These results support the assurance of using the method for equation (1) when the analytical solution is not known. In addition, for all examples, the energy dissipation law was investigated. In order to generalize the RRBFs method implementation to higher dimensional time-dependent PDEs, we need to localize the method with a modified partition of the unity method. It is behind the scope of our paper, and we leave it to our further work.

References

[1] Allen, S.M. and Cahn, J.W. *A microscopic theory for antiphase boundary motion and its application to antiphase domain coarsening*, Acta Metall 27 (1979) 1085–1095.

[2] Benes, M., Chalupecky, V. and Mikula, K. *Geometrical image segmentation by the Allen–Cahn equation*, Appl. Numer. Math. 51 (2004) 187–205.

[3] Boettinger, W.J., Warren, J.A., Beckermann, C. and Karma, A. *Phase-field simulation of solidification*, Annu. Rev. Mater. Sci. 32 (2002) 163–194.

[4] De Marchi, S., Martinez, A. and Perracchione, E. *Fast and stable rational RBF-based partition of unity interpolation*, J. Comput. Appl. Math. 349 (2019) 331–343.

[5] Driscoll, A. and Heryudono, R.H. *Adaptive residual subsampling methods for radial basis function interpolation and collocation problems*, Comput. Math. Appl. 53 (2007) 927–939.

[6] Elliott, C.M. and Stinner, B. *Computation of two-phase biomembranes with phase dependent material parameters using surface finite element*, Commun. Comput. Phys. 13 (2013) 325–360.

[7] Elliott, C.M. and Stuart, A.M. *The global dynamics of discrete semilinear parabolic equations*, SIAM J. Numer. Anal. 30 (1993) 1622–1663.

[8] Esedoglu, S. and Tsai, Y.H.R. *Threshold dynamics for the piecewise constant Mumford-Shan functional*, J. Comput. Phys. 211 (2006) 367–384.

[9] Fan, D. and Chen, L.Q. *Computer simulation of grain growth using a continuum field model*, Acta Mater. 45 (1997) 611–622.

[10] Fasshauer, G.E. *Meshfree Approximation Methods with Matlab*, World Scientific, 2007.

[11] Feng, X., Li, Y. and Zhang, Y. *Finite element methods for the stochastics Allen–Cahn equation with gradient-type multiplicative noise*, SIAM J. Numer. Anal. 55(1) (2017) 194–216.

[12] Feng, X., Song, H., Tang, T. and Yang, J. *Nonlinear stability of the implicit-explicit methods for the Allen–Cahn equation*, Inverse Probl. Imaging 7 (2013) 679–695.

[13] Golubovic, L., Levandovsky, A. and Moldovan, D. *Interface dynamics and far-from-equilibrium phase transitions in multilayer epitaxial growth and erosion on crystal surfaces: Continuum theory insights*, East Asian J. Appl. Math. 1 (2011) 297–371.

[14] Heidari, M., Mohammadi, M. and De Marchi, S. *A shape preserving quasi-interpolation operator based on a new transcendental RBF*, Dolomites Research Notes on Approximation 14 (1) (2021) 56–73.

[15] Heydari, M.H. and Hosseininia, M. *A new variable-order fractional derivative with non-singular Mittag-Leffler kernel: application to variable-order fractional version of the 2D Richard equation*, Engineering with Computers 38 (2) (2022) 1759–1770.

[16] Hosseininia, M., Heydari, M.H., Avazzadeh, Z. and Maalek Ghaini, F.M. *A hybrid method based on the orthogonal Bernoulli polynomials and radial basis functions for variable order fractional reaction-advection-diffusion equation*, Engineering Analysis with Boundary Elements 127 (2021) 18–28.

[17] Jakobsson, S., Andersson, B. and Edelvik, F. *Rational radial basis function interpolation with applications to antenna design*, J. Comput. Appl. Math. 233 (2009) 889–904.

[18] Jafari-Varzaneh, H.A. and Hosseini, S.M. *A new map for the Chebyshev pseudospectral solution of differential equations with large gradients*, Numer. Algorithms 69 (2015) 95–108.

[19] Jeong, D. and Kim, J. *An explicit hybrid finite difference scheme for the Allen–Cahn equation*, J. Comput. Appl. Math. 340 (2018) 247–255.

[20] Kay, D.A. and Tomasi, A. *Color image segmentation by the vector valued Allen–Cahn phase-field model: A multigrid solution*, IEEE Trans. Image Process. 18 (2009) 2330–2339.

[21] Kim, J. *Phase-field models for multi-component fluid flows*, Commun. Comput. Phys. 12 (2012) 613–661.

[22] Kim, J., Jeong, D., Yang, S. and Choi, Y. *A finite difference method for a conservative Allen–Cahn equation on non-flat surfaces*, J. Comput. Phys. 334 (2017) 170–181.

[23] Kobayashi, R., Warren, J.A. and Carter, W.C. *A continuum model of grain boundaries*, Phys. D 140 (2000) 141-150.

[24] Krill, C.E. and Chen, L.Q. *Computer simulation of 3-D grain growth using a phase-field model*, Acta Mater. 50 (2002) 3057–3073.

[25] Lee, H. and Lee, J. *A semi-analytical Fourier spectral method for the Allen–Cahn equation*, Comput. Math. Appl. 68 (3) (2014) 174–184.

[26] Lee, H.G., Shin, J. and Lee, J.Y. *First and second order operator splitting method for phase-field crystal equation*, J. Comput. Phys. 299 (2015) 82–91.

[27] Li, Y. and Kim, J. *Multiphase image segmentation using a phase-field model*, Comput. Math. Appl. 62 (2011) 737–745.

[28] Li, Y., Lee, H.G. and Kim, J. A fast, *robust and accurate operator splitting method for Phase-field simulation of crystal growth*, J. Cryst. Growth 321 (2011) 176–182.

[29] Mohammadi, V., Mirzaei, D. and Dehghan, M. *Numerical simulation and error estimation of the time-dependent Allen–Cahn equation on surfaces with radial basis functions*, J. Sci. Comput. 79 (2019) 493–516.

[30] Mohammadi, M., Mokhtari, R. and Schaback, R. *A meshless method for solving the 2d brusselator reaction-diffusion system*, Comput. Model. Eng. Sci. 101 (2014) 113–138.

[31] Naqvi, S.L., Levesley, J. and Ali, S. *Adaptive radial basis function for time dependent partial differential equations*, J. Prime Res. Math. 13 (2017) 90–106.

[32] Niu, J., Xu, M. and Yao, G. *An efficient reproducing kernel method for solving the Allen–Cahn equation*, Appl. Math. Lett. 89 (2019) 78–84.

[33] Perracchione, E. *Rational RBF-based partition of unity method for efficiently and accurately approximating 3D objects*, Comp. Appl. Math. 37 (2018) 4633-4648.

[34] Saberi Zafarghandi, F. and Mohammadi, M. *Numerical approximations for the Riesz space fractional advection-dispersion equations via radial basis functions*, Appl. Numer. Math. 144 (2019) 59–82.

[35] Sarra, S.A. and Bai, Y. *A rational radial basis function method for accurately resolving discontinuities and steep gradients*, Appl. Numer. Math. 130 (2018) 131–142.

[36] Schaback, R. *Kernel-Based Meshless Methods*, Gottingen, 2011.

[37] Shiralizadeh, M., Alipanah, A. and Mohammadi, M. *Numerical solution of one-dimensional Sine-Gordon equation using rational radial basis functions*, J. Math. Model. 10 (3)(2022) 387–405.

[38] Wendland, H. *Scattered data approximation*, Cambridge University Press, 2004.

How to cite this article
Parvizi, F., Razvan, M. and Alipour Fakhri, Y., Approximate solutions to the Allen–Cahn equation using rational radial basis functions method. *Iran. j. numer. anal. optim.*, 2023; 13(2): 187-204. https://doi.org/10.22067/ijnao.2022.76010.1123

Iranian Journal of Numerical Analysis and Optimization
Vol. 13, No. 2, 2023, pp 205–223
https://doi.org/10.22067/ijnao.2022.76652.1135
https://ijnao.um.ac.ir/



How to cite this article
Research Article



Regularization of the generalized auto-convolution Volterra integral equation of the first kind

S. Pishbin*and A. Ebadi

Abstract

In this paper, a generalized version of the auto-convolution Volterra integral equation of the first kind as an ill-posed problem is studied. We apply the piecewise polynomial collocation method to reduce the numerical solution of this equation to a system of algebraic equations. According to the proposed numerical method, for $n = 0$ and $n = 1, \ldots, N-1$, we obtain a nonlinear and linear system, respectively. We have to distinguish between two cases, nonlinear and linear systems of algebraic equations. A double iteration process based on the modified Tikhonov regularization method is considered to solve the nonlinear algebraic equations. In this process, the outer iteration controls the evolution path of the unknown vector U_0^δ in the selected direction $\tilde{u}_0$, which is determined from the inner iteration process. For the linear case, we apply the Lavrentiev $\tilde{m}$ times iterated regularization method to deal with the ill-posed linear system. The validity and efficiency of the proposed method are demonstrated by several numerical experiments.

AMS subject classifications (2020): 65R20, 65Q20, 45G10.

Keywords: Auto-convolution Volterra integral equations; Collocation methods; Tikhonov regularization method; Lavrentiev regularization method.

---

*Corresponding author

Received 12 May 2022; revised 15 September 2022; accepted 3 October 2022

Saeed Pishbin
Department of Mathematics, Faculty of Sciences, Urmia University, Urmia, Iran. e-mail:s.pishbin@urmia.ac.ir

Azam Ebadi
Department of Mathematics, Faculty of Sciences, Urmia University, Urmia, Iran. e-mail:azamebadi172@gmail.com

1 Introduction

Auto-convolution equations have been investigated by mathematicians in various fields. Some of the applications of auto-convolution equations are in spectroscopy, stochastic, and probability theory. In 1990, for a deeper understanding of mathematics and applied numerical analysis of stochastic [19], inverse deconvolution problems were made as a class of problems in continuous spaces. Auto-convolution Volterra integral equations (AVIEs) of the first kind,

$$g(t) = \int_0^t u(t-s)u(s)ds, \tag{1}$$

as an ill-posed problem, have been studied in [3, 5, 6, 9]. Conditions for the compactness, injectivity, and weakly closed of the associated integral operator in (1) have been studied in [9]. Also, the authors applied the Tikhonov regularization for the nonlinear ill-posed problem (1) together with an approach to defining different levels and degrees of ill-posedness in Hilbert spaces. In [3], the authors have developed a local regularization theory for the nonlinear inverse auto-convolution problem (1), which provides regularization methods to preserve the causal nature of the auto-convolution problem. Also, they investigated the convergence of the regularized solutions to the true solution as the noise level in the data shrank to zero and supplied convergence rates for the cases of both L^2 and continuous data. For confluent hypergeometric functions, a class of auto-convolution equations of the first kind was derived in [23] and a further class of auto-convolution equations of the first kind with Mittag-Leffler type functions as solutions was treated. In [5], the authors have presented new results concerning the ill-posedness character of the nonlinear auto-convolution equation (1) and discussed quasi-solutions restricted to specific (relatively) compact subsets together with limitations of Fourier transform techniques for analyzing the auto-convolution problem. Some regularization methods and specific numerical approximations applied to the first kind of integral equations as well as the equation (1) were also discussed in [10, 11, 12, 7, 18, 2, 22, 3, 1, 24, 26, 28, 29, 30, 4, 13, 25, 16]. The auto-convolution equation (1) is an interesting example of a nonlinear ill-posed operator equation with interesting properties. We can consider important properties from [9] in the following form:

The auto-convolution operator $F : L^2(0,1) \longrightarrow L^2(0,1)$ is defined by

$$[F(u)](t) := \int_0^t u(t-s)u(s)ds, \qquad 0 \le t \le 1,$$

with a domain

$$D(F) := \{u \in L^2(0,1) : u(t) \ge 0 \text{ a.e. } on\ [0,1]\}. \tag{2}$$

The operator F is continuous, weakly closed but not compact; see [9]. On the other hand, for all $u^* \in D(F)$, the Fréchet derivative of F as a compact linear operator can be defined by

$$[F'(u^*)u](t) := 2\int_0^t u^*(t-s)u(s)ds, \qquad 0 \le t \le 1,$$

and also, we have

$$\|F(u) - F(u^*) - F'(u^*)(u-u^*)\|_{L^2(0,1)} \le \|u-u^*\|^2_{L^2(0,1)}.$$

Using the Titchmarsh convolution theorem [20], the injectivity of the auto-convolution operator F and solutions of (1) can be given by the following theorem.

Theorem 1. [6] For any $u \in L^2(0,1)$, let

$$\varepsilon(u) := \sup\{0 \le \epsilon \le 1 : u(t) = 0 \text{ a.e. } on\ [0,\epsilon]\}.$$

Then the auto-convolution equation (1) together with (2) has a unique solution if and only if $\varepsilon(g) = 0$. If $u^* \in D(F)$ is the uniquely determined solution, then it fulfills the condition $\varepsilon(u^*) = 0$.

In this paper, we consider the generalized AVIEs of the first kind as

$$\int_0^t K(t,s)u(t-s)u(s)ds = g(t), \qquad t \in I := [0,1], \tag{3}$$

where g, K are given functions and $u(t)$ is unknown function. Recently, in [8], the authors have studied an ill-posed inverse problem of auto-convolution equation such that this inverse problem occurs in nonlinear optics in the context of ultrashort laser pulse characterization. They applied an iterative regularization approach, which is specifically adapted to the physical situation in pulse characterization, using a nonstandard stopping rule for the iteration process of computing regularized solutions. In [27], an AVIE of the second kind was solved by using the collocation method according to piecewise polynomials, and the (super) convergence of the mentioned method has been studied. However, as far as we know, it seems to be an open problem under what conditions we can consider for the kernel in the integral equation (3) to investigate the solutions of this equation similar to Theorem 1.

Here, we consider the numerical collocation method based on piecewise polynomials to solve (3). According to the proposed numerical method, we have to distinguish between two cases, nonlinear and linear systems. Since the auto-convolution problem (3) is ill-posed, for given g, the solutions u need not be uniquely determined and mainly small perturbations in the right-hand side g caused by noisy data may lead to arbitrarily large errors in the solution. To overcome the negative consequences of ill-posedness, we

consider a novel double iteration process from [21] and the Lavrentiev $\tilde{m}$ times iterated regularization method [14] to deal with the ill-posed nonlinear and linear algebraic systems, respectively.

2 Collocation methods

Let Π_h be a uniform partition of the interval I with grid points

$$t_n = nh, \qquad n = 0, \ldots, N,$$

and let h be the stepsize. We consider polynomial spline approximations $u_h(t)$ of the exact solution $u(t)$ in the spline space

$$S_{m-1}^{-1}(\Pi_h) = \{v : v|_{\sigma_n} \in \pi_{m-1}(0 \leq n \leq N-1)\}, \tag{4}$$

where $v|_{\sigma_n}$ is the restriction of v to the subinterval $\sigma_n = [t_n, t_{n+1}]$, and $\pi_{m-1}(m \geq 1)$ denotes the set of real polynomials of degree not exceeding $m-1$. Let the collocation parameters be $0 < c_1 < c_2 < \cdots < c_m \leq 1$ and let the collocation points be $t_{nj} = t_n + c_j h, j = 1, \ldots, m, n = 0, \ldots, N-1$. Hence, our collocation solution $u_h \in S_{m-1}^{-1}(\Pi_h)$ is determined by

$$g(t) = \int_0^t K(t,s) u_h(t-s) u_h(s) ds, \tag{5}$$

where the collocation equation is satisfied for $t = t_{nj}, j = 1, \ldots, m, n = 0, \ldots, N-1$. We consider the approximate solution on the subinterval σ_n as

$$u_h|_{\sigma_n} = u_h^n(t_n + sh) = \sum_{j=1}^{m} L_j(s) U_{n,j}, \qquad s \in (0,1], \tag{6}$$

where $U_{n,j} := u_h(t_n + c_j h)$ and

$$L_j(s) := \prod_{i=1, i\neq j}^{m} \frac{s - c_i}{c_j - c_i}, \qquad j = 1, \ldots, m.$$

Inserting (6) into (5), we obtain

$$\begin{aligned} g(t_{n,i}) &= \int_0^{t_{n,i}} K(t_{n,i}, s) u_h(t_{n,i} - s) u_h(s) ds \\ &= \int_{t_n}^{t_{n,i}} K(t_{n,i}, s) u_h^0(t_{n,i} - s) u_h^n(s) ds \\ &\quad + \sum_{k=0}^{n-1} \int_{t_k}^{t_{k,i}} K(t_{n,i}, s) u_h^{n-k}(t_{n,i} - s) u_h^k(s) ds \end{aligned}$$

$$+\sum_{k=0}^{n-1}\int_{t_{k,i}}^{t_{k+1}} K(t_{n,i},s)u_h^{n-k-1}(t_{n,i}-s)u_h^k(s)ds. \tag{7}$$

System (7) for $n=0$ is a nonlinear system, and for $n=1,\ldots,N-1$, it is a linear system. Then we have to distinguish between two cases, $n=0$ and $n=1,\ldots,N-1$.

2.1 Double iteration process to solve ill-posed nonlinear algebraic systems

Case I: $n=0$. Now (7) becomes

$$\begin{aligned} g(c_ih) =& \int_0^{c_ih} K(c_ih,s)u_h^0(c_ih-s)u_h^0(s)ds \\ =& h\int_0^{c_i} K(c_ih,sh)u_h^0((c_i-s)h)u_h^0(sh)ds \\ =& h\sum_{j,k=1}^{m}\int_0^{c_i} K(c_ih,sh)L_j(c_i-s)L_k(s)ds\ u_h^0(c_jh)u_h^0(c_kh) \\ =& h\sum_{j,k=1}^{m}\alpha_{jk}^{(i)}u_h^0(c_jh)u_h^0(c_kh) = h\sum_{j,k=1}^{m}\alpha_{jk}^{(i)}U_{0,j}U_{0,k}, \end{aligned} \tag{8}$$

where

$$\alpha_{jk}^{(i)} = \int_0^{c_i} K(c_ih,sh)L_j(c_i-s)L_k(s)ds.$$

Nonlinear system (8) can be written as

$$f(U_0)-G_0=0, \quad or \quad F(U_0)=0, \tag{9}$$

where

$$U_0=(U_{0,1},\ldots,U_{0,m})^T\in\mathbb{R}^m, \qquad G_0=(g(c_1h),\ldots,g(c_mh))^T.$$

Let us assume that g^δ is an available approximation of g satisfying

$$\|g^\delta-g\|\le\delta,$$

where δ is a given noise level.

Now, we consider a novel double iteration process from [21] to deal with the ill-posed nonlinear algebraic equation system (9). The proposed method combines the residual norm based algorithm and the modified Tikhonov's regularization method. For the outer iteration process, the evolution path

of the unknown vector follows the searching direction determined from the inner iteration process, and the process requires the path falls on the space-time manifold such that the convergence rate can be guaranteed. For the inner iteration, the direction of evolution $\tilde{u}$ is determined by solving a linear algebraic equation

$$B^T B\tilde{u} = B^T F,$$

where B is the Jacobian matrix. For an ill-posed system, this linear algebraic equation is very difficult to solve since the resulting leading coefficient matrix is ill-posed in nature. The modified Tikhonov's regularization method [15] is adapted to solve the ill-posed linear algebraic equation. However, for ill-posed problems to really find the solution of $\tilde{u}$ may require too many iteration steps for the modified Tikhonov's regularization method, which makes the whole numerical process not economical at all. Therefore, the inner iteration process stops while the direction $\tilde{u}$ makes the value of $\tilde{a}_0$ smaller than the selected margin $\tilde{a}_c$ or when the number of inner iteration steps exceeds the maximum tolerance $\tilde{I}_{\max}$. For the outer iteration process, it terminates once the root mean square error for the residual is less than the convergence criterion ε or when the number of inner iteration steps exceeds the maximum tolerance $\tilde{I}_{\max}$.

Now, first, we consider a perturbed version of the system (9) as

$$F^\delta(U_0^\delta) = 0, \tag{10}$$

where $F^\delta(U_0^\delta) = f(U_0^\delta) - G_0^\delta$. Then, we proposed a double iteration process for system (9) based on the following algorithm.

Remark 1. Let $\theta_1 = (x_1, \ldots, x_n)^T$ and let $\theta_2 = (y_1, \ldots, y_n)^T$. Then the root mean squared error (RMSE) is defined as

$$REMS(\theta_1, \theta_2) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2}.$$

Remark 2. From [21], the value of $\tilde{a}_c$ is suggested to be in the range of 2.5 to 4, and the value of $\tilde{I}_{\max}$ actually depends on the selection of $\tilde{a}_c$ such that it is suggested to be in the range of 30,000 to 80,000. The selection of ε depends on the system we want to solve. If the system is a well-posed system, then the value of ε can be very small, and also, if the system is an ill-posed system, then the value of ε should not be very big.

Algorithm 1 Double Iteration Process (DIP)

Step 1. Give initial guess $U_0^{\delta,0}$ and ε , α, $\tilde{I}_{\max}$, $\tilde{a}_c$.

Step 2.Outer Iteration: For $k = 0, 1, 2, \ldots$ Repeat

Compute $F_k^\delta = F^\delta(U_0^{\delta,k})$ and the Jacobian matrix $B_k = B(U_0^{\delta,k})$.

Step 3. Inner Iteration: Give the initial guess of $\tilde{u}$ as $\tilde{u}_0 = \dfrac{B_k^T F_k^\delta}{||B_k^T F_k^\delta||}$,

For $p = 1, 2, \ldots, \tilde{I}_{\max}$, solve $\tilde{u}_{p+1}$ by $(B_k^T B_k + \alpha I)\tilde{u}_{p+1} = B_k^T F_k^\delta + \alpha \tilde{u}_p$,

Compute $\tilde{v}_{p+1} = B_k \tilde{u}_{p+1}$,

$\eta_{p+1} = \frac{||F_k^\delta||^2 ||\tilde{v}_{p+1}||^2}{((F_k^\delta)^T \tilde{v}_{p+1})^2}$,

(a) If $\eta_{p+1} \leq \tilde{a}_c$, then $\eta_k = \eta_{p+1}$, $\tilde{u}_k = \tilde{u}_{p+1}$, $\tilde{v}_k = \tilde{v}_{p+1}$ and one terminates the inner iteration; otherwise, continue.

(b) If $p = \tilde{I}_{\max}$, then terminate the whole process.

Step 4. End of Inner Iteration:

Calculate $r_k = ||1 - \dfrac{\eta_k}{2}||$,

$U_0^{\delta,k+1} = U_0^{\delta,k} - (1 - r_k)\dfrac{(F_k^\delta)^T \tilde{v}_k}{||\tilde{v}_k||^2}\tilde{u}_k$.

If $RMSE \leq \varepsilon$ or (b) is true, then the outer iteration process stop; otherwise, continue.

End of Iteration Process.

2.2 The Lavrentiev regularization method to solve ill-posed linear algebraic system

Case I: For $n = 1, \ldots, N-1$, we can rewrite system (7) as

$$\begin{aligned} g(t_{n,i}) =& h\int_0^{c_i} K(t_{n,i}, t_n + sh)u_h^0((c_i - s)h)u_h^n(t_n + sh)ds \\ &+ h\int_0^{c_i} K(t_{n,i}, sh)u_h^n(t_n + (c_i - s)h)u_h^0(sh)ds \end{aligned} \tag{11}$$

$$+h\sum_{l=1}^{n-1}\int_0^{c_i}K(t_{n,i},t_l+sh)u_h^{n-l}(t_{n-l}+(c_i-s)h)u_h^l(t_l+sh)ds$$
$$+h\sum_{l=0}^{n-1}\int_{c_i}^{1}K(t_{n,i},t_l+sh)u_h^{n-l-1}(t_{n-l}+(c_i-s)h)u_h^l(t_l+sh)ds.$$

Using (6), we have

$$h\sum_{j,k=1}^{m}\int_0^{c_i}K(t_{n,i},t_n+sh)L_j(c_i-s)L_k(s)dsU_{0,j}U_{n,k} \tag{12}$$
$$+h\sum_{j,k=1}^{m}\int_0^{c_i}K(t_{n,i},sh)L_j(s)L_k(c_i-s)dsU_{0,j}U_{n,k}$$
$$=g(t_{n,i})-h\sum_{l=1}^{n-1}\sum_{j,k=1}^{m}\int_0^{c_i}K(t_{n,i},t_l+sh)L_j(c_i-s)L_k(s)dsU_{n-l,j}U_{l,k}$$
$$-h\sum_{l=0}^{n-1}\sum_{j,k=1}^{m}\int_{c_i}^{1}K(t_{n,i},t_l+sh)L_j(1+c_i-s)L_k(s)dsU_{n-l-1,j}U_{l,k}.$$

Linear system (12) can be written as

$$A_nU_n=Y_n, \tag{13}$$

where

$$U_n=(U_{n,1},\ldots,U_{n,m})^T\in\mathbb{R}^m,\qquad G_n=(g(t_{n,1}),\ldots,g(t_{n,m}))^T,$$

and $Y_n=\hat{f}(U_0,\ldots,U_{n-1})+G_n$. Since $U_0,\ldots,U_{n-1}$ have been determined in the previous steps, then Y_n is a definite m-dimensional vector. A perturbed version of the discrete system (13) may be considered as follows:

$$(A_n+\bar{\alpha}_nI)U_n^\delta=Y_n^\delta. \tag{14}$$

Now , we apply the Lavrentiev $\tilde{m}$ times iterated regularization method [14] for the fixed integer $\tilde{m}\geq 1$, and the regularization parameter $\bar{\alpha}_n>0$ for the equation (13), which determines U_n^δ based on the following algorithm.

Further issues related to theoretical analysis of the Lavrentiev-iterated method can be found in the references [14, 17].

3 Numerical results

Two examples are considered in this section, which provides evidence of the effectiveness of the proposed numerical methods applied to the first-kind

Algorithm 2 The Lavrentiev $\tilde{m}$ times iterated regularization method

Step 1. Give N, $\tilde{m}$.

Step 2.Outer Iteration: For $n = 1, 2, \ldots, N-1$ Repeat

Determine A_n and Y_n^δ
Choose initial guess $U_n^{\delta,0}$ and $\bar{\alpha}_n$.

Step 3. Inner Iteration: For $p = 1, 2, \ldots, \tilde{m}$,

solve $U_n^{\delta,p}$ by $(A_n + \bar{\alpha}_n I)U_n^{\delta,p} = \bar{\alpha}_n U_n^{\delta,p-1} + Y_n^\delta$

If $RMSE(U^{\delta,p}, U^{\delta,p-1}) \nrightarrow 0$, then stop the iteration, change $\bar{\alpha}_n$, and consider the inner iteration with the new $\bar{\alpha}_n$.

Step 4. End of Inner Iteration:

put $U_n^\delta = U_n^{\delta,\tilde{m}}$

End of Iteration Process.

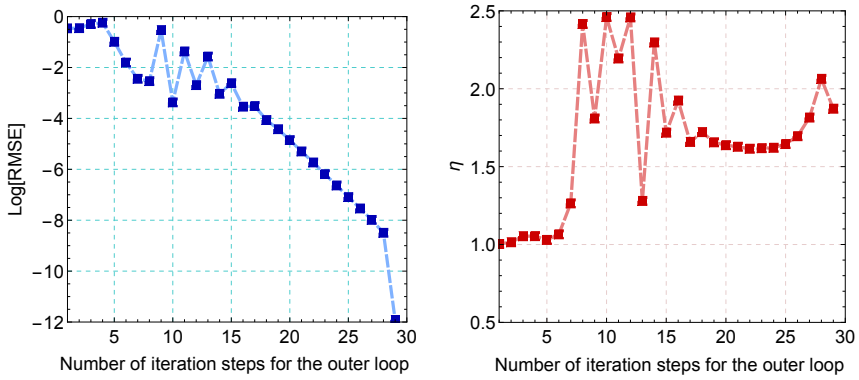


Figure 1: $Log[RMSE]$ versus the number of iteration steps with $N = 4, \delta = 0$ for the outer loop in Example 1 (left). The value of η never exceeds the critical value $\tilde{a}_c = 2.5$ in Example 1 (right).

auto-convolution equation. We produce the perturbed data g^δ by adding uniformly distributed noise from the interval $[-\delta g, \delta g]$ to the discrete values of $g(t)$ for $t = t_{n,i}, n = 0, \ldots, N-1, i = 1, \ldots, m$. All the computations were performed by Wolfram Mathematica 10.0.

Example 1. Consider the AVIE of the first kind as follows:

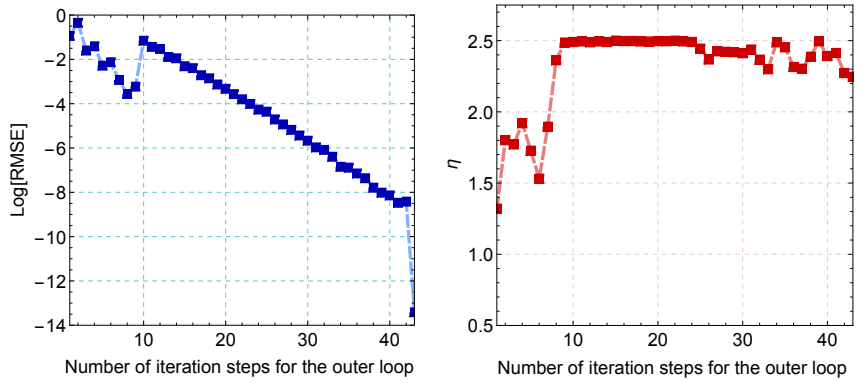


Figure 2: $Log[RMSE]$ versus the number of iteration steps with $N = 8, \delta = 0$ for the outer loop in Example 1 (left). The value of η never exceeds the critical value $\tilde{a}_c = 2.5$ in Example 1 (right).

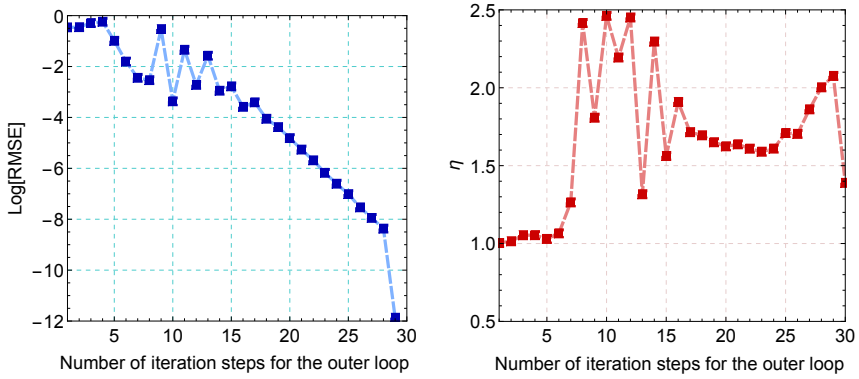


Figure 3: $Log[RMSE]$ versus the number of iteration steps with $N = 4, \delta = 10^{-6}$ for the outer loop in Example 1 (left). The value of η never exceeds the critical value $\tilde{a}_c = 2.5$ in Example 1 (right).

$$g(t) = \int_0^t \cos(t-s)u(t-s)u(s)ds, \qquad t \in [0,1], \tag{15}$$

where $g(t) = -2500e^{-t}(t + t\cos t - 2\sin t)$ and the exact solution $u(t) = 50te^{-t}$.

Let $m = 3$, let $N = 4, 8$ and let the collocation parameters be $c_1 = 0.4, c_2 = 0.6, c_3 = 1$. For $n = 0$, we have the following nonlinear system:

$$g^{\delta}(c_i h) = h \sum_{j,k=1}^{3} \alpha_{jk}^{(i)} U_{0,j}^{\delta} U_{0,k}^{\delta}, \qquad i = 1, 2, 3, \tag{16}$$

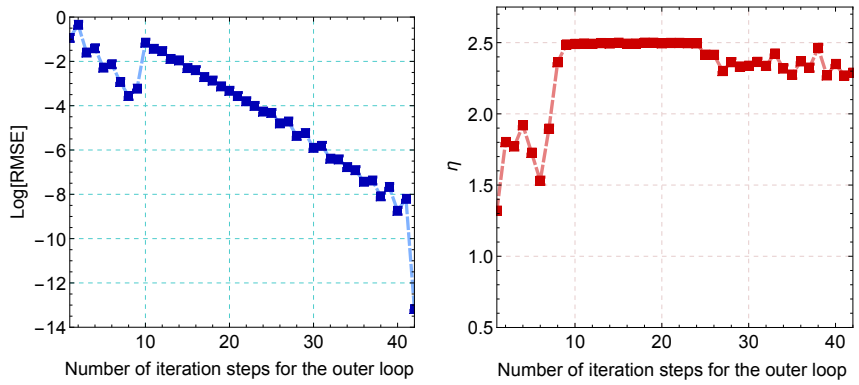


Figure 4: $Log[RMSE]$ versus the number of iteration steps with $N = 8, \delta = 10^{-6}$ for the outer loop in Example 1 (left). The value of η never exceeds the critical value $\tilde{a}_c = 2.5$ in Example 1 (right).

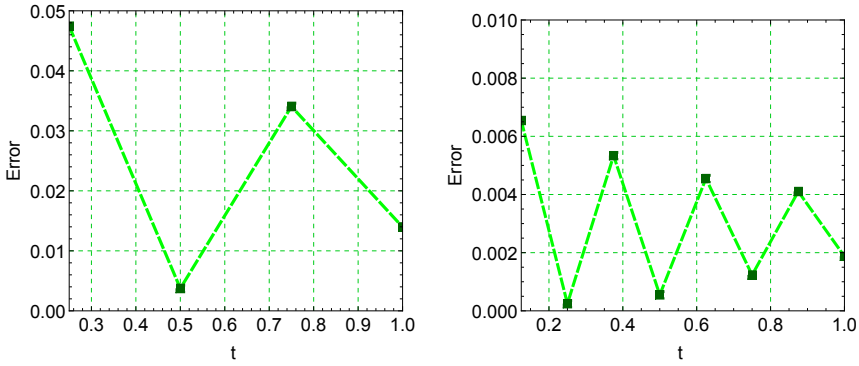


Figure 5: The absolute errors with $N = 4, \delta = 0$ at the grid-points in Example 1 (left). The absolute errors with $N = 8, \delta = 0$ at the grid-points in Example 1 (right).

where

$$\alpha_{jk}^{(i)} = \int_0^{c_i} \cos(c_i h - sh) L_j(c_i - s) L_k(s) ds, \qquad i, j, k = 1, 2, 3.$$

For $\delta = 0, 10^{-6}$, we use the double iteration process (Algorithm 1) to solve this problem with the initial guess $U_0^{\delta,0} = (1, 1, 1)^T$ and the parameters $\alpha = 0.1, \tilde{I}_{\max} = 30000, \tilde{a}_c = 2.5$ and $\epsilon = 10^{-4}$. Figures 1–4 show that for $N = 4$ and $N = 8$ with $\delta = 0, 10^{-6}$ the process terminates after 29 and 43 steps for the outer loop, respectively. Also, we observe that η never exceeds the critical value $\tilde{a}_c = 2.5$, which once more shows that this method really can guarantee the trajectory of the solution vector falls on the manifold. We report the solutions of system (16) in Table 1 and use these values to solve

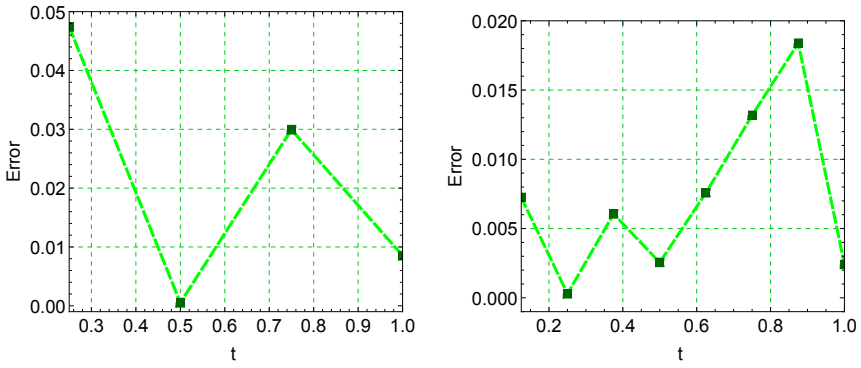


Figure 6: The absolute errors with $N = 4, \delta = 10^{-6}$ at the grid-points in Example 1 (left). The absolute errors with $N = 8, \delta = 10^{-6}$ at the grid-points in Example 1 (right).

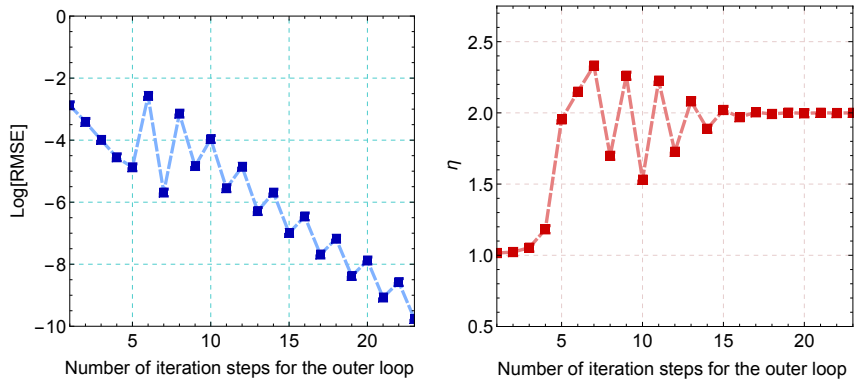


Figure 7: $Log[RMSE]$ versus the number of iteration steps with $N = 4, \delta = 0$ for the outer loop in Example 2 (left). The value of η never exceeds the critical value $\tilde{a}_c = 2.5$ in Example 2 (right).

the following linear system by the Lavrentiev $\tilde{m}$ times iterated regularization method, for $n = 1, \ldots, N-1$:

$$h \sum_{j,k=1}^{3} \int_0^{c_i} \cos[t_{n,i} - (t_n + sh)] L_j(c_i - s) L_k(s) ds U_{0,j}^{\delta} U_{n,k}^{\delta} \tag{17}$$

$$+ h \sum_{j,k=1}^{3} \int_0^{c_i} \cos[t_{n,i} - sh] L_j(s) L_k(c_i - s) ds U_{0,j}^{\delta} U_{n,k}^{\delta}$$

$$= g^{\delta}(t_{n,i}) - h \sum_{l=1}^{n-1} \sum_{j,k=1}^{3} \int_0^{c_i} \cos[t_{n,i} - (t_l + sh)] L_j(c_i - s) L_k(s) ds U_{n-l,j}^{\delta} U_{l,k}^{\delta}$$

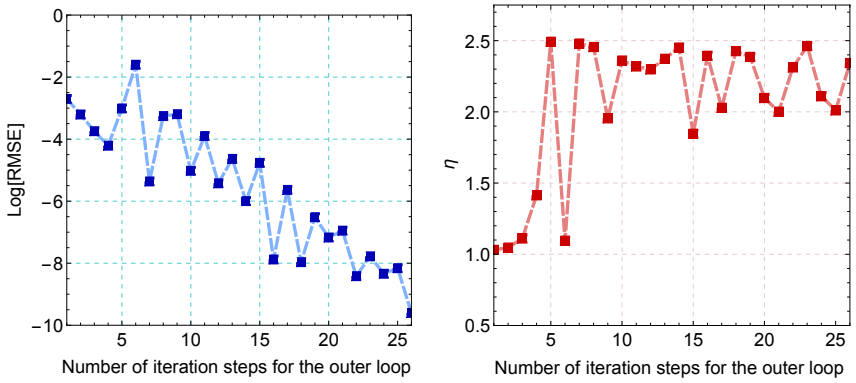


Figure 8: $Log[RMSE]$ versus the number of iteration steps with $N = 8, \delta = 0$ for the outer loop in Example 2 (left). The value of η never exceeds the critical value $\tilde{a}_c = 2.5$ in Example 2 (right).

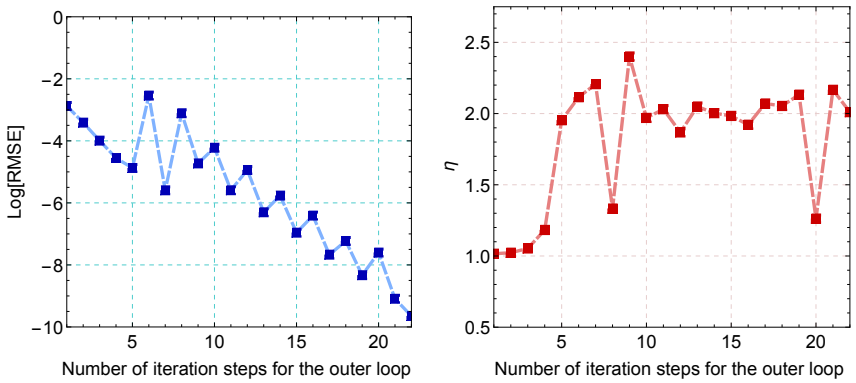


Figure 9: $Log[RMSE]$ versus the number of iteration steps with $N = 4, \delta = 10^{-4}$ for the outer loop in Example 2 (left). The value of η never exceeds the critical value $\tilde{a}_c = 2.5$ in Example 2 (right).

$$- h \sum_{l=0}^{n-1} \sum_{j,k=1}^{3} \int_{c_i}^{1} \cos[t_{n,i} - (t_l + sh)] L_j(1 + c_i - s) L_k(s) ds U^{\delta}_{n-l-1,j} U^{\delta}_{l,k},$$

$$i = 1, 2, 3.$$

In Algorithm 2, we choose initial guess $U_n^{\delta,0} = (0, 0, 0)^T, \bar{\alpha}_n = 0.01$ and $\tilde{m} = 20$. The maximum absolute errors for $N = 4, 8$ and $\delta = 0, 10^{-6}$ have been reported in Table 2. Also, Figures 5 and 6 show the absolute errors at the grid-points for $N = 4, 8$ and $\delta = 0, 10^{-6}$.

Now, we consider a special case of (3) with $K(t, s) = 1$ from [3] as follows.

Example 2. [3] Consider the AVIE of the first kind as follows:

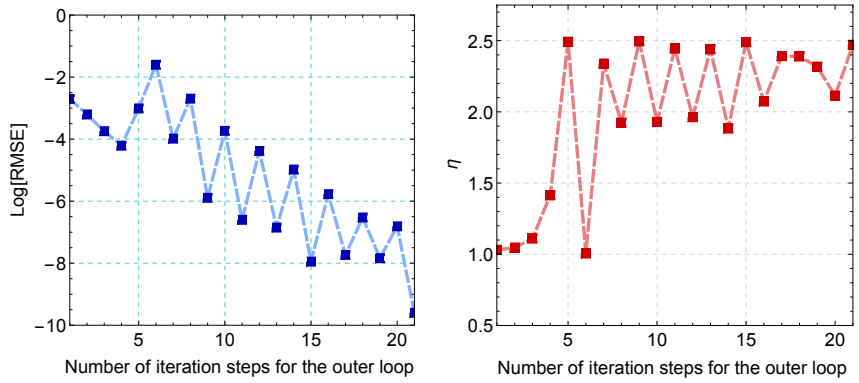


Figure 10: $Log[RMSE]$ versus the number of iteration steps with $N = 8, \delta = 10^{-4}$ for the outer loop in Example 2 (left). The value of η never exceeds the critical value $\tilde{a}_c = 2.5$ in Example 2 (right).

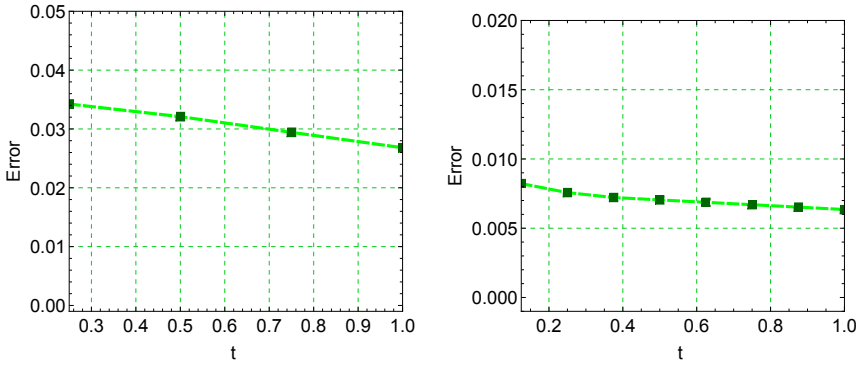


Figure 11: The absolute errors with $N = 4, \delta = 0$ at the grid-points in Example 2 (left). The absolute errors with $N = 8, \delta = 0$ at the grid-points in Example 2 (right).

	$N = 4$ $\delta = 0$	$N = 4$ $\delta = 10^{-6}$	$N = 8$ $\delta = 0$	$N = 8$ $\delta = 10^{-6}$
$U^{\delta}_{0,1}$	4.52533	4.52527	2.37834	2.37832
$U^{\delta}_{0,2}$	6.46103	6.46100	3.47984	3.47980
$U^{\delta}_{0,3}$	9.68703	9.68729	5.50843	5.50838

Table 1: The solution of system (16) in Example 1

$$g(t) = \int_0^t u(t-s)u(s)ds, \qquad t \in [0,1], \tag{18}$$

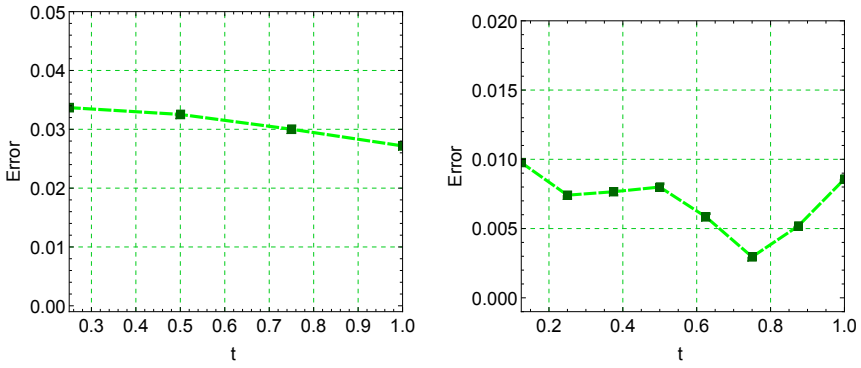


Figure 12: The absolute errors with $N = 4, \delta = 10^{-4}$ at the grid-points in Example 2 (left). The absolute errors with $N = 8, \delta = 10^{-4}$ at the grid-points in Example 2 (right).

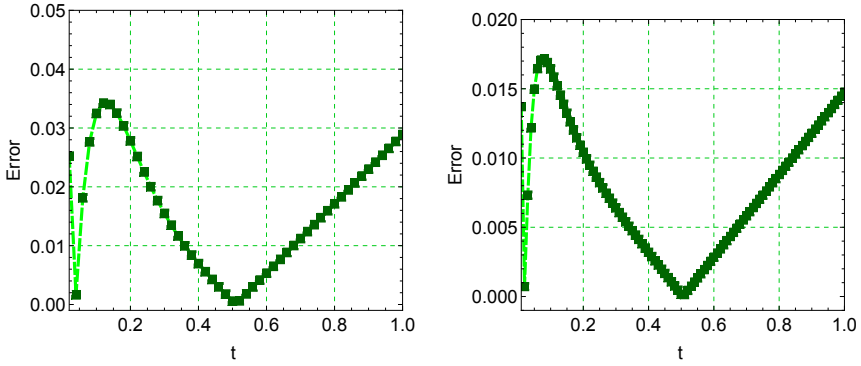


Figure 13: The absolute errors obtained by method 1 [3] with $N = 50, r = 4, \delta = 0$ at the grid-points in Example 2 (left). The absolute errors obtained by method 1 [3] with $N = 100, r = 4, \delta = 0$ at the grid-points in Example 2 (right).

N	$\delta = 0$	$\delta = 10^{-6}$
4	4.74×10^{-2}	4.74×10^{-2}
8	6.54×10^{-3}	1.83×10^{-2}

Table 2: The maximum absolute errors in Example 1

where $g(t) = \frac{3}{10}t^5 - \frac{3}{2}t^4 + t^3 + \frac{3}{4}t^2 + \frac{1}{16}t$ and the exact solution $u(t) = 1 - 3(t - \frac{1}{2})^2$.

In Example 2, assume that $m = 2, N = 4, 8, \delta = 0, 10^{-4}$, and the collocation parameters as $c_1 = 0.6, c_2 = 1$. For $n = 0$, we solve the obtained nonlinear system by the double iteration process (Algorithm 1) with the ini-

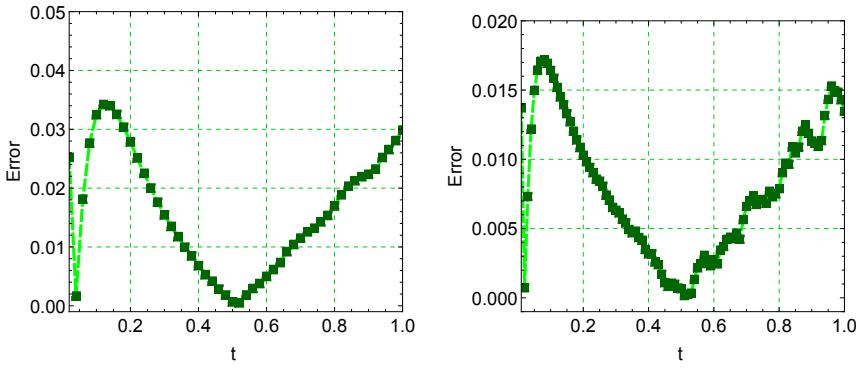


Figure 14: The absolute errors obtained by method 1 [3] with $N = 50, r = 4, \delta = 10^{-4}$ at the grid-points in Example 2 (left). The absolute errors obtained by method 1 [3] with $N = 100, r = 4, \delta = 10^{-4}$ at the grid-points in Example 2 (right).

tial guess $U_0^{\delta,0} = (1,1)^T$ and the parameters $\alpha = 0.1, \tilde{I}_{\max} = 30000, \tilde{a}_c = 2.5$ and $\epsilon = 10^{-4}$. In what follows, for $n = 1, \ldots, N-1$, we solve the obtained linear system by the Lavrentiev $\tilde{m}$ times iterated regularization method with initial guess $U_n^{\delta,0} = (0,0)^T, \bar{\alpha}_n = 0.01$ and $\tilde{m} = 20$. The maximum absolute errors for $N = 4, 8$ and $\delta = 0, 10^{-4}$ have been reported in Table 3.

N	$\delta = 0$	$\delta = 10^{-4}$
4	3.42×10^{-2}	3.56×10^{-2}
8	8.22×10^{-3}	9.76×10^{-3}

Table 3: The maximum absolute errors in Example 2

We also consider a local regularization technique from [3] by method 1 for the auto-convolution integral equation of the first kind (18) and report the maximum absolute errors for $N = 50, 100$, $r = 4$ and $\delta = 0, 10^{-4}$ in Table 4.

N	$\delta = 0$	$\delta = 10^{-4}$
50	3.42×10^{-2}	3.43×10^{-2}
100	1.71×10^{-2}	3.01×10^{-2}

Table 4: The maximum absolute errors obtained by method 1 [3] in Example 2

Similar to Example 1, in Figures 7–10, we show the number of the termination of the process in the outer loop and observe that η never exceeds the critical value $\tilde{a}_c = 2.5$. Also, Figures 11 and 12 represent the absolute errors at the grid-points for $N = 4, 8$ and $\delta = 0, 10^{-4}$. The absolute errors obtained

by method 1 [3] with $N = 50, 100, r = 4, \delta = 0, 10^{-4}$ at the grid-points are shown in Figures 13 and 14.

4 Conclusion

In this paper, we considered the numerical collocation method based on piecewise polynomials to solve the generalized version of the AVIE of the first kind as an ill-posed problem. We obtained two ill-posed nonlinear and linear systems. The double iteration process and the Lavrentiev $\tilde{m}$ times iterated regularization method have been considered nonlinear and linear systems, respectively. Numerical examples demonstrated the validity and efficiency of the proposed method. In our future work, we will investigate the existence, uniqueness, and structure of the solution related to the generalized version of the first-kind AVIEs similar to Theorem 1.

References

[1] Brunner, H. *Volterra integral equations: An introduction to theory and applications,* Cambridge University Press, Cambridge/UK, 2017.

[2] Choi, K. and Lanterman, A.D. *An iterative deautoconvolution algorithm for nonnegative functions,* Inverse Prob., 21 (2005) 981–995.

[3] Dai, Z. and Lamm P.K. *Local regularization for the nonlinear inverse autoconvolution problem,* SIAM J. Numer. Anal., 46 (2008) 832–868.

[4] Feng, L. and Yang, X. *Spectral regularization method for Volterra integral equation of the first kind with noise data,* Acta Math. Sci. Ser. A, 40 (3) (2020) 650–661.

[5] Fleischer, G. and Hofmann, B. *On inversion rates for the auto convolution equation,* Inverse Prob. 12 (1996) 419–435.

[6] Fleischer, G. and Hofmann, B. *The local degree of ill-posedness and the autoconvolution equation,* Nonlinear Anal. 30 (1997) 3323–3332.

[7] Fleischer, G., Gorenflo, R. and Hofmann, B. *On the autoconvolution equation and total variation constraints,* Z. Angew. Math. Mech., 79 (1999) 149–159.

[8] Gerth, D., Hofman, B., Birkholz, S. and Koke, S.*Regularization of an autoconvolution problem in ultrashort laser pulse characterization,* Inverse Probl. Sci. Eng. 22(2) (2013) 245–266.

[9] Gorenflo, R. and Hofmann, B. *On autoconvolution and regularization,* Inverse Prob., 10 (1994) 353–373.

[10] Janno, J. *On a regularization method for the autoconvolution equation,* Z. Angew. Math. Mech., 77 (1997) 393–394.

[11] Janno, J. *Lavrent'ev regularization of ill-posed problems containing nonlinear near-to-monotone operators with application to auto-convolution equation,* Inverse Prob., 16 (2000) 333–348.

[12] Jozi, M. and Karimi, S. *Direct implementation of Tikhonov regularization for the first kind integral equation* J. Comp. Math., 40, (2022) 335–353.

[13] Karakeev, T.T. and Imanaliev T. M. *Regularization of Volterra linear integral equations of the first kind with the smooth data,* Lobachevskii J. Math., 41 (2020) 39–45.

[14] Lamm, P.K. *A survey of regularization methods for first-kind Volterra equations.* Colton D., Engl H. W., Louis H. K., McLaughlin J. R., Rundell W, eds. Surveys on Solution Methods for Inverse Problems. Vienna: Springer, (2000) 53–82.

[15] Liu, C.S. *Optimally generalized regularization methods for solving linear inverse problems,* CMC: Computers, Materials, Continua, vol. 29 (2012) 103–127.

[16] Masouri, Z. and Hatamzadeh, S. *A regularization-direct method to numerically solve first kind Fredholm integral equation,* KYUNGPOOK Math. J. 60 (2020) 869–881.

[17] Plato, R. *Iterative and parametric methods for linear ill-posed problems,* TU Berlin, 1995.

[18] Ramlau, R. *Morozov's discrepancy principle for Tikhonov-regularization of nonlinear operators,* Numer. Funct. Anal. Optim., 23 (2002) 147–172.

[19] Richter, M. *Approximation of Gussian Random Element and Statistics,* Stuttgart, B.G. Teubner, 1992.

[20] Titchmarsh, E.C. *The zeros of certain integral functions,* Proc. London Math. Society. 25 (1926) 283–302.

[21] Weichung, Y., Chan I-Yao, Cheng-Yu Ku,Chia-Ming Fan, Pai-Chen G *A double iteration process for solving the nonlinear algebraic equations, especially for ill-posed nonlinear algebraic equations,* J. Comput. Model. Eng. Sci. 99(2) (2014) 123–149.

[22] Wolfersdorf, L.V. *Autoconvolution equations and special functions,* Integral Transforms Spec. Funct., 19 (2008) 677–686.

[23] Wolfersdorf, L.V. *Einige Klassen quadratischer Integralgleichungen [Some classes of quadratic integral equations],* Sitz.ber. Sachs. Akad. Wiss. Leipz. Math.-Nat.wiss. Kl. 2000.

[24] Wolfersdorf, L.V. *Auto convolution equations and special functions II,* Integral Transforms Spec. Funct., 21(4) (2010) 295–306.

[25] Ziyaee, F. and Tari, A. *Regularization method for the two-dimensional Fredholm integral equations of the first kind,* Int. J. Nonlinear Sci., 18 (2014) 189–194.

[26] Zhang, Y., Forssen, P., Fornstedt, T., Gulliksson M. and Dai X. *An adaptive regularization algorithm for recovering the rate constant distribution from biosensor data,* Inverse Probl. Sci. Eng., 26 (2018) 1464–1489.

[27] Zhang, R., Liang, H. and Brunner, H. *Analysis of collocation methods for generalized autoconvolution Volterra integral equations, SIAM J. Numer. Anal.*, Vol. 54 (2) (2016) 899–920.

[28] Zhang, Y., Lukyanenko, D. and Yagola, A. *Using Lagrange principle for solving two-dimensional integral equation with a positive kernel,* Inverse Probl. Sci. Eng., 24(5) (2016) 811–831.

[29] Zhang, Y., Lukyanenko, D. and Yagola, A. *An optimal regularization method for convolution equations on the sourcewise represented set,* J Inverse Ill-Pose P., 23(5) (2016) 465–475.

[30] Zhang. Y., Gulliksson, M., Hernandez, B.V. and Schaffernicht, E. *Reconstructing gas distribution maps via an adaptive sparse regularization algorithm,* Inverse Probl. Sci. Eng,. 24(7) (2016) 1186–1204.

How to cite this article
Pishbin, S. and Ebadi, A., Regularization of the generalized auto-convolution Volterra integral equation of the first kind. *Iran. j. numer. anal. optim.*, 2023; 13(2): 205-223. https://doi.org/10.22067/ijnao.2022.76652.1135

Iranian Journal of Numerical Analysis and Optimization
Vol. 13, No. 2, 2023, pp 224–242
https://doi.org/10.22067/ijnao.2022.77393.1158
https://ijnao.um.ac.ir/

CC BY NC

How to cite this article
Research Article

The interpolated variational iteration method for solving a class of nonlinear optimal control problems

Mohammad Shirazian*

Abstract

Despite the variety of methods available to solve nonlinear optimal control problems, numerical methods are still evolving to solve these problems. This paper deals with the numerical solution of nonlinear optimal control affine problems by the interpolated variational iteration method, which was introduced in 2016 to improve the variational iteration method. For this purpose, the optimality conditions are first derived as a two-point boundary value problem and then converted to an initial value problem with the unknown initial values for costates. The speed and convergence of the method are compared with the existing methods in the form of three examples, and the initial values of the costates are obtained by an efficient technique in each iteration.

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

Keywords: Optimal control problems; Interpolated variational iteration method; Pontryagin's maximum principle.

1 Introduction

One of the important problems in the field of engineering [13, 6], robotics [14, 11], biology [1, 16], and so on, is finding the best control strategy to use available resources optimally. In practice, most of these problems are formulated nonlinearly, which in general do not have analytical solutions, and

*Corresponding author

Received 27 June 2022; revised 15 September 2022; accepted 8 October 2022

Mohammad Shirazian
Department of Mathematics, University of Neyshabur, Neyshabur, Iran. e-mail: shirazian@neyshabur.ac.ir

one must look for reliable numerical solutions for them. Generally, consider the following nonlinear control system:

$$\begin{aligned} &\dot{x}(t) = A(x(t)) + B(t)u(t), \qquad t \in [t_0, t_f], \\ &x(t_0) = x_0, \quad x(t_f) = x_f, \end{aligned} \tag{1}$$

where $x(t)$ and $u(t)$ are n- and m-vector of state and control functions, respectively. The aim is to minimize the objective functional

$$J[x,u] = \frac{1}{2}\int_{t_0}^{t_f} (x^T(t)Q(t)x(t) + u^T(t)R(t)u(t))dt, \tag{2}$$

subject to the dynamical system (1). Moreover, $Q(t)$ is a positive semi-definite real matrix, and $R(t)$ is a positive definite real matrix.

There are two approaches to solving this problem. One is the direct method of parameterizing the state and control functions and solving a nonlinear optimization problem. The other is the indirect method of using optimality conditions, derived by the Pontryagin maximum principle (PMP) or Hamilton–Jacobi–Bellman equation [3, 15]. The main topic of our discussion in this paper is the optimality conditions derived by PMP, which results in the following two-point boundary value problem (TPBVP) as

$$\begin{aligned} &\dot{x}(t) = A(x(t)) + B(t)u(t), \\ &\dot{\lambda}(t) = -Q(t)x(t) - A_x^T(x(t))\lambda(t), \\ &x(t_0) = x_0, x(t_f) = x_f, \end{aligned} \tag{3}$$

where $\lambda(t)$ is an n-vector of co-state functions, $A_x^T(x(t)) = \nabla_x A^T(x(t))$, and the optimal control is given by

$$u(t) = -R^{-1}B^T(t)\lambda(t). \tag{4}$$

In 2012, Berkani, Manseur, and Maidi [4] used the variational iteration method (VIM) to solve this problem, for which the control variable can be calculated in terms of state variables. They then turned the problem into a variational problem and solved it with VIM. It is noteworthy that the VIM is an analytical-approximate method for solving initial value problems (IVPs), which was introduced by He [8] in 1999, and many researchers developed the method to increase its efficiency and validity in recent years [20, 5]. For solving (3), the author and Effati [24] used a combination of the shooting method with VIM. However, the time-consuming and complex integration of nonlinear functions reduces its efficiency, at each iteration.

Matinfar and Saeidy [17] applied a modified VIM (MVIM) that uses He polynomials. Simultaneously, Saberi Nik, Effati, and Yildirim [22] used a differential transformation method along with differential transformation polynomials to solve (3). A comparison between the existing methods,

MVIM, Homotopy perturbation method (HPM), and Adomian decomposition method was done in 2016 by Jafari. Ghasempour, and Baleanu [9].

Mirhosseini-Alizamini and Effati [19] extended the VIM for optimal delay control problems. In 2020, the author and Effati [25] used the successive approximation method (SAM) to solve (3), which has a very good computational speed but does not guarantee the convergence for problems with large domains.

An improvement of VIM, called Interpolated VIM (IVIM) introduced by Salkuyeh and Tavakoli [23], interpolates the VIM with the well-known spline functions, which make the method suitable for approximating all types of $L^2[t_0, t_f]$ solutions, especially for control and state functions. This property motivates us to apply the IVIM for solving optimal control problems. This method has low computational time in comparison with VIM since it uses an approximation for integration. It also maintains accuracy and convergence.

This paper is organized as follows. In Section 2, the basic ideas of IVIM are briefly described. Then, the application of IVIM for solving nonlinear optimal control problems (NOCPs) is explained in Section 3. Several numerical examples are illustrated in Section 4, which show the accuracy and efficiency of the proposed method compared to the existing ones.

2 Basic ideas of the interpolated variational iteration method

As stated earlier, the VIM is an analytical-approximate iterative method for solving IVPs [8]. To explain this method, consider the following one-dimensional IVP

$$\begin{aligned} &\dot{y}(t) = f(t, y(t)), \quad t \in [0, T], \\ &y(0) = y^0, \end{aligned} \tag{5}$$

where $y(\cdot) \in \mathbb{R}$ and f satisfies the Lipschitz condition:

$$|f(t, y(t)) - f(t, \tilde{y}(t))| \leq L|y(t) - \tilde{y}(t)|, \tag{6}$$

for any $t \in [0, T]$ and for some Lipschitz constant $L > 0$. We make the following iterative formula, which is known as the VIM main formula:

$$y_{m+1}(t) = y_m(t) + \int_0^t \Lambda(s, t) \left(\dot{y}_m(s) - f(s, y_m(s))\right) ds, \quad m = 0, 1, 2, \ldots, \tag{7}$$

where $\Lambda(s, t)$ is known as a Lagrange multiplier [2, 28], and $y_0(t)$ should be chosen such that it satisfies $y_0(0) = y^0$. The VIM formula produces a sequence of functions $\{y_m(t)\}$, which converges to the exact solution of (5), under suitable conditions and does not require any linearization or perturbation. Due to the attractiveness of the method for calculating accurate solutions, many researchers have used it to solve difficult problems such as

dispersive optimal solitons [7], coupled Burgers' equation [26], fractional option pricing [18], fractal Fisher's equation [27]. However, the integration of nonlinear function $f(t, y(t))$ in all these researches may greatly reduce the speed of the method or lead to the failure of the method. To overcome this problem, first, we apply the integration by parts for (7) to get:

$$y_{m+1}(t) = G_m(t) - \int_0^t H_m(s,t)ds, \tag{8}$$

where

$$G_m(t) = (1 + \Lambda(t,t))y_m(t) - \Lambda(0,t)y_m(0),$$

and

$$H_m(s,t) = \frac{\partial \Lambda}{\partial s}(s,t)y_m(s) + \Lambda(s,t)f(s, y_m(s)).$$

Now, we interpolate $y_m(t)$ by first order B-spline functions [12], at equidistance points t_i, $i = 1, \ldots, n$ on $[0, T]$. For any $i = 2, 3, \ldots, n-1$, let us define

$$\mathcal{B}_i(t) = \frac{t - t_{i-1}}{h}\chi_{[t_{i-1}, t_i)}(t) + \frac{t_{i+1} - t}{h}\chi_{[t_i, t_{i+1})}(t),$$

and

$$\mathcal{B}_n(t) = \frac{t - t_{n-1}}{h}\chi_{[t_{n-1}, t_n)}(t),$$

where $\chi_{\mathcal{A}}$ is the characteristic function defined on a set $\mathcal{A}$. Its value is 1 on $\mathcal{A}$ and is 0 otherwise. It is obvious from the definition that $\mathcal{B}_i(t_i) = 1$ and $\mathcal{B}_i(t_j) = 0$, for $j \neq i$. Then, we employ the interpolant

$$\hat{y}_m(t) = \sum_{i=2}^{n} y_m(t_i)\mathcal{B}_i(t).$$

Thus, $\hat{y}_m(t_i) = y_m(t_i)$, for $i = 2, \ldots, n$. Now, we substitute the interpolant $\hat{y}_m$ into (7) and after some simplifications, we arrive at

$$\begin{aligned} y_{m+1}(t_i) &= \hat{y}_{m+1}(t_i) \\ &= G_m(t_i) - h\sum_{r=2}^{i-1} H_m(t_r, t_i) - \frac{h}{2}\left(H_m(t_1, t_i) + H_m(t_i, t_i)\right). \end{aligned} \tag{9}$$

This is the main formula of IVIM, which overcomes the nonlinear integration of VIM and will be used as a tool for solving NOCPs in the next section.

Now, we state the convergence of VIM and IVIM sequences as the next two theorems, which were proved in [23].

Theorem 1. Let f be a continuous function on $[0, T]$. If $y_m(t) \in \mathbb{C}^1([0, T])$ for $m = 0, 1, 2, \ldots$, then the VIM sequence (7) converges to the exact solution of (5).

Theorem 2. Let f be a two times continuously differentiable function. If the VIM sequence (7) is to the exact solution of (5), then so is the IVIM sequence (9).

3 The application of IVIM for solving NOCPs

For the sake of simplicity, we rewrite the TPBVP in (3) as

$$\begin{aligned} &\dot{x} = \mathcal{F}(t, x, \lambda), \\ &\dot{\lambda} = \mathcal{G}(t, x, \lambda), \\ &x(t_0) = x_0, \quad x(t_f) = x_f, \end{aligned} \tag{10}$$

where

$$\begin{aligned} &\mathcal{F}(t, x, \lambda) = A(x(t)) - B(t)R^{-1}B^T(t)\lambda(t), \\ &\mathcal{G}(t, x, \lambda) = -Q(t)x(t) - A_x^T(x(t))\lambda(t). \end{aligned}$$

This nonlinear TPBVP has no analytical solution in general. Thus, researchers are looking for reliable and almost accurate solutions to it. The technique we are studying here is IVIM, which was described in Section 2. Given that the above problem is TPBVP, we first convert it to an IVP as follows [24]:

$$\begin{aligned} &\dot{x} = \mathcal{F}(t, x, \lambda), \\ &\dot{\lambda} = \mathcal{G}(t, x, \lambda), \qquad t \in [t_0, t_f] \\ &x(t_0) = x_0, \quad \lambda(t_0) = \alpha, \end{aligned} \tag{11}$$

where α is an n-vector unknown variable. This variable could be identified after sufficient iterations of IVIM.

As Section 2, we create the VIM correction functional as follows:

$$x_{m+1}(t) = x_m(t) + \int_{t_0}^{t} \Lambda_1(t, s) \left\{ \dot{x}_m(s) - \mathcal{F}(s, x_m(s), \lambda_m(s)) \right\} ds, \tag{12}$$

$$\lambda_{m+1}(t) = \lambda_m(t) + \int_{t_0}^{t} \Lambda_2(t, s) \left\{ \dot{\lambda}_m(s) - \mathcal{G}(s, x_m(s), \lambda_m(s)) \right\} ds, \tag{13}$$

with $x_0(t) = x_0$, $\lambda_0(t) = \alpha$ and $m \geq 0$. To determine general multipliers $\Lambda_1(t, s)$ and $\Lambda_2(t, s)$, we restrict the variations for $\mathcal{F}$ and $\mathcal{G}$, that is, $\delta\widetilde{\mathcal{F}} = \delta\widetilde{\mathcal{G}} = 0$. Therefore, the optimal multipliers become $\Lambda_1(t, s) = \Lambda_2(t, s) = -1$. Compared with (8), consider $G_1(t)$ and $H_1(t)$ corresponding to $x(t)$ and

$$\begin{aligned} &G_{1,m}(t) = (1 + \Lambda_1(t, t))x_m(t) - \Lambda_1(0, t)x_0, \\ &H_{1,m}(s, t) = \frac{\partial \Lambda_1}{\partial s}(s, t)x_m(s) + \Lambda_1(s, t)\mathcal{F}(s, x_m(s), \lambda_m(s)). \end{aligned}$$

Similarly, consider $G_2(t)$ and $H_2(t)$ corresponding to $\lambda(t)$ and

$$G_{2,m}(t) = (1 + \Lambda_2(t,t))x_m(t) - \Lambda_2(0,t)\alpha,$$
$$H_{2,m}(s,t) = \frac{\partial \Lambda_2}{\partial s}(s,t)x_m(s) + \Lambda_2(s,t)\mathcal{G}(s, x_m(s), \lambda_m(s)).$$

For a given $n \in \mathbb{N}$, assume $t_i = t_0 + (i-1)h$, $i = 2, 3, \ldots$, where $h = \frac{t_f - t_0}{n-1}$. Then, the B-spline interpolants of $x_{m+1}(t)$ and $\lambda_{m+1}(t)$ become

$$x_{m+1}^{(h)}(t) = \sum_{i=2}^{n} \hat{x}_{m+1}(t_i)\mathcal{B}_i(t),$$
$$\lambda_{m+1}^{(h)}(t) = \sum_{i=2}^{n} \hat{\lambda}_{m+1}(t_i)\mathcal{B}_i(t),$$

where the coefficients are calculated as

$$\begin{aligned} x_{m+1}(t_i) &= \hat{x}_{m+1}(t_i) \\ &= G_{1,m}(t_i) - h\sum_{r=2}^{i-1} H_{1,m}(t_r, t_i) - \frac{h}{2}\left(H_{1,m}(t_1, t_i) + H_{1,m}(t_i, t_i)\right), \end{aligned} \tag{14}$$

$$\begin{aligned} \lambda_{m+1}(t_i) &= \hat{\lambda}_{m+1}(t_i) \\ &= G_{2,m}(t_i) - h\sum_{r=2}^{i-1} H_{2,m}(t_r, t_i) - \frac{h}{2}\left(H_{2,m}(t_1, t_i) + H_{2,m}(t_i, t_i)\right), \end{aligned} \tag{15}$$

for any $i = 2, 3, \ldots, n$. We call these two equations as the IVIM main formulas. The convergence of the IVIM is a straightforward conclusion of Theorems 1–2, which we state in the next theorem.

Theorem 3. Let $\mathcal{F}$ and $\mathcal{G}$ be two times continuously differentiable functions with respect to their first variable and let they satisfy the Lipschitz condition. If $x_m(t), \lambda_m(t) \in \mathbb{C}^1([t_0, t_f])$, then the IVIM of (14)–(15) is convergent to the exact solutions of (11), say $x^*(t,\alpha)$ and $\lambda^*(t,\alpha)$. Accordingly, $x^*(t,\alpha^*)$, $\lambda^*(t,\alpha^*)$ are the exact solution of (10) when α^* is the real root of $x^*(t_f,\alpha) = x_f$.

Proof. Since $\mathcal{F}$ and $\mathcal{G}$ are continuous on $[t_0, t_f]$, and $x_m(t), \lambda_m(t) \in \mathbb{C}^1([t_0, t_f])$, Theorem 3 guarantees the convergence of the VIM formulas (12)–(13) to the exact solution of (11). Besides, since $\mathcal{F}$ and $\mathcal{G}$ are two times continuously differentiable functions, Theorem 4 ensures that the IVIM is convergent to the exact solution of (11), say $x^*(t,\alpha)$ and $\lambda^*(t,\alpha)$. Clearly, for every α, $x^*(t_0,\alpha) = x_0$ and $\lambda^*(t_0,\alpha) = \alpha$. Then, from (10), we have $x(t_f) = x_f = x^*(t_f,\alpha)$. Therefore, to get the exact solution of (10), one should solve $x^*(t_f,\alpha) = x_f$ for some appropriate real α. □

Theorem 4. Under the assumptions of Theorem 3, the sequences $\{u_m(t)\}$ and $\{J_m\}$ defined by

$$u_m(t) = -R^{-1}B^T(t)\lambda_m(t), \tag{16}$$

$$J_m = \frac{h}{2}\left(\mathcal{L}(x_0, u_m(t_1)) + \mathcal{L}(x_m(t_f), 0)\right) + h\sum_{i=2}^{n-1}\mathcal{L}(x_m(t_i), u_m(t_i)), \tag{17}$$

converge to the optimal control and optimal objective value, respectively, where $\mathcal{L}(x,u) = \frac{1}{2}(x^TQx + u^TRu)$.

Proof. Taking limit from (16) and putting $\alpha = \alpha^*$ yield

$$\begin{aligned} u^*(t) := \lim_{m\to\infty} u_m(t) &= -R^{-1}B^T(t)\lim_{m\to\infty}\lambda_m(t) \\ &= -R^{-1}B^T(t)\lambda^*(t,\alpha^*), \end{aligned}$$

where $u^*(t)$ is the optimal control function, since $x^*(t,\alpha^*)$ and $\lambda^*(t,\alpha^*)$ are the exact solutions of extreme conditions in (10).

Now, taking limit from (17), as $n, m \to \infty$, and using the trapezoidal integration rule result in

$$\begin{aligned} &\lim_{m\to\infty}\lim_{n\to\infty} J_{m,n} \\ &= \lim_{m\to\infty}\lim_{n\to\infty}\frac{h}{2}\left(\mathcal{L}(x_0, u_m(t_1)) + \mathcal{L}(x_m(t_f), 0)\right) + h\sum_{i=2}^{n-1}\mathcal{L}(x_m(t_i), u_m(t_i)) \\ &= \lim_{m\to\infty}\int_{t_0}^{t_f}\mathcal{L}(x_m(t), u_m(t))dt \\ &= \lim_{m\to\infty}\frac{1}{2}\int_{t_0}^{t_f}\left(x_m(t)Q(t)x_m(t) + u_m(t)^TR(t)u_m(t)\right)dt \\ &= \frac{1}{2}\int_{t_0}^{t_f}\left(\lim_{m\to\infty}x_m(t)Q(t)\lim_{m\to\infty}x_m(t) + \lim_{m\to\infty}u_m^T(t)R(t)\lim_{m\to\infty}u_m(t)\right)dt \\ &= \frac{1}{2}\int_{t_0}^{t_f}\left(x^*(t,\alpha^*)Q(t)x^*(t,\alpha^*) + u^*(t)^TRu^*(t)\right)dt \\ &= J^*. \end{aligned}$$

It follows that J^* is the optimal objective value and the proof is completed. □

Now, we present an algorithm to summarize our proposed method. We choose the stopping criterion

$$\bar{E}_m = \left|\frac{J_m - J_{m-1}}{J_m}\right| < \epsilon, \tag{18}$$

which maintains the accuracy of the objective value for any sufficiently small tolerance limit $\epsilon > 0$.

Algorithm of IVIM for solving nonlinear OCPs

Step 1. Let $m := 1$, $x_0(t) := x_0$, $\lambda_0(t) := \alpha$ and $\epsilon > 0$.
Step 2. Calculate state and costate functions, $x_{m+1}(t, \alpha)$ and $\lambda_{m+1}(t, \alpha)$, through the IVIM formulas (14)–(15).
Step 3. Solve $x_{m+1}(t_f, \alpha) = x_f$ for α, to get α^*.
Step 4. Find the suboptimal control $u_{m+1}(t, \alpha^*)$ and the suboptimal objective value J_{m+1}, by (16)–(17).
Step 5. If $\bar{E}_{m+1} < \epsilon$, then goto **Step 6**; otherwise let $m := m + 1$ and goto **Step 2**.
Step 6. Stop the Algorithm. $u_{m+1}(t)$ and J_{m+1} are the desired optimal control and objective value.

Remark 1. If the state function has free final condition, then the costate final condition $\lambda(t_f) = 0$ must be applied. Therefore, to calculate α in **Step 3**, the equation $\lambda_{m+1}(t_f, \alpha) = 0$ must be used instead of $x_{m+1}(t_f, \alpha) = x_f$.

4 Illustrative examples

In this section, we give two nonlinear examples to illustrate the reliability and efficiency of our proposed method. In the first example, we describe the algorithm process in detail. Also, we present the second the third example to compare the method with some other existing methods. Since there are no exact solutions for these three examples, we use criterion (18) to stop the algorithm and check the accuracy of the method. Besides, we use the Maple "dsolve" package as a reliable solution for comparison. The codes were developed using symbolic software Maple 18 on a machine with Intel Core 2 Due Processor 2.53 GHz and 4 GB RAM.

Example 1. Consider the following nonlinear control system [24, 25]:

$$\dot{x} = \frac{1}{2}x^2(t)\sin x(t) + u(t), \qquad t \in [0, 1],$$
$$x(0) = 0, \quad x(1) = 0.5,$$

under which the following objective functional should be minimized:

$$J = \int_0^1 u^2(t)dt.$$

Conforming to (3), we have

$$\dot{x}(t) = \frac{1}{2}x^2(t)\sin x(t) - \frac{1}{2}\lambda(t), \tag{19}$$

$$\dot{\lambda}(t) = -\lambda(t)x(t)\sin x(t) - \frac{1}{2}\lambda(t)x^2(t)\cos x(t), \quad t \in [0,1] \tag{20}$$

$$x(0) = 0, \quad x(1) = 0.5, \tag{21}$$

and the control function is $u(t) = -\frac{1}{2}\lambda(t)$. We substitute the final state condition $x(1) = 0.5$, by $\lambda(0) = \alpha$ to get

$$\begin{aligned}
&\dot{x}(t) = \frac{1}{2}x^2(t)\sin x(t) - \frac{1}{2}\lambda(t),\\
&\dot{\lambda}(t) = -\lambda(t)x(t)\sin x(t) - \frac{1}{2}\lambda(t)x^2(t)\cos x(t), \qquad t \in [0,1],\\
&x(0) = 0, \qquad \lambda(0) = \alpha,
\end{aligned}$$

where α is an unknown real variable. In view of (11), $\mathcal{F}$ and $\mathcal{G}$ are the right-hand sides of (19)–(20), respectively. They are analytic functions with respect to x and λ, and they satisfy the Lipschitz condition (6), with $L_{\mathcal{F}} = 9/16$ and $L_{\mathcal{G}} = 5/16$, whenever $|x| \le 0.5$ and $\lambda \le 1$. Hence, Theorem 3 guarantees the convergence of both VIM and IVIM formulas. The VIM formulas (12)–(13) become

$$\begin{aligned}
x_{m+1}(t) =& x_m(t) - \int_0^t \{\dot{x}_m(s) - \frac{1}{2}x_m^2(s)\sin x_m(s) + \frac{1}{2}\lambda_m(s)\}ds,\\
\lambda_{m+1}(t) =& \lambda_m(t) - \int_0^t \{\dot{\lambda}_m(s) + \lambda_m(s)x_m(s)\sin x_m(s)\\
&+ \frac{1}{2}\lambda_m(s)x_m^2(s)\cos x_m(s)\}ds,\\
x_0(t) =& 0, \qquad \lambda_0(t) = \alpha, \quad m \ge 0.
\end{aligned}$$

Now, interpolating VIM at $t_j = 0, \frac{1}{2}, 1$, the IVIM formulas (14)–(15) give the following results, after the first iteration:

$$\begin{aligned}
&x_1(t_j) = 0, \quad -\frac{1}{4}\alpha - \frac{1}{128}\alpha^2\sin(\frac{1}{4}\alpha), \quad -\frac{1}{64}\alpha^2\sin(\frac{1}{4}\alpha) - \frac{1}{2}\alpha - \frac{1}{32}\alpha^2\sin(\frac{1}{2}\alpha),\\
&\lambda_1(t_j) = \alpha, \quad \alpha - \frac{1}{16}\alpha^2\sin(\frac{1}{4}\alpha) - \frac{1}{128}\alpha^3\cos(\frac{1}{4}\alpha),\\
&\alpha - \frac{1}{8}\alpha^2\sin(\frac{1}{4}\alpha) - \frac{1}{64}\alpha^3\cos(\frac{1}{4}\alpha) - \frac{1}{8}\alpha^2\sin(\frac{1}{2}\alpha) - \frac{1}{32}\alpha^3\cos(\frac{1}{2}\alpha).
\end{aligned}$$

The final state condition, $x_1(1) = 0.5$, results in

$$-\frac{1}{64}\alpha^2\sin(\frac{1}{4}\alpha) - \frac{1}{2}\alpha - \frac{1}{32}\alpha^2\sin(\frac{1}{2}\alpha) = 0.5, \tag{22}$$

Table 1: Calculation of α for $n = 3$ and $m = 1$, Example 1.

α	Objective functional
-0.9659445089	0.2146370242
-1.023796574	0.2387946125

which obtains two real roots. These real roots along with their corresponding objective values are summarized in Table 1. We suggest choosing the parameter α, which gives a better objective value, that is, $\alpha = -0.9659445089$. This parameter could be a good starting point for root-finding iterative methods, such as Newton, at the next iteration of the IVIM.
Table 2 shows the results of IVIM procedure up to $m = 5$ iterations with $n = 10$ nodes.

Table 2: Simulation results for $n = 10$, Example 1.

m	J_m	$\bar{E}_m$	CPU time	α
1	0.22264786	-	0.000	-0.9717459824
2	0.23465678	5.1177E-02	0.016	-0.9985940802
3	0.23541609	3.2254E-03	0.078	-0.9997821239
4	0.23513137	1.2109E-03	0.359	-0.9991493164
5	0.23512939	8.4519E-06	2.266	-0.9991493164

To compare the IVIM with two recent methods, VIM [24] and Modal series [10], we summarize their results in Table 3 after five iterations with the same stopping criterion (18), with $\epsilon = 2 \times 10^{-5}$. Clearly, the IVIM converges rapidly, in contrast with VIM and Modal series, since it reaches a better tolerance limit. Also, the suboptimal control and state of the proposed VIM and Modal series methods are demonstrated in Figure 1.

Table 3: Comparing results of IVIM, VIM, Modal series, Example 1.

Method	J_m	$\bar{E}_m$
IVIM (m=5)	0.2351	8.5E-06
VIM (m=5)	0.2353	1.7E-03
Modal series method (m=5)	0.2346	1.3E-03

Example 2. Consider the nonlinear control system described by

$$\begin{aligned} &\dot{x}_1 = x_2 + x_1 x_2, \\ &\dot{x}_2 = -x_1 + x_2 + x_2^2 + u, \\ &x_1(0) = -0.8, \quad x_2(0) = 0, \end{aligned} \tag{23}$$

and the objective functional

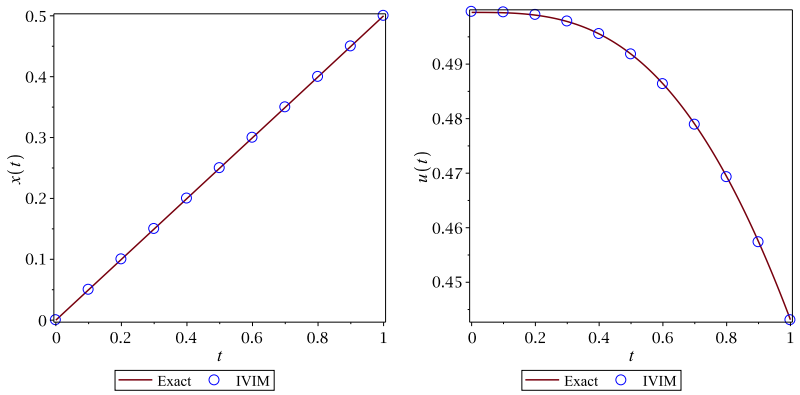


Figure 1: Suboptimal state and control functions, Example 1.

$$J = \frac{1}{2}\int_0^1 (x_1^2 + x_2^2 + u^2)dt. \tag{24}$$

One can obtain the optimality conditions as

$$\begin{aligned}
&\dot{\lambda}_1 = -(x_1 + \lambda_1 x_2 - \lambda_2),\\
&\dot{\lambda}_2 = -(x_2 + \lambda_1(1 + x_1) + \lambda_2(1 + 2x_2),\\
&\dot{x}_1 = x_2 + x_1 x_2,\\
&\dot{x}_2 = -x_1 + x_2 + x_2^2 - \lambda_2,\\
&x_1(0) = -0.8, \quad x_2(0) = 0,
\end{aligned}$$

and the optimal control law as $u = -\lambda_2$. Since $x_1(t)$ and $x_2(t)$ are free at $t = 1$, we use the Remark 1 and put $\lambda_1(1) = \lambda_2(1) = 0$.

Now, due to the polynomial nature of the system and their validity in the assumptions of Theorem 3, we apply the proposed algorithm for $\epsilon = 10^{-3}$ to get an accurate optimal objective value. It is worth noting that for some iterations, $m = 2$, solving $\lambda_{1,m}(1, \alpha) = 0$ and $\lambda_{2,m}(1, \alpha) = 0$, results in different values of α. In these cases, as shown in Table 4, we clearly select the vector α that yields the minimum objective functional, that is, $\alpha_1 = -1.227402971$ and $\alpha_2 = 0.5053886518$.

Table 4: Calculation of α for $m = 2$, Example 2.

α_1	α_2	Objective functional
-1.227402971	0.5053886518	0.4113400952
4.538547207	-1.207206466	2.664175385
1.875104138	-1.907110672	4.484594474

Table 5: Simulation results for different iterations, Example 2.

m		VIM			SAM			IVIM	
(Itr.)	J_m	E_m	Time	J_m	E_m	Time	J_m	E_m	Time
1	0.8533	-	0.031	0.5357	-	0.015	0.8560	-	0
2	0.4108	1.1E-0	0.031	0.7575	2.9E-1	0.031	0.4113	1.1E-0	0.015
3	0.4325	5.0E-2	0.046	0.9930	2.4E-1	0.063	0.4340	5.2E-2	0.078
4	0.4395	1.6E-2	0.125	0.6750	4.7E-1	0.046	0.4413	1.7E-2	0.124
5	0.4422	6.1E-3	1.544	0.3503	9.3E-1	0.094	0.4443	6.9E-3	0.296
6	0.4430	1.8E-3	45.693	0.4245	1.7E-1	0.078	0.4453	2.3E-3	1.279
7	-	-	>1800	0.6025	3.0E-1	0.078	0.4456	6.1E-4	7.504

Then, one can use these α_1 and α_2 as initial guesses for solving $\lambda_{1,3}(1,\alpha) = 0$ and $\lambda_{2,3}(1,\alpha) = 0$. This technique greatly improves system speed and reduces computational complexity.

The objective values, relative errors and the CPU time of the algorithm iterations are summarized in Table 5. Moreover, to clarify the reliability of the algorithm, we compare the results with VIM [24] and SAM [25]. As shown in Table 5, SAM has a divergent behavior, and although it has less computational time than the other two methods, it does not reach the desired tolerance limit. Also, the VIM requires more than 1800 seconds of computational time for $m = 7$ and does not reach the desired tolerance. This is while the IVIM reaches the desired tolerance limit after $m = 7$ iterations, and it only takes 7.504 seconds of CPU time.

From the beginning of the algorithm, if we considered $\epsilon = 5 \times 10^{-3}$, both IVIM and VIM would get the desired solutions after $m = 6$ iterations, but IVIM in 1.279 seconds and VIM in 45.693 seconds of CPU time, which clearly shows the speed of convergence of IVIM.

Figure 2 shows the solutions after $m = 7$ algorithm iterations, in comparison with "dsolve" package of Maple as the exact solutions.

Example 3. Consider the optimal maneuvers of a rigid asymmetric spacecraft given by [21]

$$\begin{aligned}
\dot{x}_1(t) &= -\frac{I_3 - I_2}{I_1}\, x_2(t)x_3(t) + \frac{1}{I_1}u_1(t),\\
\dot{x}_2(t) &= -\frac{I_1 - I_3}{I_2}\, x_1(t)x_3(t) + \frac{1}{I_2}u_2(t),\\
\dot{x}_3(t) &= -\frac{I_2 - I_1}{I_3}\, x_1(t)x_2(t) + \frac{1}{I_3}u_3(t),
\end{aligned} \tag{25}$$

with the boundary conditions

$$\begin{aligned}
&x_1(0) = 0.01\, r/s, \quad x_2(0) = 0.005\, r/s, \quad x_3(0) = 0.001\, r/s,\\
&x_1(100) = x_2(100) = x_3(100) = 0\, r/s,
\end{aligned} \tag{26}$$

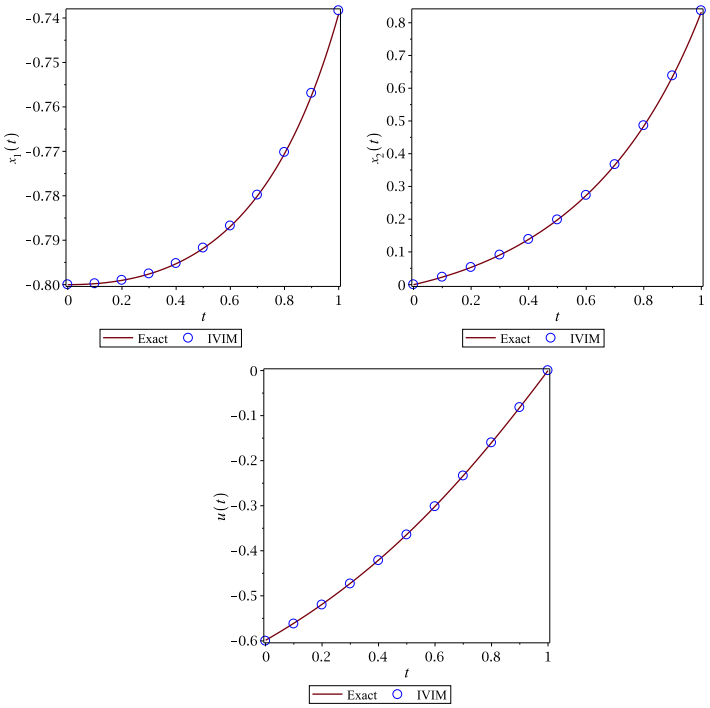


Figure 2: The optimal trajectories and control of Example 2.

where x_1, x_2, and x_3 are angular velocities of the spacecraft, u_1, u_2, and u_3 are control torques, $I_1 = 86.24$, $I_2 = 85.07$, and $I_3 = 113.59$ kg m^2 are the spacecraft principle inertia. The controls $u_i(t), i = 1, 2, 3$, should be chosen to minimize the following quadratic performance index:

$$J[x, u] = \frac{1}{2}\int_0^{100} (u_1^2(t) + u_2^2(t) + u_3^2(t))dt, \tag{27}$$

In view of (1)–(2), we have:

$$A(x(t)) = \begin{bmatrix} -\frac{I_3 - I_2}{I_1}\, x_2(t)x_3(t) \\ -\frac{I_1 - I_3}{I_2}\, x_1(t)x_3(t) \\ -\frac{I_2 - I_1}{I_3}\, x_1(t)x_2(t) \end{bmatrix}.$$

Also, $B(t) = diag(\frac{1}{I_1}, \frac{1}{I_2}, \frac{1}{I_3})$, $Q(t) = O_{3\times 3}$, and $R(t) = I_{3\times 3}$. Then, the optimality conditions become

$$\dot{x}_1(t) = -\frac{I_3 - I_2}{I_1}\, x_2(t)x_3(t) - \frac{\lambda_1(t)}{I_1^2},$$

$$\dot{x}_2(t) = -\frac{I_1 - I_3}{I_2}\, x_1(t)x_3(t) - \frac{\lambda_2(t)}{I_2^2},$$
$$\dot{x}_3(t) = -\frac{I_2 - I_1}{I_3}\, x_1(t)x_2(t) - \frac{\lambda_3(t)}{I_3^2},$$
$$\dot{\lambda}_1(t) = \frac{I_1 - I_3}{I_2}\, x_3(t)\lambda_2(t) + \frac{I_2 - I_1}{I_3}\, x_2(t)\lambda_3(t),$$
$$\dot{\lambda}_2(t) = \frac{I_3 - I_2}{I_1}\, x_3(t)\lambda_1(t) + \frac{I_2 - I_1}{I_3}\, x_1(t)\lambda_3(t),$$
$$\dot{\lambda}_3(t) = \frac{I_3 - I_2}{I_1}\, x_2(t)\lambda_1(t) + \frac{I_1 - I_3}{I_2}\, x_1(t)\lambda_2(t),$$

with the boundary conditions (26), and the optimal control laws are given by

$$u_i^*(t) = -\frac{\lambda_i(t)}{I_i}, \quad i = 1, 2, 3, \quad t \in [0, 100]. \tag{28}$$

By using the proposed algorithm with $\epsilon = 10^{-6}$, we reach the desired solution after $m = 6$ iterations.

Table 6: Simulation results for $n = 10$, Example 3.

m	J_m	$\bar{E}_m$	α_1	α_2	α_3
1	0.00468914	-	0.7314359	0.3851119	0.1356719
2	0.00468570	7.3285E-04	0.7453951	0.3581385	0.1272553
3	0.00468711	2.9978E-04	0.7437911	0.3617697	0.1281838
4	0.00468791	1.7023E-04	0.7437022	0.3619104	0.1291345
5	0.00468782	1.9383E-05	0.7437007	0.3619099	0.1290549
6	0.00468781	7.9781E-07	0.7437039	0.3619034	0.1290513

The simulation results along with the estimation of unknown parameters $\alpha_i, i = 1, 2, 3$, are given in Table 6.

Table 7: The maximum error of $x_1(t)$ in comparison to HPM [21], Example 3.

m		HPM			IVIM	
(Itr.)	Max Error		CPU time	Max Error		CPU time
2	7.3102E-05		7.328	9.7727E-06		0.031
3	9.2612E-06		12.542	5.7522E-07		0.109
4	1.9301E-06		18.664	4.4215E-08		0.188
5	2.2560E-07		36.729	3.7427E-08		0.655
6	3.1420E-08		46.401	1.6156E-08		3.994

To compare the solutions with HPM [21], we have summarized the maximum error of $x_1(t)$ in both methods compared to "dsolve" package of Maple as the exact solution, In Table 7. As is evident, the IVIM is faster and

more accurate than HPM, so as to reach the error of 5×10^{-8}, HPM needs $m = 6$ iterations with 46.401 seconds of calculation time, but IVIM, needs 4 iterations with 0.188 seconds of calculation time.

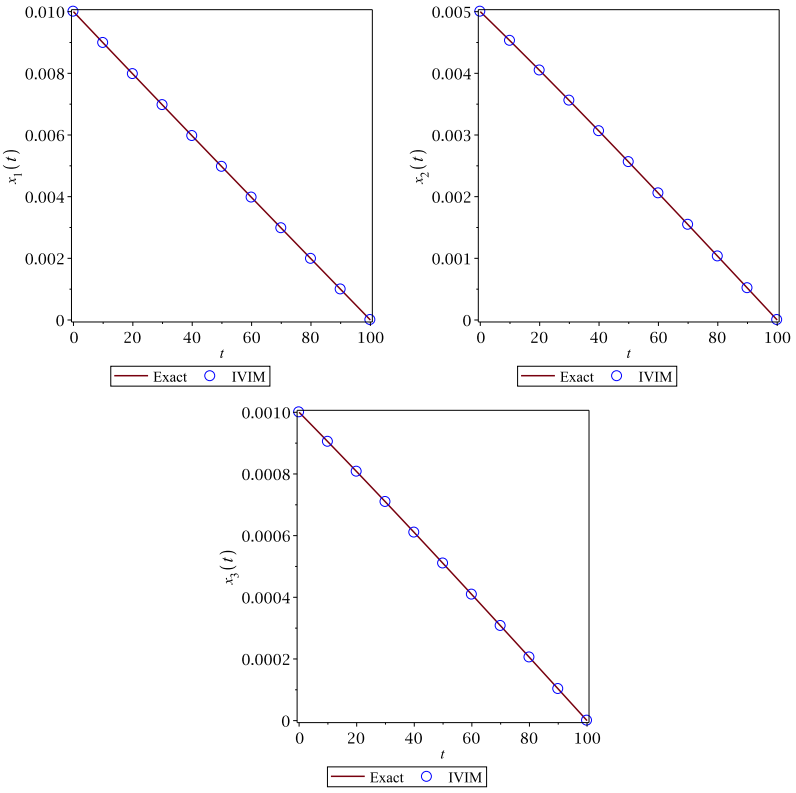


Figure 3: The optimal states $x_1(t)$,$x_2(t)$, and $x_3(t)$, Example 3.

Figures 3–4 show the optimal states and controls of IVIM for $m = 6$ iterations compared to "dsolve" solutions. The maximum errors of our proposed method are 1.6156×10^{-8}, 1.0231×10^{-8}, and 1.39963×10^{-8}, for state functions x_1, x_2, and x_3, respectively, and 3.4660×10^{-7}, 6.8377×10^{-7}, and 2.1526×10^{-7}, for control functions u_1, u_2, and u_3, respectively. Meanwhile, only 3.994 seconds are needed to perform this operation.

5 Conclusions

In this paper, the interpolated VIM was used to solve NOCPs. This method used basic spline functions as interpolants, which is desirable for the approximate calculation of state and control functions. This method did not require

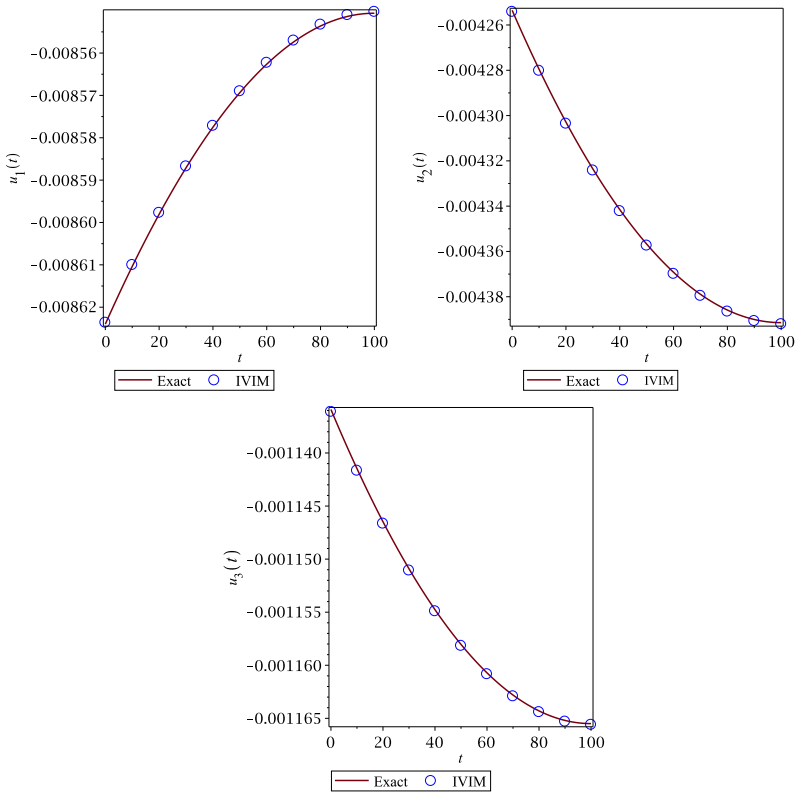


Figure 4: The optimal controls $u_1(t)$,$u_2(t)$, and $u_3(t)$, Example 3.

complex analytical integrations and therefore was faster than VIM. Furthermore, the convergence range of the IVIM was wider, compared to the new SAM. In each iteration of the method, the co-state initial value was quickly calculated with an efficient technique described in the Examples section. Finally, the efficiency and reliability of the method were discussed, and the "dsolve" package of Maple 18 was used as the exact solution to check the validity of the method.

Acknowledgments

This research is supported by a grant from University of Neyshabur (Grant No. s/96/5173).

References

[1] Ali, A., Iqbal, Q., Asamoah, J. K. K., and Islam, S. *Mathematical modeling for the transmission potential of Zika virus with optimal control strategies.* Eur. Phys. J. Plus 137 (1) (2022), 1–30.

[2] Anjum, N., and He, J.-H.*Laplace transform: making the variational iteration method easier.* Appl. Math. Lett. 92 (2019), 134–138.

[3] Ashchepkov, L., Dolgy, D., Kim, T., and Agarwal, R. Optimal Control. *Springer International Publishing*, 2022.

[4] Berkani, S., Manseur, F., and Maidi, A. *Optimal control based on the variational iteration method.* Comput. Math. Appl. 64 (4) (2012), 604–610.

[5] El-Sayed, T. A., and El-Mongy, H. H. *A new numeric–symbolic procedure for variational iteration method with application to the free vibration of generalized multi-span Timoshenko beam.* J. Vib. Control 28 (7-8) (2022), 799–811.

[6] Ermakov, A., Marie, A., and Ringwood, J. V. *Optimal control of pitch and rotational velocity for a cyclorotor wave energy device.* IEEE Trans. Sustain. Energy 13 (3) (2022), 1631–1640.

[7] Gonzalez-Gaxiola, O., Biswas, A., Ekici, M., and Khan, S. *Highly dispersive optical solitons with quadratic–cubic law of refractive index by the variational iteration method.* J. Opt. 51 (1) (2022), 29–36.

[8] He, J.-H. *Variational iteration method–a kind of non-linear analytical technique: some examples.* Int. J. Non-Linear Mech. 34 (4) (1999), 699–708.

[9] Jafari, H., Ghasempour, S., and Baleanu, D. *On comparison between iterative methods for solving nonlinear optimal control problems.* J. Vib. Control 22 (9) (2016), 2281–2287.

[10] Jajarmi, A., Pariz, N., Kamyad, A. V., and Effati, S. *A novel modal series representation approach to solve a class of nonlinear optimal control problems.* Int. J. Innov. Comput. Inf. Control. 1 (2011), 2.

[11] Kim, Y., and Singh, T. *Energy-time optimal control of wheeled mobile robots.* J. Frankl. Inst. (2022).

[12] Kincaid, D., Kincaid, D. R., and Cheney, E. W. *Numerical analysis: mathematics of scientific computing*, vol. 2. American Mathematical Soc., 2009.

[13] Liu, S., You, H., Liu, Y., Feng, W., and Fu, S. *Research on optimal control strategy of wind–solar hybrid system based on power prediction.* ISA trans.123 (2022), 179–187.

[14] Lohéac, J., Varma, V. S., and Morărescu, I. C. *Time optimal control for a mobile robot with a communication objective.* Math. Comput. Simul. 201 (2022), 96–120.

[15] Ma, Z., and Zou, S. *Optimal Control Theory: The Variational Method.* Springer Singapore, 2021.

[16] Majumder, M., Tiwari, P. K., and Pal, S. *Impact of saturated treatments on HIV-TB dual epidemic as a consequence of COVID-19: Optimal control with awareness and treatment.* Nonlinear Dyn. (2022), 1–34.

[17] Matinfar, M., and Saeidy, M. *A new analytical method for solving a class of nonlinear optimal control problems.* Optim. Control Appl. Methods 35 (3) (2014), 286–302.

[18] Meng, L., Kexin, M., Ruyi, X., Mei, S., and Cattani, C. *Haar wavelet transform and variational iteration method for fractional option pricing models.* Math. Methods Appl. Sci. (2022).

[19] Mirhosseini-Alizamini, S. M., and Effati, S. *An iterative method for sub-optimal control of a class of nonlinear time-delayed systems.* Int. J. Control 92 (12) (2019), 2869–2885.

[20] Nadeem, M., and He, J.-H. *He–Laplace variational iteration method for solving the nonlinear equations arising in chemical kinetics and population dynamics.* J. Math. Chem. 59 (5) (2021), 1234–1245.

[21] Saberi Nik, H., Effati, S., Motsa, S. S., and Shirazian, M. *Spectral homotopy analysis method and its convergence for solving a class of nonlinear optimal control problems.* Numer. Algorithms 65 (1) (2014), 171–194.

[22] Saberi Nik, H., Effati, S., and Yildirim, A. *Solution of linear optimal control systems by differential transform method.* Neural. Comput. Appl. 23 (5) (2013), 1311–1317.

[23] Salkuyeh, D. K., and Tavakoli, A. *Interpolated variational iteration method for initial value problems.* Appl. Math. Model. 40 (5-6) (2016), 3979–3990.

[24] Shirazian, M., and Effati, S. *Solving a class of nonlinear optimal control problems via he's variational iteration method.* Int. J. Control Autom. Syst 10 (2) (2012), 249–256.

[25] Shirazian, M., and Effati, S. *A novel successive approximation method for solving a class of optimal control problems.* Caspian Journal of Mathematical Sciences (CJMS) peer 9 (1) (2020), 124–136.

[26] Singh, G., and Singh, I. *Semi-analytical solutions of three-dimensional (3d) coupled Burgers' equations by new Laplace variational iteration method.* Partial Differential Equations in Applied Mathematics (2022), 100438.

[27] Sun, J.-S. *Approximate analytic solution of the fractal Fisher's equation via local fractional variational iteration method.* Therm. Sci. 26 (3 Part B) (2022), 2699–2705.

[28] Tao, Z.-L., Chen, G.-H., and Chen, Y.-H. *Variational iteration method with matrix Lagrange multiplier for nonlinear oscillators.* J. Low Freq. Noise Vib. Act. Control. 38 (3-4) (2019), 984–991.

How to cite this article
Shirazian, M., The interpolated variational iteration method for solving a class of nonlinear optimal control problems. *Iran. j. numer. anal. optim.*, 2023; 13(2): 224-242. https://doi.org/10.22067/ijnao.2022.77393.1158

Iranian Journal of Numerical Analysis and Optimization
Vol. 13, No. 2, 2023, pp 243–261
https://doi.org/10.22067/ijnao.2022.76652.1178
https://ijnao.um.ac.ir/

CC BY NC

How to cite this article
Research Article

A robust method for optimal control problems governed by system of Fredholm integral equations in mechanics

A. Ebrahimzadeh

Abstract

This essay considers an optimal control problem (**OCP**) governed by a system of Fredholm integral equations (**FIE**). In this paper, collocation approach with utilizing Lagrange polynomials is introduced to transform the OCP into a nonlinear programming problem (**NLP**). An efficient optimization method in Mathematica software is utilized to solve NLP. The convergence analysis is discussed, which show the theoretical structure behind the propounded technique under some assumptions. In this essay, computational outcomes are given to demonstrate the adaptability, forthrightness, and relationship of the calculations manufactured. A practical real-world problem involving hanging chain in classical mechanic is also dissolved utilizing the approach proposed.

AMS subject classifications (2020): 49M25, 90C30.

Keywords: Collocation method; Lagrange polynomials; Optimal control; System of Fredholm integral equation; Convergence analysis.

1 Introduction

The mathematical formulations of most physical and mechanical engineering models lead to FIE. FIEs occur in different real world models like the mass dispensation of polymers in polymeric melt, linear forward modeling, signal processes etc [7, 26]. The OCPs inclusive integral equations are a different class of OCPs that have a substantial role in population dynamics, economy, heat conduction, continuum mechanics, electrical power centers, water

Received 30 August 2022; revised 13 October 2022; accepted 31 October 2022

Asiyeh Ebrahimzadeh
Department of Educational Mathematics, Farhangian university, Tehran, Iran. e-mail: a.ebrahimzadeh@cfu.ac.ir

resource management, and mass transport [13, 19]. In particular, optimal control of systems managed by FIE is significant in usages like the OCP regarding the Ornstein-Uhlenbeck process which arises from statistical communication theory [18]. In recent years, many methods for solving OCPs of FIE, whether direct or indirect, have been proposed. One can read about this subject in references [2, 1, 8, 10, 16, 20, 24].

Problem$\mathcal{B}$: Determine the optimal control vector u^* and the corresponding optimal state vector x^* for a category of OCPs governed by a system of nonlinear FIEs described as follows:

$$\min \ J(x,u) = \int_0^T F(t, x_1(t), \cdots, x_n(t), u_1(t), \cdots, u_m(t))dt, \qquad (1)$$

subject to

$$x_j(t) = y_j(t) + \int_0^T K_j(t, s, x_1(s), \cdots, x_n(s), u_1(s), \cdots, u_m(s))ds, \quad j = 1, \cdots, n, \qquad (2)$$

where $x = [x_1, x_2, \cdots, x_n]^T \in \Re^n$, $u = [u_1, u_2, \cdots, u_m]^T \in \Re^m$, $F \in C([0,T], \Re^n, \Re^m)$ and $K_j \in C([0,T], [0,T], \Re^n, \Re^m)$.

Many researchers have been solved and carefully examined non-linear system of FIEs by utilizing orthogonal basis functions. For example, see the references [3, 4, 21, 22]. The orthogonal nature causes the solution of the preliminary problem to transmute into an algebraic system of equations. More ever, by utilizing an operational matrix of integral or derivative of these functions, the computational convolution of the problem is substantially diminished.

For dissolving the mentioned problem, the Lagrange polynomials and Gauss–-Legendre (**GL**) quadrature rule are used. A robust direct approach is established to transform the problem $\mathcal{B}$ into a NLP. This method is based on the expansion of control and state functions with Lagrange polynomials together with the GL integration method. We have utilized Wolfram Mathematica 12 for obtaining the solution of the NLP. Convergence analysis and related theorems are discussed in details. The reliability of the approach is investigated numerically by dissolving some illustrative examples. The OCP, which modeled the problem of classical mechanics for power lines are also described and solved by propounded approach.

The remaining structure of our article is given as follows: In Section 2, by using approximation of control and state variables, we obtain a NLP corresponding to problem $\mathcal{B}$ and also discuss about the method of solving the resulted NLP. Convergence analysis is given in Section 3. In Section 4, we demonstrate the efficiency of the propounded method to several examples and show the computational results of the method. Finally, in Section 5, concluding comments and future extensions are summarized.

2 Proposed method

Consider the problem $\mathcal{B}$ in equations (1) and (2). In this section, by utilizing the properties of Lagrange polynomials and GL integration, the considered OCP is changed to a NLP, then the function NMinimze in Wolfram Mathematica software is used to solve the resulted NLP.

2.1 Discretization of problem $\mathcal{B}$

Firstly, we approximate the control function $u(t)$ and corresponding state function $x(t)$ over the interval $[0,T]$ by the following linear combination of the Lagrange polynomials

$$x_j(t) \approx x_{jN}(t) = \sum_{i=0}^{N} x_{ji} Q_i(t) = {X_j}^T Q(t), \quad j = 1, \cdots, n, \tag{3}$$

and

$$u_k(t) \approx u_{kN}(t) = \sum_{i=0}^{N} u_{ki} Q_i(t) = {U_k}^T Q(t), \quad k = 1, \cdots, m, \tag{4}$$

where $Q(t) = [Q_0(t), Q_1(t), \cdots, Q_N(t)]^T$, $X_j = [x_{j0}, x_{j1}, x_{j2}, \cdots, x_{jN}]^T$ and $U_k = [u_{k0}, u_{k1}, u_{k2}, \cdots, u_{kN}]^T$, in which x_{ji} and u_{ki} for $i = 0, 1, \cdots, N$, $j = 1, \cdots, n$ and $k = 1, \cdots, m$ are the unknown coefficients and, $Q_i(t)$ are the Lagrange polynomials which are defined as follows:

$$Q_i(t) = \prod_{\substack{0 \le s \le N \\ s \ne i}} \frac{t - t_s}{t_i - t_s} = \frac{w(t)}{(t - t_i) w'(t_i)}, \tag{5}$$

where $w(t) = \prod_{s=0}^{N} (t - t_s)$ and t_i, for $i = 0, 1, 2, \cdots, N$, are the zeros of Legendre polynomials [11] transformed to the interval $[0,T]$. The transfer is done according to the interval in which the problem is defined. By discritizing the FIE given in Eq (2) by using (3) and (4) and collocation nodes t_q ,we obtain for $j = 1, \cdots, n$ and $q = 0, \cdots, N$

$$x_{jN}(t_q) = y(t_q) + \int_0^T K_j(t_q, s, x_{1N}(s), \cdots, x_{nN}(s), u_{1N}(s), \cdots, u_{mN}(s)) ds. \tag{6}$$

For using GL quadrature formula, a simple linear transmutation must be produced with the following form

$$s = \frac{T}{2}(\mu + 1), \tag{7}$$

Then, the dynamical system (2) is converted to

$$x_{jN}(t_q) = y(t_q) + \frac{T}{2}\int_{-1}^{1} K_j(t_q, \frac{T}{2}(\mu+1), x_{1N}(\frac{T}{2}(\mu+1)), \cdots, x_{nN}(\frac{T}{2}(\mu+1)), u_{1N}(\frac{T}{2}(\mu+1)), \cdots, u_{mN}(\frac{T}{2}(\mu+1)))d\mu. \tag{8}$$

We apply the GL quadrature for approximating the integral given in Eq (8)

$$x_j(t_q) = y(t_q) + \frac{T}{2}\sum_{l=0}^{N} w_l K_j(t_q, \frac{T}{2}(\mu_l+1), x_{1N}(\frac{T}{2}(\mu_l+1)), \cdots, x_{nN}(\frac{T}{2}(\mu_l+1)), u_{1N}(\frac{T}{2}(\mu_l+1)), \cdots, u_{mN}(\frac{T}{2}(\mu_l+1))), \tag{9}$$

where w_l is defined as follows:

$$w_l = \frac{2}{(1-(\mu_l)^2)(P'_{N+1}(\mu_l))},$$

in which P_{N+1} is the Legendre polynomial of order $N+1$ [11]. By using GL quadrature and equations (3) and (4), we can approximate the cost functional in equation (1) as follows:

$$J(x,u) = \int_0^T F\left(t, x_1(t), \cdots, x_n(t), u_1(t), \cdots, u_m(t)\right) dt \approx \frac{T}{2} \tag{10}$$

$$\int_{-1}^{1} F\left(\frac{T}{2}(\mu+1), x_{1N}\left(\frac{T}{2}(\mu+1)\right), \cdots, x_{nN}\left(\frac{T}{2}(\mu+1)\right), u_{1N}\left(\frac{T}{2}(\mu+1)\right), \cdots, u_{mN}\left(\frac{T}{2}(\mu+1)\right)\right) d\mu \approx \frac{T}{2}\sum_{l=0}^{N} w_l F\left(\frac{T}{2}(\mu_l+1), x_{1N}\left(\frac{T}{2}(\mu_l+1)\right), \cdots, x_{nN}\left(\frac{T}{2}(\mu_l+1)\right), u_{1N}\left(\frac{T}{2}(\mu_l+1)\right), \cdots, u_{mN}\left(\frac{T}{2}(\mu_l+1)\right)\right).$$

Eventually, the OCP governed by a system of FIE in problem $\mathcal{B}$ has been reduced to a NLP by (10) as objective functional and (9) as constraints. We call this NLP as problem $\mathcal{B}_{\mathcal{N}}$.

2.2 Optimization algorithm

In this paper, for solving NLP, which we call as problem $\mathcal{B}_N$, given in equations (9) and (10), the **Nminimize** function in Wolfram Mathematica has been utilized. **Nminimize** find global optimum numerically. This function implement several algorithms for solving optimization problems according to the kind of optimization problem. Now by solving this NLP and determining the vectors X_j and U_k for $j = 1, \cdots, n$ and $k = 1, \cdots, m$, we can find a numerical solution of problem $\mathcal{B}$ given in equations (1) and (2).

3 Convergence analysis

We suppose that the finite set, $\mathcal{P}_N$, consists of all Lagrange polynomials from degree less than or equal to N. The set of $\mathcal{P}_N$ is linearly independent and a basis of the vector space for polynomials of degree at most equal to N. If $\mathcal{F}(t)$ is a function in $L^2[0,T]$, $\mathcal{F}(t)$ has a best estimation $\mathcal{F}_0 \in \mathcal{P}_N$, so we possess [17]

$$\forall\ g \in \mathcal{L}^2[0,T]: \quad ||\mathcal{F} - \mathcal{F}_0||_2 \leq ||\mathcal{F} - g||_2. \tag{11}$$

Suppose $\mathcal{S}_n \in \mathcal{P}_N$, then there is coefficients s_l, $l = 0, 1, \cdots, N$, whereas

$$\mathcal{S}_n = \sum_{i=0}^{N} s_l \mathcal{Q}_l(t) = \mathcal{S}^T \mathcal{Q}(t), \tag{12}$$

where $S = [s_0, s_1, \cdots, s_N]$, and $\mathcal{Q}(t) = [Q_0(t), Q_1(t), \cdots, Q_N(t)]$.

Theorem 1. Let $\mathcal{F} \in L^2[0,1]$ be estimated by Lagrange polynomials $\{\mathcal{Q}_i\}_{i=0}^N$ that is, $\mathcal{F}_N := \sum_{i=0}^N C_i \mathcal{Q}_i(t)$. If e_N is absolutely error of $\mathcal{F}_N$, then $||\mathcal{F} - \mathcal{F}_N||_2 \leq \frac{L}{(N+2)!}$ and $\lim_{N\to\infty} e_N = 0$ and by increasing N; the errors quickly tends to zero. L is a constant.

Proof. The proof is given in [23]. □

Theorem 1 indicates the accuracy estimation of Lagrange polynomials.

Definition 1. The vectors $x(t) \in \Re^n$ and $u(t) \in \Re^m$ are called admissible, if they satisfy in equation (2). The set of admissible pairs $\{(x,u)\}$, for $j = 1, \cdots, n$, is defined as follows:

$$\psi = \{(x(t), u(t)) \Big| x_j(t) = y_j(t) + \int_0^T K_j(t, s, x_1(s), \cdots, x_n(s), u_1(s), \cdots, u_m(s))ds\}. \tag{13}$$

Define U as the admissible control functions. Let

$$x_N = (x_{1N}, x_{2N}, \cdots, x_{nN}), \qquad u_N = (u_{1N}, u_{2N}, \cdots, u_{mN}), \tag{14}$$

then we have

$$x_{jN}(t) = X_j^T \mathcal{Q}(t), \quad j = 1, \cdots, n; \qquad u_{iN}(t) = U_i^T \mathcal{Q}(t), \quad i = 1, \cdots, m, \tag{15}$$

and

$$\psi_N = \Big\{(x_N(t), u_N(t)) \Big| x_{jN}(t) = y(t) + \int_0^T K_j(t, s, x_{1N}(s), \cdots, x_{nN}(s), u_{1N}(s), \cdots, u_{mN}(s))ds\Big\}. \tag{16}$$

The proof of the following theorem is given by utilizing of the same argument in [25], but some revisions are given according to the problem of our paper.

Theorem 2. Assume $J_i^* = \inf_{\psi_i} J$ and $J^* = \inf_{\psi} J$ and J^* is finite and unique, then the following relation is hold.

$$J_1^* \geq J_2^* \geq \cdots J_r^* \geq \cdots \geq J^* = inf_{\psi} J(x, t). \tag{17}$$

Proof. From the definition of ψ_i, the following relation is hold

$$\psi_1 \subset \psi_2 \subset \psi_3 \cdots \subset \psi_N \subset \cdots \subset \psi. \tag{18}$$

According to (18), $\{J_i^*\}$ is a non-increasing sequence that is bounded from below, so it is convergent to J^*. □

Remark 1. The computational distance can be transmuted from $[0, T]$ to any ideal interval via an affine alteration.

Lemma 1. Suppose that t_i and w_i for $1 \leq i \leq N$ are GL nodes and weights respectively, Let $f(t)$ be Rimann integrable. Then [9]

$$\int_{-1}^{1} f(t)dt = \lim_{N \to \infty} \sum_{i=1}^{N} f(t_i) w_i. \tag{19}$$

Theorem 3. Assume that $x^* \in \Re^n$ is exact state solution of equation (2) with $u^* \in \Re^m$ as exact control functions, let x_N^* and u_N^* be numerical solutions of

$$x_{jN}(t_q) = y(t_q) + \int_0^T K(t_q, s, x_{1N}(s) \cdots, x_{nN}(s), u_{1N}, \cdots, u_{mN}))ds. \tag{20}$$

Also assume that the function $K(t, s, x_1(s), \cdots, x_n(s), u_1(s), \cdots, u_m(s))$ satisfies Lipschitz condition with respect to variables x and u as follows:

$$\begin{aligned} &||K(t, s, x_1(s), \cdots, x_n(s), u_1(s), \cdots, u_m(s)) \\ &\quad - K(t, s, x'_1(s), \cdots, x'_n(s), u'_1(s), \cdots, u'_m(s))||_2 \\ &\qquad \leq L_K\{||x_1(s) - x'_1(s)||_2 + \cdots + ||x_n(s) - x'_n(s)||_2 + ||u_1(s) - u'_1(s)||_2 \end{aligned}$$

$$+\cdots+||u_n(s)-u'_m(s)||_2\},\tag{21}$$

where L_K is Lipschitz constant. Then we have:

$$||x^*(t_p)-x^*_N(t_p)||_2\leq\left(\frac{\mathcal{A}}{(N+1)!}\right),\tag{22}$$

where A is a constant which is non-dependent to N.

Proof. We have for $j=1,\cdots,n$,

$$\begin{aligned}
||x^*_j(t)-x^*_{jN}(t)||\leq||&\int_0^T K\left(t,s,x^*_1(s),\cdots,x^*_n(t),u^*_1(s),\cdots,u^*_m(s)\right)ds\\
&-\int_0^T K\left(t,s,x^*_{1N}(s),\cdots,x^*_{nN}(s),u^*_{1N}(s),\cdots,u^*_{mN}(s)\right)ds||_2\\
\leq&\int_0^T||K\left(t,s,x^*_1(s),\cdots,x^*_n(t),\ u^*_1(s),\cdots,u^*_m(s)\right)ds\\
&-K\left(t,s,x^*_{1N}(s),\ \cdots,x^*_{nN}(s),u^*_{1N}(s),\cdots,u^*_{mN}(s)\right)ds||_2\\
\leq& L_K\int_0^T||x^*_1(s)-x^*_{1N}(s)||_2+\cdots+||x^*_n(s)-x^*_{nN}(s)||_2\\
&+||u^*_1(s)-u^*_{1N}(s)||_2+\cdots+,||u^*_m(s)-u^*_{mN}(s)||_2ds\\
=& L_K\sum_{i=1}^n\int_0^T||x^*_i(s)-x^*_{iN}(s)||_2ds\\
&+L_K\sum_{i=1}^m\int_0^T||u^*_i(s)-u^*_{iN}(s)||_2ds
\end{aligned}\tag{23}$$

$$\leq L_KT\sum_{i=1}^n||x^*_i(s)-x^*_{iN}(s)||_2+L_KT\sum_{i=1}^m||u^*_i(s)-u^*_{iN}(s)||_2\tag{24}$$

$$\begin{aligned}
&\leq 2L_KT\left(\frac{\sum_{i=1}^n L_i}{(N+1)!}\right)+2L_KT\left(\frac{\sum_{i=1}^m L'_i}{(N+1)!}\right)\\
&\leq\frac{2L_KT}{(N+1)!}\left(\sum_{i=1}^n L_i+\sum_{i=1}^m L'_i\right)\leq\frac{\mathcal{A}}{(N+1)!}.
\end{aligned}\tag{25}$$

In (25) $\mathcal{A}=2L_KT\left(\sum_{i=1}^n L_i+\sum_{i=1}^m L'_i\right)$, so by increasing N for $j=1,2,\cdots,n$, $||x^*_j(t)-x^*_{jN}(t)||$ goes to zero. □

4 Numerical experiments

In this section, the proficiency of the proposed technique is shown by applying to some given numerical examples of OCP for systems governed by FIE.

Example 1. Observe the following OCP

$$\min\ J(x(t),u(t)) = \int_0^1 (x(t) - 0.8182 - 2.7273t^2)^2 + (u(t) - t^2)^2 dt, \quad (26)$$

subject to

$$x(t) = t^2 + \int_0^1 (t^2 + u(s))x(s)ds, \quad (27)$$

with the exact solutions, $u^*(t) = t^2$ and $x^*(t) = 0.8182 + 2.7273t^2$.

By exerting the propounding method, the approximate solutions are approximately $x^*(t) = 0.8182 - 8.78094 * 10^{-15}t + 2.7273t^2$ and $u^*(t) = 1.47439 * 10^-6 + 1.t^2$. The optimal value of J^* for $N = 2$ is $1.20133 * 10^{-11}$.

Example 2. Investigate the following minimization problem

$$J(x_1(t),x_2(t),u(t)) = \int_{-1}^1 (x_1(t)-t^2+1)^2+(x_2(t)-1+t^2)^2+(u(t)-\sin(t))^2 dt, \quad (28)$$

subject to FIE system

$$x_1(t) = t^2 - 1 + \int_{-1}^1 t^3 u(s)\,(x_1(s) + x_2(s))\,ds,$$
$$x_2(t) = 1 - t^2 + \int_{-1}^1 (s - u(s))(x_1(s) + x_2(s))ds.$$

The precise optimal solutions of this OCP are $x_1^*(t) = t^2 - 1$, $x_2^*(t) = 1 - t^2$ and $u^*(t) = \sin(t)$.

After dissolving with the propounded approach, we obtain the following result

$$x_1{}^*(t) = -1 + 5.55112 \times 10^{-17}t + t^2,$$
$$x_2{}^*(t) = 1 + 5.55112 \times 10^{-17} - 1t^2$$
$$u^*(t) = -1.50469 \times 10^{-24} + 0.902957t - 1.11022 \times 10^{-16}t^2,$$

which approximates the exact functions with good approximation. The Maclaurin series of sine function is $\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots$. As we can see in the approximate control function $u^*(t) = -1.50469 \times 10^{-24} + 0.902957t - 1.11022 \times 10^{-16}t^2$, the constant sentence and the coefficient of sentence includes x^2 is approximately equal to zero. The optimal value of objective

function for different value of N is given in Table 1. The numerical results of this example with sinc wavelet for $k = 5$ and $N = 10,\ \ 20$ given in [16] are reported in Table 1, which is comparable with our method. The value of the objective function for $N = 2$ is more accurate than its value for $N = 3$, because for $N = 2$, the quadratic functions of the state are approximated with better accuracy. The plot of approximate function $u^*(t)$ for $N = 3$ is given in Figure 1. The plots of the absolute error of state functions x_1 and x_2 for $N = 3$ are demonstrated in Figures 2 and 3.

Table 1: Optimal value of J^* for Example 2

N	2	3
J^*	8.55969×10^{-33}	1.07179×10^{-31}
N [16]	10	20
J^*	6.88×10^{-12}	6.88×10^{-12}

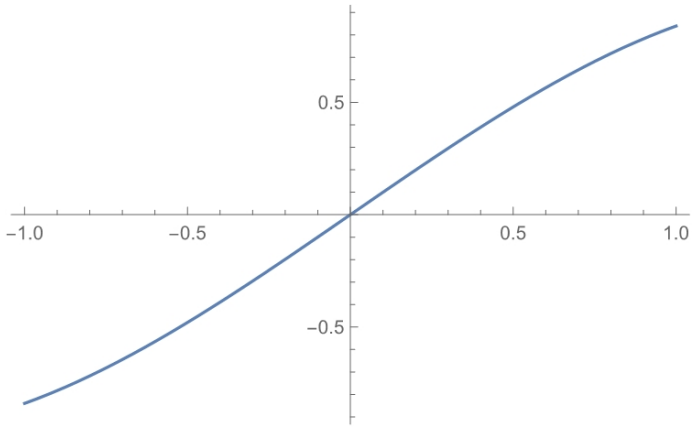


Figure 1: The approximate function of u^* in Example 2

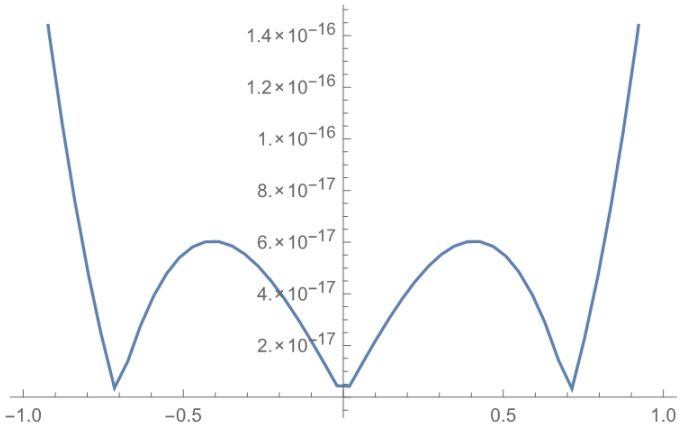


Figure 2: The absolute error of x_1^* in Example 2

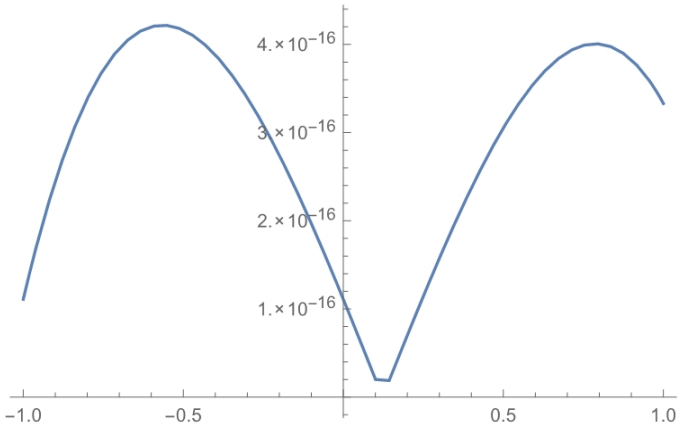


Figure 3: The absolute error of x_2^* in Example 2

Example 3. Consider the following OCP governed by FIE systems

$$\min \int_0^1 (x_1(t) - t) + \left(x_2(t) - t^2\right) + \left(u(t) - e^{-t}\right)$$

subject to

$$x_1(t) = t - \frac{e^{-t}}{4} + \int_0^1 u(t)x_1(s)x_2(s)ds,$$

$$x_2(t) = t^2 - \frac{e^{-t}}{5} + \int_0^1 u(t){x_1}^2(s)x_2(s)ds$$

The exact solution of this OCP is ${x_1}^*(t) = t$, ${x_2}^*(t) = t^2$ and $u^*(t) = e^{-t}$.

The approximate and exact optimal control and state functions for $N = 2$ are shown in Figure 4, 5 and 6, respectively. The optimal values of J^* for different values of N are given in the Table 2.

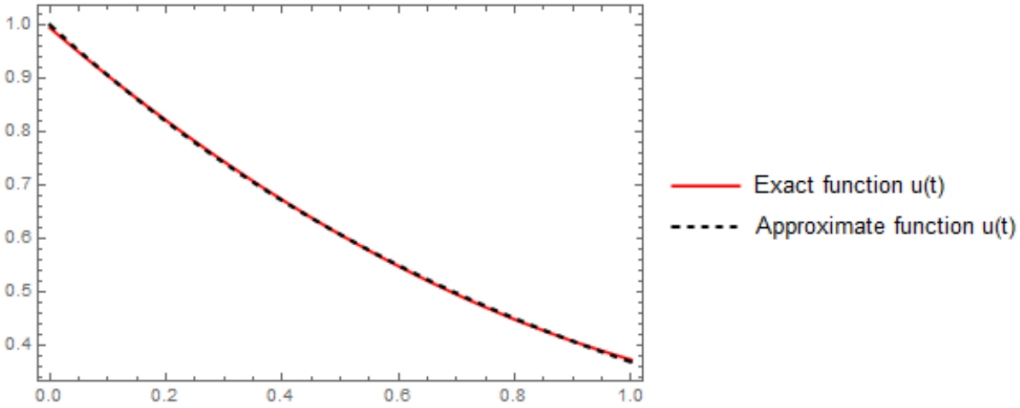


Figure 4: The exact and approximate optimal control for $N = 2$ in Example 3

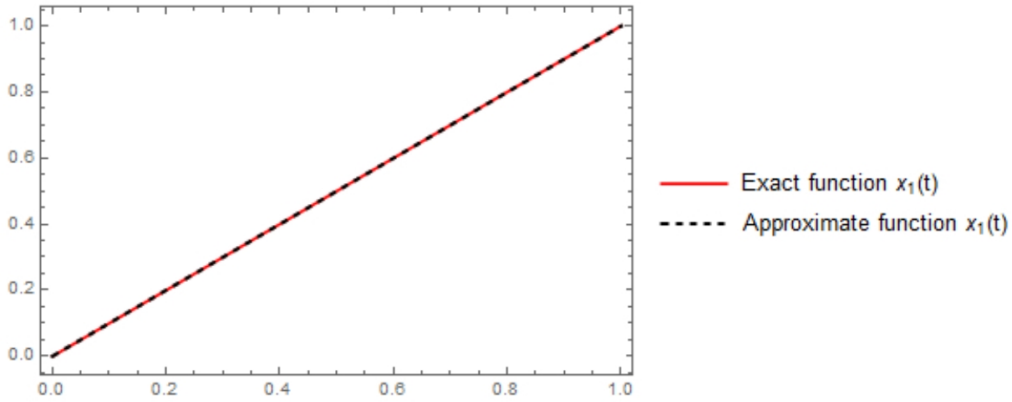


Figure 5: The exact and approximate optimal state $x_1(t)$ for $N = 2$ in Example 3

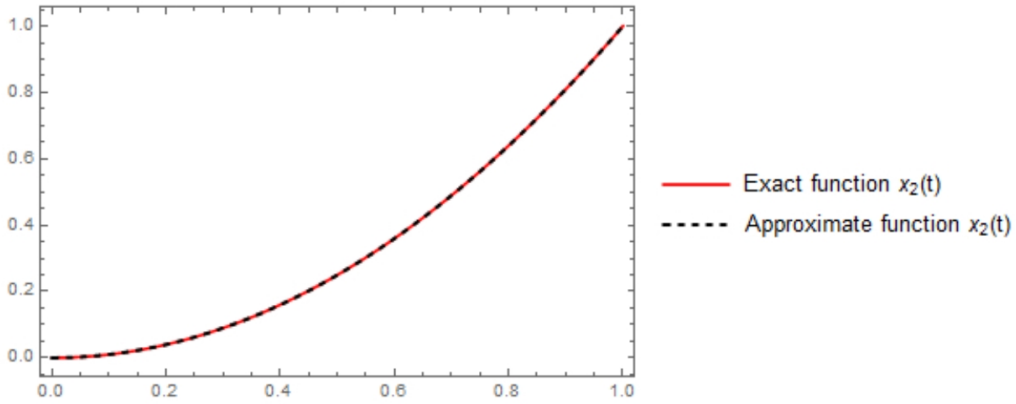


Figure 6: The exact and approximate optimal state $x_2(t)$ for $N = 2$ in Example 3

Table 2: Optimal value of J^* for Example 3

N	2	3	4
J^*	1.10430×10^{-7}	5.40995×10^{-10}	1.11407×10^{-12}

Example 4 (Hanging chain)**.** In this example, we investigate a problem from classical mechanics, which appears in power lines. If a flexible chain is hung from two points, it forms a U shape that is called a catenary curve. The resulting shape of the chain depends on the mass distribution along the chain. Consider the following FIE from [14],

$$x(s) = g \int_0^L G(s,t)\phi(t)dt, \tag{29}$$

Here, $x(s)$ and $\phi(s)$ are demonstrated the displacement of the chain and the mass density of the chain respectively. g is the gravitational constant and

$$G(s,t) = \begin{cases} \frac{s(L-t)}{T_0 L} & 0 \le s \le t \\ \frac{t(L-s)}{T_0 L} & t \le s \le L \end{cases} \tag{30}$$

In (30), T_0 is the constant tension. In this example, we are looking for the mass density distribution as the control function $\phi(t)$ that leads to a prescribed shape $x(t)$ as its corresponding state. Therefore, the OCP problem is given as the minimization of the cost functional

$$J(x,u) = \int_0^L \left(\frac{g(tL^2 - t^3)}{6T_0} - x(t) \right)^2 + \left(\phi(x) - x\right)^2 dx \tag{31}$$

subject to equation (29). The optimal control and state solutions are $\phi^*(x) = x$ and $x^*(t) = \frac{g(tL^2 - t^3)}{6T_0}$. In this example, we consider the following parameters as $L = 1$, $g = 9.8$, $T0 = 1$. By applying the proposed algorithm for $N = 3$, the graphs of the exact and approximate controls and trajectories are given in figures 7 and 8. The optimal value of J^* is 3.59596×10^{-4}. The graph of error for control and state function is given in figures 9a and 9b.

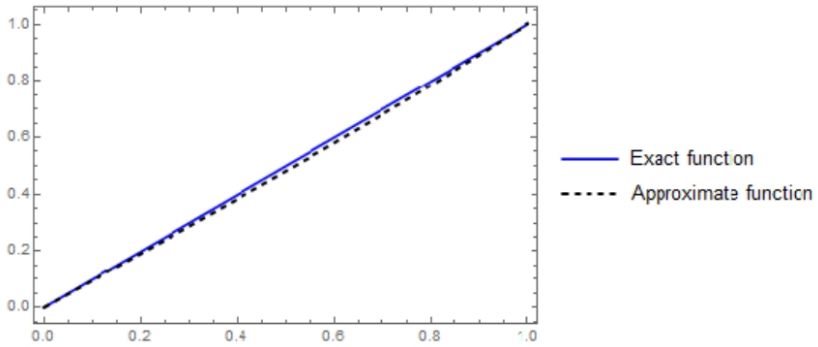


Figure 7: The exact and approximate optimal control in Example 4

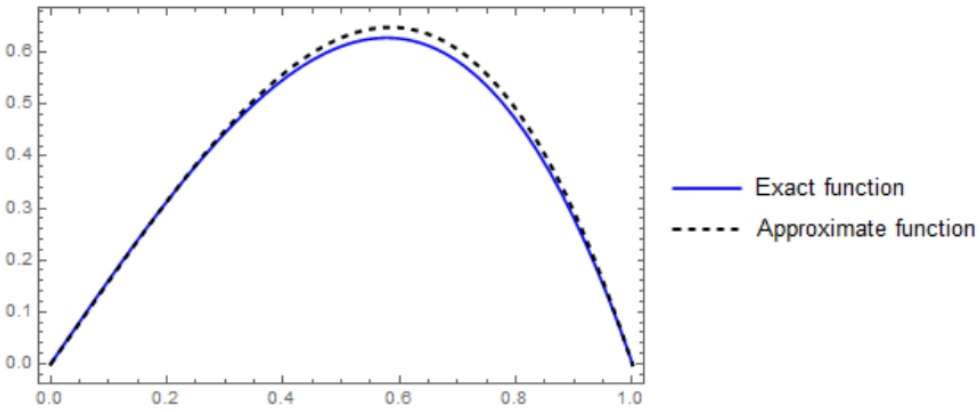
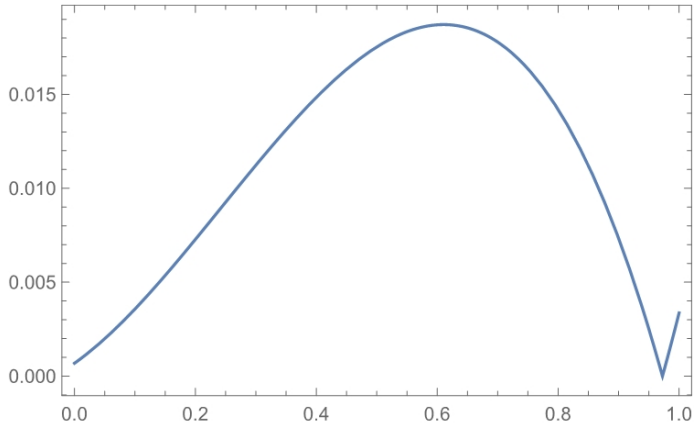
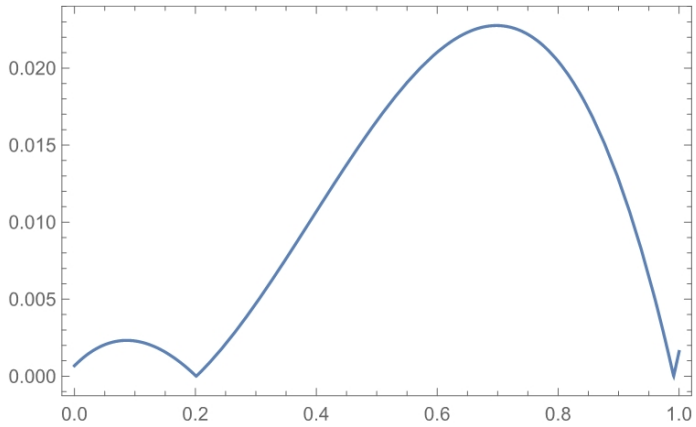


Figure 8: The exact and approximate optimal state in Example 4



(a) The error of state function in Example 4



(b) The errors of control function in Example 4

Figure 9

Example 5. Find the optimal control u^* and corresponding optimal state x^* that minimize objective function

$$\min\ J = \int_0^1 (x_1(t) - \sin(t))^2 + \left(x_2(t) - e^{-3t}\right)^2 + (u(t) - t)^2\, dt$$

subject to

$$x_1(t) = h_1(t) - \int_0^1 \left(-t(u(s))^2 x_1(s) + tu(s)x_2(s)\right) ds,$$
$$x_2(t) = h_2(t) - \int_0^1 (t(u(s)+1)x_1(s) + t^2 u(s)x_2(s))ds.$$

where $h_1(t) = -\sin(5t) - \frac{23\cos(5)}{125}t + \frac{2\sin(5)}{15}t + \frac{107}{1125}t - \frac{4e^{-3}}{9}t$ and $h_2(t) = e^{-3t} + \frac{2cos(5)}{5}t - \frac{sin(5)}{25}t - \frac{t}{5} - \frac{4e^{-3}}{9}t^2 + \frac{t^2}{9}$. The exact solutions are $x_1^*(t) = \sin(t)$, $x_2^*(t) = e^{-3t}$ and $u^*(t) = t$. Figures 10, 11 and 12 are the graph of exact and approximate solutions for $N = 3$ and $N = 5$, respectively.

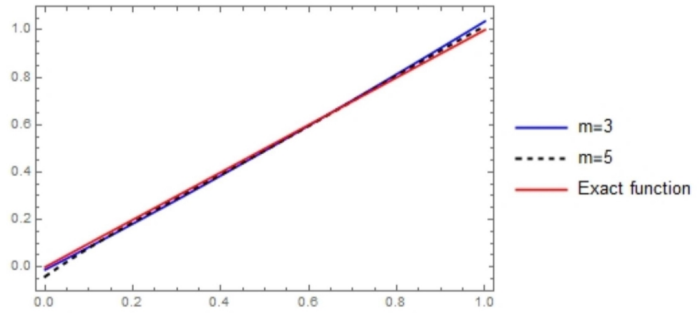


Figure 10: The precise and proximate optimal control in Example 5

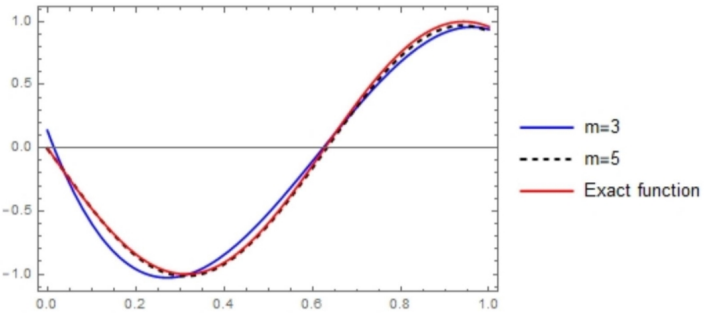


Figure 11: The accurate and proximate optimal state x_1 in Example 5

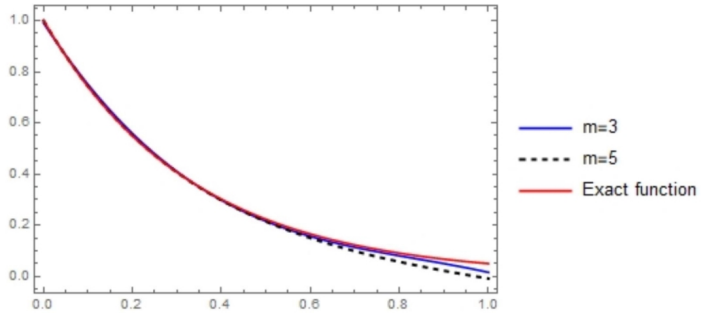


Figure 12: The accurate and proximate optimal state x_2 in Example 5

Example 6. Find the optimal solutions that minimize objective function

$$\min J = \int_0^1 \left(x_1(t) - (t + e^t)\right)^2 + \left(x_2(t) - e^t\right)^2 + \left(x_3(t) - (1 + \cos(t))\right)^2 + \left(u(t) - e^t\right)^2 dt$$

subject to

$$x_1(t) + \int_0^1 3su(t)x_2(s)ds = t + 4e^t,$$
$$x_2(t) - \int_0^1 6te^s x_1(s) - t^2 x_3(s)ds = e^t - 3t(e^2 + 1) + t^2(\sin(1) + 1)$$
$$x_3(t) - \int_0^1 (4(x_3(s) - 1) + x_2(s)ds = \cos(t) - (e + 4\sin(1)) + 2$$

The exact solutions of this example are $x_1^*(t) = t + e^t$, $x_2^*(t) = e^t$, $x_3^*(t) = 1 + \cos(t)$ and $u^*(t) = e^t$. The absolute errors of u^*, x_1^*, x_2^* and x_3^* for $N = 3$, $N = 4$ and $N = 5$ are represented in Tables 3 and 4.

Table 3: Absolute errors of control and state functions in Example 6

	N	$t = 0$	$t = 0.25$	$t = 0.5$	$t = 0.75$	$t = 1$
E_{x_1}	3	7.51115×10^{-4}	2.45764×10^{-4}	4.67421×10^{-4}	2.56842×10^{-4}	9.41652×10^{-4}
	4	3.99592×10^{-5}	7.55614×10^{-6}	5.88693×10^{-7}	9.47055×10^{-6}	486114×10^{-5}
	5	1.83038×10^{-6}	9.19475×10^{-7}	8.9541×10^{-7}	9.96471×10^{-7}	2.13043×10^{-6}
E_{x_2}	3	4.91524×10^{-4}	2.60616×10^{-4}	5.32509×10^{-4}	3.01636×10^{-4}	5.72133×10^{-4}
	4	2.49372×10^{-5}	1.28883×10^{-5}	1.78345×10^{-7}	1.4376×10^{-5}	2.98331×10^{-5}
	5	1.05009×10^{-6}	1.08341×10^{-6}	1.13318×10^{-6}	1.15422×10^{-6}	1.19212×10^{-6}
E_{x_3}	3	3.36834×10^{-4}	1.04755×10^{-4}	3.23772×10^{-4}	9.70093×10^{-4}	3.05991×10^{-4}
	4	6.42153×10^{-6}	3.47200×10^{-6}	9.92872×10^{-6}	4.26175×10^{-6}	8.97689×10^{-6}
	5	6.20439×10^{-7}	6.10935×10^{-7}	5.82297×10^{-7}	5.87852×10^{-7}	5.74393×10^{-7}
E_u	3	4.34033×10^{-4}	2.24118×10^{-4}	6.13891×10^{-4}	2.32091×10^{-4}	5.65502×10^{-4}
	4	1.93484×10^{-5}	1.40343×10^{-5}	7.62504×10^{-7}	1.68747×10^{-5}	2.46499×10^{-5}
	5	8.13269×10^{-7}	1.17233×10^{-6}	1.16398×10^{-6}	1.27319×10^{-6}	9.68081×10^{-7}

Table 4: Absolute errors of objective functions in Example 6

N	3	4	5
J^*	1.01824×10^{-7}	2.36092×10^{-10}	4.34198×10^{-13}

5 Conclusion

The FIEs have rich physical and Engineering backgrounds. These equations got superior fondness across plenty of disciplines and broadly utilized in dynamical machinery with chaotic or quasi-chaotic treatment. In this article, a

robust and efficient numerical approach based on Lagrange polynomials and collocation approach is applied to obtain the solutions of OCP governed by FIE system. At last, the prescribed problem is converted to a NLP. The resulted NLP is solved by optimization algorithm. The preciseness of Lagrange collocation approach can be readily concluded from the ameliorated results of our presented approach. The propounded approach can be extended for the approximation of solutions in OCPs involving FIE with fractional [5, 6, 12, 27] and singular [15] kernels, however some modifications will be needed in the method and convergence analysis.

References

[1] Abu Arqub, O. and Shawagfeh, N. *Solving optimal control problems of Fredholm constraint optimality via the reproducing kernel Hilbert space method with error estimates and convergence analysis*, Math. Methods Appl. Sci. 44(10) (2021), 7915–7932.

[2] Alipour, M. and Soradi-Zeid, S. *Optimal control of time delay Fredholm integro-differential equations*, J. Math. Model. 9(2) (2021), 277–291.

[3] Almasieh, H. and Roodaki, M. *Triangular functions method for the solution of Fredholm integral equations system*, Ain Shams Eng. J. 3(4) (2012), 411–416.

[4] Babolian, E., Biazar, J. and Vahidi, A.R. *The decomposition method applied to systems of Fredholm integral equations of the second kind*, Appl. Math. Comput. 148(2) (2004), 443–452.

[5] Baghani, O. *Second Chebyshev wavelets (SCWs) method for solving finite-time fractional linear quadratic optimal control problems*, Math. Comput. Simul. 190 (2021), 343–361.

[6] Baghani, O. *SCW-iterative-computational method for solving a wide class of nonlinear fractional optimal control problems with Caputo derivatives*, Math. Comput. Simul. 202 (2022), 540–558.

[7] Basit, M. and Khan, F. *An effective approach to solving the system of Fredholm integral equations based on Bernstein polynomial on any finite interval*, Alex. Eng. J. 61(4) (2022), 2611–2623.

[8] Borzabadi, A.H., Fard, O.S. and Mehne, H.H. *A hybrid algorithm for approximate optimal control of nonlinear Fredholm integral equations*, Int. J. Math. 89(16) (2012), 2259–2273.

[9] Canuto, C., Hussaini, M.Y., Quarteroni, A. and Zang, T.A. *Spectral methods in fluid dynamics*, Springer Verlag, New York, 1988.

[10] Ebrahimzadeh, A., Khanduzi, R., Panjeh Ali Beik, S. and Baleanu, D. *Research on a collocation approach and three metaheuristic techniques based on MVO, MFO, and WOA for optimal control of fractional differential equation*, J. Vib. Control (2021), 10.1177/10775463211051447.

[11] Ebrahimzadeh, A. and Panjeh Ali Beik, S. *Application of spectral methods to solve nonlinear buckling analysis of an elastic beam*, Casp. J. Math. Sci., (CJMS) peer 10(1) (2021), 68–76.

[12] Habibli, M., Noori Skandari, M.H. *Fractional Chebyshev pseudospectral method for fractional optimal control problems*, Optim Control Appl Meth. 40(3) (2019), 558–572.

[13] Heydari, M.H., Razzaghi, M. and Avazzadeh, Z. *Orthonormal piecewise Bernoulli functions: Application for optimal control problems generated using fractional integro-differential equations*, J. Vib. Control (2022), 10775463211059364.

[14] Jerri, A.J. *Introduction to Integral Equations with Applications,Wiley-Interscience*, NewYork, 1999.

[15] Kaneko, H. and Xu, Y. *Numerical solutions for weakly singular Fredholm integral equations of the second kind*, Applied Numerical Mathematics 7 (1991), 167–177.

[16] Keyanpour, M. and Akbarian, T. *Optimal control of Fredholm integral equations*, Int. J. Applied Math. and Information Sci 5(3) (2011), 514–524.

[17] Kreyszig, E. *Introductory Functional Analysis With Applications*, John Wiley, New York, 1978.

[18] Lefebvre, M. *Optimal control of an Ornstein-Uhlenbeck process*, Stochastic Processes and their Applications 24(1) (1987), 89–97.

[19] Maleknejad, K. and Ebrahimzadeh, A. *Optimal control of Volterra integro-differential systems based on Legendre wavelets and collocation method*, Int. J. Comput. Sci 8(7) (2014), 1040–1044.

[20] Maleknejad, K. and Ebrahimzadeh, A. *The use of rationalized Haar wavelet collocation method for solving optimal control of Volterra integral equations*, J. Vib. Control 21(10) (2015), 1958–1967.

[21] Maleknejad, K., Aghazadeh, N. and Rabbani, M. *Numerical solution of second kind Fredholm integral equations system by using a Taylor-series expansion method*, Appl. Math. Comput. 175(2) (2006), 1229–1234.

[22] Rashidinia, J. and Zarebnia, M. *Convergence of approximate solution of system of Fredholm integral equations*, J. Math. Anal. Appl. 333(2) (2007), 1216–1227.

[23] Rivlin, T.J. *An introduction to Approximation of functions*, Dover, Network, 1969.

[24] Roubíček, T. *Optimal control of nonlinear Fredholm integral equations*, J. Optim. Theory Appl. 97(3) (1998), 707–729.

[25] Safaie, E., Farahi, M.H. and Farmani Ardehaie, M. *An approximate method for numerically solving multi-dimensional delay fractional optimal control problems by Bernstein polynomials*, Comput. Appl. Math. 34(3) (2015), 831–846.

[26] Wazwaz, A.M. *Nonlinear Singular Integral Equations. In Linear and Nonlinear Integral Equations*, Springer, Heidelberg, Berlin, 2011.

[27] Xiaobing, P., Yang, X., Skandari, M.H.N., Tohidi, E. and Shateyi, S. *A new high accurate approximate approach to solve optimal control problems of fractional order via efficient basis functions*, Alex. Eng. J. 61(8) (2022), 5805–5818.

How to cite this article
Ebrahimzadeh, A., A robust method for optimal control problems governed by system of Fredholm integral equations in mechanics. *Iran. j. numer. anal. optim.*, 2023; 13(2): 243-261. https://doi.org/10.22067/ijnao.2022.76652.1178

Iranian Journal of Numerical Analysis and Optimization
Vol. 13, No. 2, 2023, pp 262–284
https://doi.org/10.22067/ijnao.2022.76652.1161
https://ijnao.um.ac.ir/



How to cite this article
Research Article



Fitted scheme for singularly perturbed time delay reaction-diffusion problems

M. Amsalu Ayele*, , A. Andargie Tiruneh and G. Adamu Derese

Abstract

In this article, we constructed a numerical scheme for singularly perturbed time-delay reaction-diffusion problems. For the discretization of the time derivative, we used the Crank-Nicolson method and a hybrid scheme, which is a combination of the fourth-order compact difference scheme and the central difference scheme on a special type of Shishkin mesh in the spatial direction. The proposed scheme is shown to be second-order accurate in time and fourth-order accurate with a logarithmic factor in space. The uniform convergence of the proposed scheme is discussed. Numerical investigations are carried out to demonstrate the efficacy and uniform convergence of the proposed scheme, and the obtained numerical results reveal the better performance of the present scheme, as compared with some existing methods in the literature.

AMS subject classifications (2020): Primary 65M06, Secondary 65M12, 65M15.

Keywords: Singular perturbation; Time-delay, Parabolic differential equation, Reaction-diffusion problem, Hybrid scheme.

---

*Corresponding author

Received 30 June 2022; revised 28 October 2022; accepted 9 November 2022

Mulunesh Amsalu Ayele
Department of Mathematics, College of Sciences, Bahir Dar University, Bahir Dar, Ethiopia. e-mail: mulu287@yahoo.com

Awoke Andargie Tiruneh
Department of Mathematics, College of Sciences, Bahir Dar University, Bahir Dar, Ethiopia. e-mail: awoke248@yahoo.com

Getachew Adamu Derese
Department of Mathematics, College of Sciences, Bahir Dar University, Bahir Dar, Ethiopia. e-mail: getachewsof@yahoo.com

1 Introduction

Delay differential equations play a crucial role in the mathematical modeling of various practical phenomena and are widely applicable in fields such as biosciences, control theory, economics, material science, medicine, robotics, etc. [17, 16, 15]. If we restrict the delay differential equations to a class in which the highest order term is multiplied by a small parameter ε, then we call it singularly perturbed delay differential equations of the retarded type. Nowadays, there has been a growing interest in the study of singularly perturbed delay differential equations due to their occurrence in many practical models, for instance, Hutchinson's equation which is a model for the evolution of the population in mathematical ecology[1] and, the Wazewska-Czyzeska and Lasota equation that describes the survival of red blood cells in animals [27].

The solutions of singularly perturbed time-delayed problems are fundamentally different from those problems without time delay because small lags can have large effects. The solution of delay differential equations requires knowledge of both the current state and the state at a certain time in the past. Finding the solution to singularly perturbed delay differential equations using classical numerical methods fails to give stable and accurate results because of the presence of the perturbation parameter ε.

Many researchers have treated time-dependent singularly perturbed parabolic partial differential equations,with or without time-delay: to mention some [3, 4, 14, 12, 2, 10, 20, 21, 22, 23, 28]. However, time-delayed reaction diffusion problems have not been extensively investigated. Authors in [13] solved singularly perturbed time-delay parabolic partial differential equations of the reaction-diffusion type. The problem is discretized by a hybrid scheme on a generalized Shishkin mesh in the spatial direction and the implicit Euler scheme on a uniform mesh in the time direction, and to increase the order of convergence in the time direction, Richardson extrapolation is applied. The authors in [12] studied a similar problem with [13]. The scheme uses the Crank-Nicolson scheme for the time variable, and the spatial variable is discretized by the tension spline scheme on a non-uniform Shishkin mesh.

Recently, in [19] Crank-Nicolson method for the time derivative on the uniform mesh and the central difference scheme for the spatial derivative on the Shishkin-type meshes is used for the problem in [13], and to enhance the order of convergence Richardson extrapolation technique is used. Also, in [7] singularly perturbed delay parabolic reaction-diffusion problem with mixed type boundary condition is solved. The problem is discretized by the implicit Euler method on a uniform mesh in the time and the extended cubic B-spline collocation method on a Shishkin mesh in the space variable.

In real-world applications, higher-order methods are preferred to their lower-order counterparts since they provide better accuracy at a lower computing cost. This paper aims to design a uniformly convergent scheme with a higher order of convergence in both space and time variables or singularly

perturbed time delay reaction-diffusion problems. To discretize the problem, a similar approach to [14] is applied, that is, the Crank-Nicolson scheme for the discretization of the time derivative, and for the spatial variable, a hybrid scheme on a special type of Shishkin mesh is used, which is gradually condensing starting from the center of the domain to both the right and left boundary layers.

In this article, we considered the following singularly perturbed time-delay parabolic partial differential equation of the reaction-diffusion type of the form:

$$(\frac{\partial}{\partial t} + \mathcal{L}_{\varepsilon,x})u(x,t) = -b(x,t)u(x,t-\tau) + f(x,t), (x,t) \in D \tag{1}$$

subject to the initial and interval conditions:

$$u(x,t) = \phi_b(x,t), \text{ for } (x,t) \in \Gamma_b, \tag{1a}$$

and boundary conditions:

$$u(0,t) = \phi_l(t) \quad \text{on } \Gamma_l = \{(0,t) : 0 \le t \le T\}, \tag{1b}$$

$$u(1,t) = \phi_r(t) \quad \text{on } \Gamma_r = \{(1,t) : 0 \le t \le T\}, \tag{1c}$$

where

$$\mathcal{L}_{\varepsilon,x}u(x,t) = -\varepsilon u_{xx}(x,t) + a(x,t)u(x,t),$$

$0 < \varepsilon << 1$ and $\tau > 0$ is the delay parameter. Also, $D = \Omega_x \times \Omega_t = (0,1) \times (0,T]$ and $\Gamma = \Gamma_l \cup \Gamma_b \cup \Gamma_r$, $\Gamma_b = [0,1] \times [-\tau, 0]$, Γ_l and Γ_r are the left and the right sides of the rectangular domain D corresponding to $x = 0$ and $x = 1$, respectively. The functions $a(x,t), b(x,t), f(x,t), \phi_b, \phi_l$ and ϕ_r are assumed to be sufficiently smooth and bounded and satisfying,

$$a(x,t) \ge \alpha > 0, b(x,t) \ge \beta > 0, (x,t) \in \overline{D}.$$

For the existence and uniqueness of the solution of (1), see [1]. Under the above assumptions the solution of (1) exhibits boundary layers along $x = 0$ and $x = 1$.

The article is organized as follows: In Section 2, we provide some properties of the analytical solution and its derivatives. In Section 3, temporal semi-discretization, spatial discretization, and the derivation of the scheme are discussed. In Section 4, we investigate the uniform convergence of the fully discrete scheme. Numerical results and discussion are presented in Section 5. Finally, in Section 6, the conclusion of the paper is provided.

Notation : Throughout this paper, C denotes a generic positive constant that is independent of ε and mesh sizes. Also $||.||$ denotes the standard supremum norm, which is defined as $||f|| = \underset{(x,t)\in D}{Sup} |f(x,t)|$, for a function f defined on some domain D.

2 Analytical results

In this section, the analytical aspects of the solution of problem (1) and its derivatives are studied.

Lemma 1 (Maximum principle)**.** Let $\Phi(x,t) \in C^{2,1}(\overline{D})$ and assume that $\Phi(x,t) \geq 0$ on Γ.
Then $\left(\frac{\partial}{\partial t} + \mathcal{L}_{\varepsilon,x}\right)\Phi(x,t) \geq 0, \forall (x,t) \in D$ imply that $\Phi(x,t) \geq 0, \forall (x,t) \in \overline{D}$.

Proof. Let the point $(x^*,t^*) \in \overline{D}$, such that $\Phi(x^*,t^*) = \min \Phi(x,t)$ and assume that $\Phi(x^*,t^*) < 0$. Clearly, $(x^*,t^*) \notin \Gamma$ these implies that $(x^*,t^*) \in D$.

The function $\Phi(x,t)$ attains its minimum at (x^*,t^*), then $\Phi_x = 0, \Phi_t = 0$ and $\Phi_{xx} \geq 0$ at (x^*,t^*) . Therefore, from (1), it is easy to observe that

$$(\frac{\partial}{\partial t} + \mathcal{L}_{\varepsilon,x})\Phi(x,t) \leq 0,$$

which is a contradiction as $(\frac{\partial}{\partial t} + \mathcal{L}_{\varepsilon,x})\Phi(x,t) \geq 0$. Hence $\Phi(x,t) \geq 0$ for all $(x,t) \in \overline{D}$. □

Lemma 2. The solution u of (1) satisfies the following bounds,

$$||\frac{\partial^{i+j}u}{\partial x^i \partial t^j}|| \leq C\varepsilon^{-i/2}, \quad \text{for } 0 \leq i+2j \leq 8. \tag{2}$$

Proof. See [13]. □

In the proof of the error analysis, sharper bounds of the solution and its derivatives are required so that the solution u is decomposed into a regular component v, and a singular component w, as follows:

$$u = v + w,$$

for more detail see [1, 18].

Lemma 3. The solution u of (1) satisfies the following bounds,

$$||\frac{\partial^{i+j}v}{\partial x^i \partial t^j}|| \leq C(1+\varepsilon^{2-i/2}), \tag{3}$$

$$||\frac{\partial^{i+j}w}{\partial x^i \partial t^j}|| \leq C\varepsilon^{-i/2}(\exp(-\sqrt{\alpha}x/\sqrt{\varepsilon}) + \exp(-\sqrt{\alpha}(1-x)/\sqrt{\varepsilon})), \tag{4}$$

$$\text{for } \ (x,t) \in D, 0 \leq i+2j \leq 8.$$

Proof. The proof of this Lemma is described in [1]. □

3 Numerical scheme

3.1 Temporal semi-discretization

On the time domain $[0,T]$, we use uniform mesh with step size Δt, such that $\Omega_t^M = \{t_j : t_j = j\Delta t, \Delta t = \frac{T}{M},\ \text{for } j = 0,1,...,M\}$, $T = y\tau$, $\tau = m_\tau \Delta t$, where y and m_τ are positive integer, and M is the number of mesh intervals in $[0,T]$. To descretize the time variable for problem (1), we use Crank-Nicolson method, given by

$$\begin{cases} u^{-j}(x) = \phi_b(x,-t_j),\ \text{for } j = 0,1,2,...,m_\tau, 0 \le x \le 1, \\ \left(I + \frac{\Delta t}{2}\mathcal{L}_{\varepsilon,x}\right)u^{j+1}(x) = \left(I - \frac{\Delta t}{2}\mathcal{L}_{\varepsilon,x}\right)u^j(x) + \frac{\Delta t}{2}\left[F^j(x) + F^{j+1}(x)\right], \\ u^{j+1}(0) = \phi_l, u^{j+1}(1) = \phi_r,\ \text{for } j = 1,2,...,M-1, \end{cases} \tag{5}$$

where $F^j(x) = f(x,t_j) - b(x,t_j)u(x,t_{j-m_\tau})$ and, $u^{j+1}(x) = u(x,t_{j+1})$ is the semi-discrete approximation to the exact solution $u(x,t)$ of (1) at the time level $t_j = j\Delta t$. The local truncation error of the semi-discrete method (5) is given by

$$e_{j+1} = u(x,t_{j+1}) - \widetilde{u}^{j+1}(x),$$

where $\tilde{u}^j(x)$ is the solution evaluated after one step of the semi-discrete scheme (5) taking the exact value $u(x,t_j)$ instead of u^j as the initial data. Re-write (5) in the form

$$\begin{cases} \left(I + \frac{\Delta t}{2}\mathcal{L}_{\varepsilon,x}\right)\widetilde{u}^{j+1}(x) = g^j(x), x \in \Omega_x, \\ \widetilde{u}^{j+1}(0) = \phi_l,\ \ \widetilde{u}^{j+1}(1) = \phi_r, \end{cases} \tag{6}$$

where $g^j(x) = \frac{\Delta t}{2}\left(F^j(x) + F^{j+1}(x) - \mathcal{L}_{\varepsilon,x}u^j(x)\right) + u(x,t_j)$.

For the stability of the Crank-Nicolson method, one can refer to [6, 8].

Lemma 4. Suppose that $\left|\frac{\partial^i}{\partial t^i}u(x,t)\right| \le C, \quad (x,t) \in \overline{D}, \quad i = 0,1,2,3,$ the local error associated to scheme (5) satisfies:

$$||e_{j+1}|| \le C_1(\Delta t)^3, \quad j = 1,2,...,M. \tag{7}$$

Proof. Using Taylor's series expansion,

$$\begin{aligned} \frac{u(x,t_{j+1}) - u(x,t_j)}{\Delta t} &= u_t(x,t_{j+1/2}) + O((\Delta t)^2), \\ &= -\mathcal{L}_{\varepsilon,x}u(x,t_{j+1/2}) + F(x,t_{j+1/2}) + O((\Delta t)^2), \end{aligned} \tag{8}$$

where $F(x,t_{j+1/2}) = -b(x,t_{j+1/2})u(x,t_{j+1/2-m_\tau}) + f(x,t_{j+1/2})$ and

$$f(x,t_{j+1/2}) = \frac{f(x,t_{j+1})+f(x,t_j)}{2} + O((\Delta t)^2),$$
$$a(x,t_{j+1/2}) = \frac{a(x,t_{j+1})+a(x,t_j)}{2} + O((\Delta t)^2),$$
$$b(x,t_{j+1/2}) = \frac{b(x,t_{j+1})+b(x,t_j)}{2} + O((\Delta t)^2).$$

From the above, we obtain

$$\mathcal{L}_{\varepsilon,x}u(x,t_{j+1/2}) = \mathcal{L}_{\varepsilon,x}\frac{u(x,t_{j+1})+u(x,t_j)}{2} + O((\Delta t)^2).$$

Clearly, the local error $||e_{j+1}||$ is the solution of the following BVP:

$$\begin{cases} \left(I + \frac{\Delta t}{2}\mathcal{L}_{\varepsilon,x}\right)e_{j+1} = O((\Delta t)^3), \\ e_{j+1}(0) = 0, \quad e_{j+1}(1) = 0. \end{cases} \tag{9}$$

Next, using the maximum principle for the operator $\left(I + \frac{\Delta t}{2}\mathcal{L}_{\varepsilon,x}\right)$ proves the result, for more detail please refer [6]. □

All the results in the above insure the second order convergence of the scheme (5). The global error of the time semi-discretization is given by

$$E_j = u(x,t_j) - w^j(x) = \sum_{k=1}^{j} e_k.$$

Theorem 1 (Global error estimate)**.** The global error estimate at t_j is

$$||E_j|| \leq C(\Delta t)^2, \quad j = 1,2,...,M.$$

Proof. see [11]. □

3.2 Spatial discretization

In this section, we discretized problem (5), using a hybrid scheme. First, we define a special type of Shishkin mesh to discretize the domain $\overline{\Omega}_x$, then the required scheme can be constructed.

3.2.1 Shishkin mesh

For $N \geq 2^k$, for integer $k \geq 2$, the fitted mesh $\overline{\Omega}_x^N$ is constructed by dividing the domain $\overline{\Omega}_x = [0,1]$ in to three subintervals such that

$$\overline{\Omega}_x^N = [0,\sigma] \cup [\sigma, 1-\sigma] \cup [1-\sigma, 1].$$

Consequently, the mesh points and the transition parameter σ are defined as described in [14, 26]. Let, $L = L(N)$ satisfying,

$$\ln(\ln N) \leq L \leq lnN \quad \text{and} \quad \exp(-L) \leq \frac{L}{N}.$$

Then, the location of the transition point is

$$\sigma = \min\left\{\frac{1}{4}, \eta\sqrt{\varepsilon}L\right\}, \tag{10}$$

where $\eta \geq 4/\sqrt{\alpha}$. Moreover, mesh points in the spatial direction is given by

$$x_i = \begin{cases} \frac{4\sigma}{N} i, & \text{for } i = 1, 2, ..., \frac{N}{4}, \\ \sigma + r(\frac{i}{N} - 1/4)^3 + 4\sigma(\frac{i}{N} - 1/4), & \text{for } \frac{N}{4}+1, \frac{N}{4}+2, ..., \frac{N}{2} \\ 1 - x_{N-i}, & \text{for } i = \frac{N}{2}+1, \frac{N}{2}+2, ..., N. \end{cases} \tag{11}$$

where the coefficient r is determined from $x_{N/2} = 1/2$.
Let $h_{\max} = \max_{\forall i} h_i$, where $h_i = x_i - x_{i-1}, i = 1, 2, ..., N$, from (11) it is clear that the maximum mesh width h_{max} always correspond to $h_{\frac{N}{2}}$ and $h_{\frac{N}{2}+1}$ of the domain $\overline{\Omega}_N$, that is,

$$h_{\max} = h_{\frac{N}{2}} = h_{\frac{N}{2}+1} \leq CN^{-1}.$$

In the coarse part $[\sigma, 1-\sigma]$, the mesh width satisfies the following,

$$h_{i+1} \leq CN^{-1}, \tag{12}$$

$$|h_{i+1} - h_i| \leq CN^{-2}, \tag{13}$$

for more detail see [26].

3.2.2 Derivation of the scheme

We use a hybrid scheme which is a combination of the central difference scheme and the fourth-order compact difference scheme to approximate the semi-discrete problem (5), where the coefficients are determined to make the

scheme exact for polynomials up to degree four and satisfy the normalization condition, $q_i^- + q_i^c + q_i^+ = 1$, for $i = 1, 2, ..., N-1$.

$$\mathcal{L}_\varepsilon^N \widetilde{U}_i^{j+1} \equiv r_{i,j}^- \widetilde{U}_{i-1}^{j+1} + r_{i,j}^c \widetilde{U}_i^{j+1} + r_{i,j}^+ \widetilde{U}_{i+1}^{j+1} = Q(g_i^j) = q_i^- g_{i-1}^j + q_i^c g_i^j + q_i^+ g_{i+1}^j, \tag{14}$$

where

$$r_{i,j}^- = \frac{\Delta t}{2}\widetilde{r}_{i,j}^- + q_i^-, \quad r_{i,j}^c = \frac{\Delta t}{2}\widetilde{r}_{i,j}^c + q_i^c, \quad r_{i,j}^+ = \frac{\Delta t}{2}\widetilde{r}_{i,j}^+ + q_i^+,$$

and the coefficients $\widetilde{r}_{i,j}^-$, $\widetilde{r}_{i,j}^c$, $\widetilde{r}_{i,j}^+$,q_i^-, q_i^c and q_i^+ will be determined later. The development of the method is based on computing the local truncation error as follows:

$$\begin{aligned} T_{i,\widetilde{u}^{j+1}} &= \mathcal{L}_\varepsilon^N(\widetilde{u}_i^{j+1}) - \mathcal{L}_\varepsilon^N(\widetilde{U}_i^{j+1}) \\ &= \mathcal{L}_\varepsilon^N(\widetilde{u}_i^{j+1}) - Q\left((I + \frac{\Delta t}{2}\mathcal{L}_{\varepsilon,x})\widetilde{u}_i^{j+1}\right) \\ &= r_{i,j}^- \widetilde{u}_{i-1}^{j+1} + r_{i,j}^c \widetilde{u}_i^{j+1} + r_{i,j}^+ \widetilde{u}_{i+1}^{j+1} - q_i^-(\widetilde{u}_{i-1}^{j+1} + \frac{\Delta t}{2}\mathcal{L}_{\varepsilon,x}\widetilde{u}_{i-1}^{j+1}) \\ &- q_i^c(\widetilde{u}_i^{j+1} + \frac{\Delta t}{2}\mathcal{L}_{\varepsilon,x}\widetilde{u}_i^{j+1}) - q_i^+(\widetilde{u}_{i+1}^{j+1} + \frac{\Delta t}{2}\mathcal{L}_{\varepsilon,x}\widetilde{u}_{i+1}^{j+1}). \end{aligned} \tag{15}$$

Let's denote the local truncation error corresponds to the discretization of the spatial variable by $T_{i,\widetilde{u}}^x$, and $T_{i,\widetilde{u}^{j+1}}$ represent, $T_{i,\widetilde{u}^{j+1}} = \frac{\Delta t}{2} T_{i,\widetilde{u}}^x$

$$T_{i,\widetilde{u}}^x = T_i^0 \widetilde{u} + T_i^1 \widetilde{u}' + ... + T_i^6 \widetilde{u}^6 + O(h_{\max}^5), \tag{16}$$

where

$$\begin{aligned} T_i^0 &= (\widetilde{r}_{i,j}^- + \widetilde{r}_{i,j}^c + \widetilde{r}_{i,j}^+) - (q_i^- a_{i-1}^{j+1} + q_i^c a_i^{j+1} + q_i^+ a_{i+1}^{j+1}), \\ T_i^1 &= (h_{i+1}\widetilde{r}_{i,j}^+ - h_i \widetilde{r}_{i,j}^-) - (h_{i+1} q_i^+ a_{i+1}^{j+1} - h_i q_i^- a_{i-1}^{j+1}), \\ T_i^2 &= (h_{i+1}^2 \widetilde{r}_{i,j}^+ + h_i^2 \widetilde{r}_{i,j}^-) - 2\varepsilon - (h_{i+1}^2 q_i^+ a_{i+1}^{j+1} + h_i^2 q_i^- a_{i-1}^{j+1}), \\ &\vdots \\ T_i^k &= \frac{h_{i+1}^k}{k!}\widetilde{r}_{i,j}^+ + (-1)^k \frac{h_i^k}{k!}\widetilde{r}_{i,j}^- + q_i^+\left(\frac{h_{i+1}^{k-2}}{(k-2)!}\varepsilon - \frac{h_{i+1}^k}{k!}a_{i+1}^{j+1}\right) \\ &+ (-1)^k q_i^-\left(\frac{h_i^{k-2}}{(k-2)!}\varepsilon - \frac{h_i^k}{k!}a_{i-1}^{j+1}\right), \text{ for } k = 3, 4, 5, 6. \end{aligned}$$

The truncation error is said to be of order p if $T_{i,\widetilde{u}}^x = O(h_{\max}^p)$ as $h_{\max} \to 0$, for $i = 1, 2, ..., N-1$. The method is constructed under the conditions,

$$\begin{aligned} T_i^k &= 0, \quad k = 0, 1, 2, \\ T_i^k &= O(h_{\max}^4), \quad k = 3, 4. \end{aligned} \tag{17}$$

Therefore, the local truncation error $T_{i,\widetilde{u}}^x$ can be written in the form,

$$\begin{aligned} T_{i,\widetilde{u}}^x = T_i^3 \widetilde{u}''' + T_i^4 \widetilde{u}^4 + \left(\frac{1}{5!}(h_{i+1}^5 \widetilde{r}_{i,j}^+ - h_i^5 \widetilde{r}_{i,j}^-) + \frac{\varepsilon}{3!}(h_{i+1}^3 q_i^+ - h_i^3 q_i^-) \right) \widetilde{u}^5 \\ + \left(\frac{1}{6!}(h_{i+1}^6 \widetilde{r}_{i,j}^+ + h_i^6 \widetilde{r}_{i,j}^-) + \frac{\varepsilon}{4!}(h_{i+1}^4 q_i^+ + h_i^4 q_i^-) \right) \widetilde{u}^6. \end{aligned} \tag{18}$$

As a result, to fulfill the condition stated in (17), the coefficients are determined as follows:

$$\begin{aligned} \widetilde{r}_{i,j}^- &= \left(\frac{-2\varepsilon}{h_i(h_i + h_{i+1})} + q_i^- a_{i-1}^{j+1} \right), \\ \widetilde{r}_{i,j}^+ &= \left(\frac{-2\varepsilon}{h_{i+1}(h_i + h_{i+1})} + q_i^+ a_{i-1}^{j+1} \right), \\ \widetilde{r}_{i,j}^c &= \left(\frac{2\varepsilon}{h_i h_{i+1}} + q_i^c a_i^{j+1} \right), \quad 1 \le i \le N-1. \end{aligned} \tag{19}$$

The coefficients q_i^+, q_i^c and q_i^-, $i = 1, 2, ..., N-1$ are defined in two different ways,
(i) For the inner layer region, i.e $(0, \sigma) \cup (1-\sigma, 1)$, the coefficients are given by

$$\begin{aligned} q_i^- &= \frac{1}{6} - \frac{h_{i+1}^2}{6h_i(h_i + h_{i+1})}, \\ q_i^+ &= \frac{1}{6} - \frac{h_i^2}{6h_{i+1}(h_i + h_{i+1})}, \\ q_i^c &= \frac{h_i^2 + h_{i+1}^2 + 3h_i h_{i+1}}{6h_i h_{i+1}} \end{aligned} \tag{20}$$

(ii) For the outer layer region $[\sigma, 1-\sigma]$, depending on the maximum step length(h_{max}) and ε, the coefficients are defined in two different cases. Define $\widetilde{a} = (a + \frac{2}{\Delta t})$, γ is a positive constant independent of ε and Δt,
Case-1 When $\gamma h_{\max}^2 ||\widetilde{a}|| \le \varepsilon$, the coefficients q_i^+, q_i^c and q_i^-, are defined by (20)
Case-2 When $\gamma h_{\max}^2 ||\widetilde{a}|| \ge \varepsilon$, the coefficients q_i^+, q_i^c and q_i^-, are given by

$$q_i^- = 0, \quad q_i^c = 1, \quad q_i^+ = 0. \tag{21}$$

Remark 1. The coefficients q_i^+, q_i^c and q_i^- are determined by the fourth order compact difference scheme in the boundary layer region, and in the regular region when $\gamma h_{\max}^2 ||\widetilde{a}|| \leq \varepsilon$. While in the regular region these coefficients are determined by central difference scheme when $\gamma h_{\max}^2 ||\widetilde{a}|| \geq \varepsilon$. For more detail see [3, 14, 8, 5, 24, 25].

3.2.3 Stability and error analysis

To estimate the $\varepsilon-$ uniform convergence of the proposed scheme, we decompose the approximate solution of (14) into regular and singular components,

$$\widetilde{U} = \widetilde{V} + \widetilde{W}.$$

Now, the truncation error of the method (14) can be decomposed as

$$|T_{i,\widetilde{u}^{j+1}}| \leq |T_{i,\widetilde{v}^{j+1}}| + |T_{i,\widetilde{w}^{j+1}}|, \tag{22}$$

where $|T_{i,\widetilde{v}^{j+1}}|$ and $|T_{i,\widetilde{w}^{j+1}}|$ are the errors corresponding to the regular component $\widetilde{v}^{j+1}$ and the singular component $\widetilde{w}^{j+1}$, respectively.

Lemma 5. Let $N \geq N_0$, where N_0 is the smallest positive integer such that

$$\frac{4\eta^2}{3}(||a|| + 2/\Delta t) < \frac{N_0^2}{L^{*2}}, \tag{23}$$

where $L^* = L(N_0)$. Then for any $N \geq N_0$, the coefficients

$$r_{i,j}^- < 0, \quad r_{i,j}^c > 0, \quad r_{i,j}^+ < 0 \quad \text{and} \quad r_{i,j}^- + r_{i,j}^c + r_{i,j}^+ \geq 0, \quad \text{for } i = 1, 2, ..., N-1.$$

Proof. From (20) and (21) it is clear that $q_i^- \geq 0$, $q_i^c > 0$, and $q_i^+ \geq 0$, then to show $r_{i,j}^- < 0$, $r_{i,j}^+ < 0$ and $r_{i,j}^c > 0$, we have

$$\begin{aligned}
r_{i,j}^- &= \frac{\Delta t}{2}\left(\frac{-2\varepsilon}{h_i(h_i + h_{i+1})} + q_i^-\left(a_{i-1}^{j+1} + \frac{2}{\Delta t}\right)\right),\\
r_{i,j}^+ &= \frac{\Delta t}{2}\left(\frac{-2\varepsilon}{h_{i+1}(h_i + h_{i+1})} + q_i^+\left(a_{i+1}^{j+1} + \frac{2}{\Delta t}\right)\right),\\
r_{i,j}^c &= \frac{\Delta t}{2}\left(\frac{2\varepsilon}{h_i h_{i+1}} + q_i^c\left(a_i^{j+1} + \frac{2}{\Delta t}\right)\right).
\end{aligned} \tag{24}$$

Using (20) and (21), the mesh width and the condition (23) proves $r_{i,j}^- < 0$, $r_{i,j}^+ < 0$ and $r_{i,j}^c > 0$, then from (24), we obtain
$r_{i,j}^- + r_{i,j}^c + r_{i,j}^+ > 0, \quad$ for $i = 1, 2, ..., N-1$. □

Remark 2. From Lemma 5 it is clear that the matrix associated with the discrete operator $\mathcal{L}_\varepsilon^N$ defined in (14) is an irreducible M-matrix and so has

a positive inverse. The operator $\mathcal{L}_\varepsilon^N$ in (14) satisfies the following maximum principle on $D^{N,M}$. Hence, the scheme (14) is uniformly stable in the maximum norm.

Lemma 6. Let Φ satisfies $\Phi \geq 0$ on $\Gamma^{N,M}$. Then $\mathcal{L}_\varepsilon^N \Phi \geq 0$ on $D^{N,M}$ implies that $\Phi \geq 0$ at each point of $D^{N,M}$.

Proof. The proof follows using Lemma 5 and Remark 2. □

Lemma 7. Let $\widetilde{u}^{n+1}$ be the solution of (5) and $\widetilde{U}^{n+1}$ be the solution of discrete scheme (14). Then the global error satisfies

$$||\widetilde{u}^{n+1}(x_i) - \widetilde{U}_i^{n+1}|| \leq C\Delta t(L/N)^4. \tag{25}$$

Proof. For uniform mesh i.e $\sigma = 1/4$, classical analysis can be used to prove the convergence of the scheme. So, we only consider the case $\sigma = \sigma_0\sqrt{\varepsilon}L$.
Case-1 Inner region $[0,\sigma] \cup [1-\sigma, 1]$, $h_i = h_{i+1} = \sigma_0\sqrt{\varepsilon}N^{-1}L$, from the truncation error in (18), we have

$$\begin{aligned} T_{i,\widetilde{u}}^x \leq C\Delta t\varepsilon\Bigg(& |h_{i+1} - h_i|(h_{i+1} + h_i)^2 \left|\left| \frac{d^5\widetilde{u}^{n+1}}{dx^5} \right|\right|_{[x_{i-1},x_{i+1}]} \\ & + (h_i^4 + h_{i+1}^4) \left|\left| \frac{d^6\widetilde{u}^{n+1}}{dx^6} \right|\right|_{[x_{i-1},x_{i+1}]} \Bigg). \end{aligned} \tag{26}$$

Using Lemma 3, we can get

$$T_{i,\widetilde{v}}^x \leq C\Delta t(L/N)^4 \quad \text{and} \quad T_{i,\widetilde{w}}^x \leq C\Delta t(L/N)^4 \tag{27}$$

Therefore, (22) gives

$$T_{i,\widetilde{u}}^x \leq C\Delta t(L/N)^4.$$

Case-2 Outer region $[\sigma, 1-\sigma]$, determined by the relation between $h_{\max}$ and ε
(i) When $\gamma h_{\max}^2 ||\widetilde{a}|| \leq \varepsilon$

$$\begin{aligned} T_{i,\widetilde{u}}^x \leq C\Delta t\varepsilon\Bigg(& |h_{i+1} - h_i|(h_{i+1} + h_i)^2 \left|\left| \frac{d^5\widetilde{u}^{n+1}}{dx^5} \right|\right|_{[x_{i-1},x_{i+1}]} \\ & + (h_i^4 + h_{i+1}^4) \left|\left| \frac{d^6\widetilde{u}^{n+1}}{dx^6} \right|\right|_{[x_{i-1},x_{i+1}]} \Bigg). \end{aligned} \tag{28}$$

Using (3), (12) and (13), we obtain the bound of the truncation error with respect to the regular component $\widetilde{v}$ is,

$$T_{i,\widetilde{v}}^x \leq C\Delta t N^{-4}. \tag{29}$$

Similarly for the singular component $\widetilde{w}$ using (4), (12) and (13), we obtain the bound of the truncation error with respect to singular component $\widetilde{w}$,

$$\begin{aligned} T^x_{i,\widetilde{w}} &\leq C\varepsilon^{-2}\Delta t N^{-4}(\exp(-\sqrt{\alpha}x/\sqrt{\varepsilon}) + \exp(-\sqrt{\alpha}(1-x)/\sqrt{\varepsilon})) \\ &\leq C\Delta t(L/N)^4, \end{aligned} \tag{30}$$

since

$$\left(\exp(-\sqrt{\alpha}x/\sqrt{\varepsilon}) + \exp(-\sqrt{\alpha}(1-x)/\sqrt{\varepsilon})\right) \leq C\exp(-a(x,t)L),$$

and

$$\exp(-L) \leq L/N.$$

(ii) When $\gamma h^2_{\max}||\widetilde{a}|| \geq \varepsilon$, Assume $L^{-4} \leq C\Delta t$,

$$T^x_{i,\widetilde{u}} \leq C\Delta t\varepsilon\left(|h_{i+1}-h_i|\left|\left|\frac{d^3\widetilde{u}^{n+1}}{dx^3}\right|\right|_{[x_{i-1},x_{i+1}]} + (h_i^2+h_{i+1}^2)\left|\left|\frac{d^4\widetilde{u}^{n+1}}{dx^4}\right|\right|_{[x_{i-1},x_{i+1}]}\right). \tag{31}$$

Using Lemma 3, (12) and (13), we obtain the bound of the truncation error with respect to regular component $\widetilde{v}$ and layer component $\widetilde{w}$ as follows

$$T^x_{i,\widetilde{v}} \leq C\varepsilon\Delta t N^{-2}.$$

Since $\gamma h^2_{\max} > \varepsilon$,

$$T^x_{i,\widetilde{v}} \leq CN^{-4} \tag{32}$$

and

$$\begin{aligned} T^x_{i,\widetilde{w}} \leq C\varepsilon\Delta t\left|\left|\frac{d^2w}{dx^2}\right|\right|_{[x_{i-1},x_{i+1}]} &\leq C\Delta t(\exp(-\sqrt{\alpha}x/\sqrt{\varepsilon}) + \exp(-\sqrt{\alpha}(1-x)/\sqrt{\varepsilon})) \\ &\leq C\Delta t(L/N)^4. \end{aligned} \tag{33}$$

Hence, from (32), (33) and (22), we get

$$T^x_{i,\widetilde{u}} \leq C\varepsilon\Delta t(L/N)^4. \tag{34}$$

Therefore, taking the estimate in case (i) and case (ii) then the truncation error estimate of the scheme (14) is given by

$$T^x_{i,\widetilde{u}} \leq C\varepsilon\Delta t(L/N)^4. \tag{35}$$

Finally, combining the uniform stability result and (35), concludes the proof. □

4 The full discrete scheme

A combination of time and the spatial semi-discretization gives the full discrete scheme. Let U^j_i be the numerical approximation for $u(x_i,t_j)$ then the fully discrete scheme on the mesh $D^{N,M}$ is of the form

$$\begin{cases} U^{-j}(x_i) = \phi_b(x_i, -t_j), \text{ for } j = 0, 1, 2, ..., m_\tau, \text{ and } 0, 1, 2, ..., N, \\ \mathcal{L}_\varepsilon^{N,M} U_i^{j+1} \equiv r_{i,j}^- U_{i-1}^{j+1} + r_{i,j}^c U_i^{j+1} + r_{i,j}^+ U_{i+1}^{j+1} = q_i^- G_{i-1}^j + q_i^c G_i^j + q_i^+ G_{i+1}^j, \\ U_0^{j+1} = \phi_l(t_{j+1}), \quad U_N^{j+1} = \phi_r(t_{j+1}), \\ \text{for} \quad i = 1, 2, ..., N-1 \quad \text{and } j = 1, 2, ..., M-1. \end{cases} \tag{36}$$

Where

$$\mathcal{L}_{\varepsilon,x}^N U_i^{j+1} = -\mathcal{L}_{\varepsilon,x}^N U_i^j - 2\frac{U_i^{j+1} - U_i^j}{\Delta t} + F^j(x_i) + F^{j+1}(x_i),$$
$$G_i^j = U_i^j + \frac{\Delta t}{2}\left[F^j(x_i) + F^{j+1}(x_i) - \mathcal{L}_{\varepsilon,x}^N U_i^j\right],$$

and the coefficients $r_{i,j}^-, r_{i,j}^c$, $r_{i,j}^+$, q_i^-, q_i^c and q_i^+, are described in subsection 3.2.2.

The next theorem is for the $\varepsilon-$ uniform convergence of the proposed scheme (36).

Theorem 2. Let u be the exact solution of (1) and U_i^j be the approximate solution of (36). Under the hypothesis of Lemma (5). Then the global $\varepsilon-$ uniform error estimate of the scheme (36) satisfies the following bound

$$||u(x_i, t_j) - U_i^j||_{\overline{D}^{N,M}} \leq C\left((\Delta t)^2 + (L/N)^4\right), \quad 1 \leq i \leq N \text{ and } 1 \leq j \leq M,$$

where L is as described in section 3.2.1.

Proof. Case-1 For $t \leq \tau$, the delay term $u(x, t-\tau)$ is not dependent on ε, since the right hand side of problem (1) is $-b(x,t)\phi(x,t-\tau) + f(x,t)$. Let $\overline{D}_\tau^{N,M} = \Omega_x^N \times \Omega_\tau^M$, where Ω_τ^M denotes the uniform mesh elements of $[0, \tau]$, then the global error estimate on $\overline{D}_\tau^{N,M}$ can be computed as follows

$$\begin{aligned} ||u(x_i, t_j) - U_i^j||_{\overline{D}_\tau^{N,M}} \leq ||u(x_i, t_j) - \widetilde{u}_i^j||_{\overline{D}_\tau^{N,M}} + ||\widetilde{u}_i^j - \widetilde{U}_i^j||_{\overline{D}_\tau^{N,M}} \\ + ||\widetilde{U}_i^j - U_i^j||_{\overline{D}_\tau^{N,M}}. \end{aligned} \tag{37}$$

The combination of (7) and Lemma 3.2.3 result,

$$||u(x_i, t_j) - U_i^j||_{\overline{D}_\tau^{N,M}} \leq ||\widetilde{U}_i^j - U_i^j||_{\overline{D}_\tau^{N,M}} + C\Delta t\left((\Delta t)^2 + (L/N)^4\right). \tag{38}$$

Taking the stability of the fully discrete method, we conclude that

$$||\widetilde{U}_i^j - U_i^j||_{\overline{D}_\tau^{N,M}} \leq ||u(x_i, t_{j-1}) - U_i^{j-1}||_{\overline{D}_\tau^{N,M}}. \tag{39}$$

Now from (38) and (39) results a recurrence relation for the global error, so that

$$||u(x_i,t_j)-U_i^j||_{\overline{D}^N} \leq C\Big((\Delta t)^2+(L/N)^4\Big),\ \ 1\leq i\leq N \text{ and } 1\leq j\leq M. \quad (40)$$

Case-2 When $t\geq\tau$, in this case the delay term $u(x,t-\tau)$ depends on ε. The proof is done following the approach given in [1, 25]. The solution $U(x_i,t_j)$ found on $\overline{D}_{\tau}^{N,M}$ is denoted by $U_\tau(x_i,t_j)$.To determine the estimate of the global error on $\overline{D}_{2\tau}^{N,M}=\Omega^N\times\Omega_{2\tau}^M$, where $\Omega_{2\tau}^M$ denotes the uniform mesh with m_τ mesh elements used on $[\tau,2\tau]$,

$$\begin{aligned}|\mathcal{L}_\varepsilon^{N,M}(u(x_i,t_j)-U_i^j)| &= q_i^- b_{i-1}^j\big(U_\tau(x_{i-1},t_{j-m_\tau})-u(x_{i-1},t_{j-m_\tau})\big)+q_i^- E_j\\ &\quad + q_i^c b_i^j\big(U_\tau(x_i,t_{j-m_\tau})-u(x_i,t_{j-m_\tau})\big)+q_i^c E_j\\ &\quad + q_i^+ b_{i+1}^j\big(U_\tau(x_{i+1},t_{j-m_\tau})-u(x_{i+1},t_{j-m_\tau})\big)+q_i^+ E_j\\ &\quad + T_{i,\widetilde{u}^j},\end{aligned} \quad (41)$$

where E_j and $T_{i,\widetilde{u}^j}$ are the local error associated to the Crank-Nicolson method and the method in (14) respectively. Then using the bound (40),

$$\begin{aligned}|\mathcal{L}_\varepsilon^{N,M}(u(x_i,t_j)-U_i^j)| &\leq ||E_j||+|T_{i,\widetilde{u}^j}|+||u(x_i,t_j)-U_i^j||_{\overline{D}^N}\\ &\leq C\Big((\Delta t)^2+(L/N)^4\Big),\\ &\quad 1\leq i\leq N \text{ and } 1\leq j\leq M.\end{aligned}$$

By introducing barrier functions and applying the discrete maximum principle over $\overline{D}_{2\tau}^{N,M}$, we found

$$||u(x_i,t_j)-U_i^j||_{\overline{D}_{2\tau}^{N,M}} \leq C\Big((\Delta t)^2+(L/N)^4\Big),\ \ 1\leq i\leq N \text{ and } 1\leq j\leq M,$$

Lastly, by applying an induction argument we can get the required estimate. □

5 Numerical results and discussion

To test the performance of the proposed scheme, we applied it to three different problems of the form (1). The maximum absolute errors (point-wise) and the rate of the convergence are calculated and tabulated. For those problems whose exact solution is unknown, the maximum absolute error is calculated using the formula:

$$E_\varepsilon^{N,M}=\max_{0\leq i,j\leq N,M}|U^{N,M}(x_i,t_j)-U^{2N,2M}(x_{2i},t_{2j})|.$$

where $U^{N,M}(x_i,t_j)$ denote the numerical solution obtained by N mesh intervals in the spatial direction and M mesh intervals in the time direction, such that $M=T/\Delta t$. The corresponding rate of convergence rate is given by the formula:

$$R_{\varepsilon}^{N,M}=\log 2\left(\frac{E_{\varepsilon}^{N,M}}{E_{\varepsilon}^{2N,2M}}\right).$$

Also, the $\varepsilon-$ uniform maximum point -wise error $E^{N,M}$ is given by the formula:

$$E^{N,M}=\max_{\varepsilon}E_{\varepsilon}^{N,M}.$$

and the corresponding $\varepsilon-$uniform rate of convergence $R^{N,M}$ is given by the formula:

$$R^{N,M}=\log 2\left(\frac{E^{N,M}}{E^{2N,2M}}\right).$$

Example 1. Consider the following reaction-diffusion problem [9]:

$$\begin{aligned}
&u_t-\varepsilon u_{xx}+\frac{1}{2}u(x,t)=-2e^{-1}u(x,t-1)+f(x,t), \quad (x,t)\in(0,1)\times(0,2],\\
&u(x,t)=e^{-\left(t+\frac{x}{\sqrt{\varepsilon}}\right)}+e^{-\left(t+\frac{(1-x)}{\sqrt{\varepsilon}}\right)}, \quad (x,t)\in[0,1]\times[-1,0],\\
&u(0,t)=e^{-t}+e^{-\left(t+\frac{1}{\sqrt{\varepsilon}}\right)}, \quad u(1,t)=e^{-\left(t+\frac{1}{\sqrt{\varepsilon}}\right)}+e^{-t}, \ t\in[0,2].
\end{aligned}$$

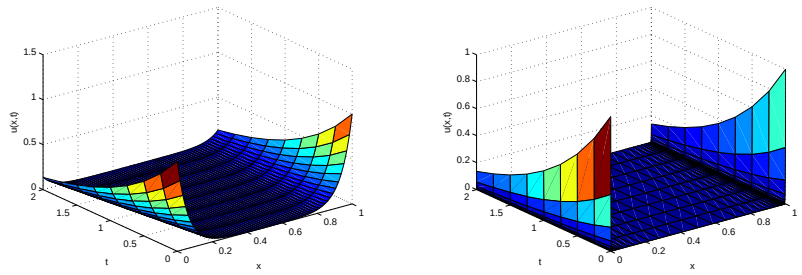
The exact solution of problem (1) is $u(x,t)=e^{-\left(t+\frac{x}{\sqrt{\varepsilon}}\right)}+e^{-\left(t+\frac{(1-x)}{\sqrt{\varepsilon}}\right)}$.

Table 1: Computed $E^{N,\Delta t}$ and $R^{N,\Delta t}$ for Example 1

$\varepsilon\downarrow$	$N\to$64	128	256	512	1024
	$\Delta t\to 0.25$	$0.25/2$	$0.25/2^2$	$0.25/2^3$	$0.25/2^4$
2^0	7.4762e-04	1.5843e-5	3.7666e-5	9.3469e-06	2.3224e-06
	2.2385	2.0725	2.0107	2.0027	
2^{-2}	1.0284e-03	2.4731e-04	6.1734e-5	1.5411e-5	3.8514e-6
	2.0560	2.0022	2.0021	2.0005	
2^{-4}	7.8737e-4	1.9537e-4	4.8763e-5	1.2199e-5	3.0495e-6
	2.0108	2.0024	1.9990	2.0001	
2^{-6}	5.6837e-4	1.4103e-4	3.5236e-5	8.8022e-6	2.2004e-6
	2.0108	2.0009	2.0011	2.0001	
2^{-8}	5.6713e-4	1.4042e-4	3.5088e-5	8.7639e-6	2.1905e-6
	2.0139	2.0007	2.0013	2.0003	
2^{-10}	5.8228e-04	1.4152e-04	3.5095e-05	8.7677e-06	2.1907e-06
	2.0407	2.0117	2.0010	2.0008	
	⋮	⋮	⋮	⋮	⋮
2^{-20}	5.8228e-04	1.4152e-04	3.5095e-05	8.7677e-06	2.1907e-06
	2.0407	2.0117	2.0010	2.0008	
$\mathbf{E^{N,t}}$	**5.8228e-04**	**1.4152e-04**	**3.5095e-05**	**8.7677e-06**	**2.1907e-06**
$\mathbf{R^{N,t}}$	**2.0407**	**2.0117**	**2.0010**	**2.0008**	

Table 2: Computed $E^{N,\Delta t}$ and $R^{N,\Delta t}$ for Example 1, to show the order of convergence of space variable

$\varepsilon\downarrow$	$N\to$32	64	128	256	512
	$\Delta t\to 0.25$	$0.25/4$	$0.25/4^2$	$0.25/4^3$	$0.25/4^4$
2^0	7.4762e-04	3.7666e-05	2.3324e-06	1.4570e-07	9.1115e-09
	4.3110	4.0134	4.0007	3.9992	
2^{-2}	1.0284e-03	6.1735e05	3.8514e-06	2.4071e-07	1.5047e-08
	4.0582	4.0026	4.0000	3.9998	
2^{-4}	7.8752e-04	4.8772e-05	3.0501e-06	1.9063e-07	1.1915e-08
	4.0132	3.9991	4.0000	3.9999	
2^{-6}	5.7001e-04	3.5282e-05	2.2070e-06	1.3793e-07	8.6217e-09
	4.0140	3.9988	4.0001	3.9999	
2^{-8}	5.9289e-04	3.6752e-05	2.2962e-06	1.4374e-07	8.9838e-09
	4.0119	4.0005	3.9977	4.0000	
2^{-10}	1.0075e-03	6.2749e-05	3.9321e-06	2.4595e-07	1.7804e-08
	4.0050	3.9962	3.9989	3.7881	
2^{-12}	1.1750e-03	7.2942e-05	4.5593e-06	2.8753e-07	1.7959e-08
	4.0098	3.9999	3.9870	4.0009	
2^{-14}	1.1750e-03	7.2942e-05	4.5593e-06	2.8753e-07	1.7959e-08
	4.0098	3.9999	3.9870	4.0009	
2^{-16}	1.1750e-03	7.2942e-05	4.5593e-06	2.8753e-07	1.7959e-08
	4.0098	3.9999	3.9870	4.0009	
$\mathbf{E^{N,t}}$	**1.1750e-03**	**7.2942e-05**	**4.5593e-06**	**2.8753e-07**	**1.7959e-08**
$\mathbf{R^{N,t}}$	**4.0098**	**3.9962**	**3.9870**	**3.7881**	



(a) $\varepsilon = 2^{-8}, N = 64, \Delta t = 0.25$ (b) $\varepsilon = 2^{-16}, N = 64, \Delta t = 0.25$

Figure 1: Surface plot of numerical solution of Example 1

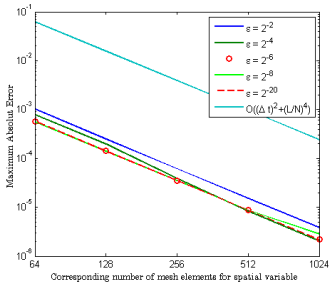


Figure 2: Log-log plot of Example(1)

Example 2. Consider the following reaction-diffusion problem [1, 12, 13]:

$$u_t - \varepsilon u_{xx} + \frac{1+x^2}{2}u(x,t) = -u(x,t-1) + t^3, \quad (x,t) \in (0,1) \times (0,2],$$

with

$$u(x,t) = 0, (x,t) \in [0,1] \times [-1,0], \quad u(0,t) = 0, u(1,t) = 0, \quad t \in (0,2].$$

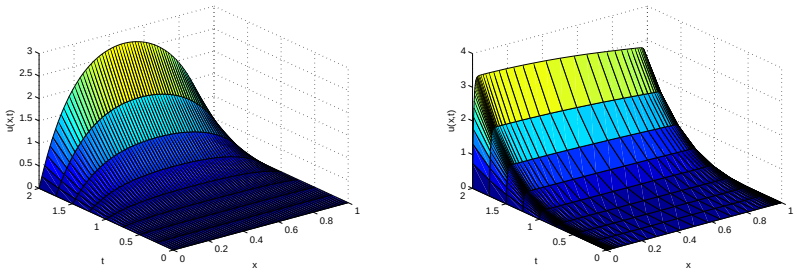
Table 3: Computed $E^{N,\Delta t}$ and $R^{N,\Delta t}$ for Example 2

Method	$\varepsilon \downarrow$	$N \rightarrow$**128**	**256**	**512**	**1024**	**2048**
		t $\rightarrow$ **0.25**	**0.25/2**	**0.25/2²**	**0.25/2³**	**0.25/2⁴**
present	10^0	2.7387e-04	6.7253e-05	1.6808e-05	4.2019e-06	1.0505e-06
		2.0258	20004	2.0001	2.0000	
	10^{-2}	1.6900e-02	4.2067e-03	1.0448e-03	2.6493e-04	9.2519e-05
		2.0062	2.0095	1.9795	1.5178	
	10^{-4}	1.9293e-02	4.8267e-03	1.2067e-03	3.0157e-04	9.3848e-05
		1.9989	2.0000	2.0005	1.6841	
	10^{-6}	1.9357e-02	4.8428e-03	1.2067e-03	3.0275e-04	9.3848e-05
		1.9989	2.0048	1.9999	1.6895	
	10^{-8}	1.9358e-02	4.8432e-03	1.2111e-03	3.0280e-04	9.3833e-05
		1.9989	1.9998	1.9999	1.6902	
	10^{-10}	1.9358e-02	4.8432e-03	1.2111e-03	3.0280e-04	9.3833e-05
		1.9989	1.9998	1.9999	1.6902	
	10^{-12}	1.9358e-02	4.8432e-03	1.2111e-03	3.0280e-04	9.3833e-05
		1.9989	1.9998	1.9999	1.6902	
	$\mathbf{E^{N,t}}$	**1.9358e-02**	**4.8432e-03**	**1.2111e-03**	**3.0280e-04**	**9.3833e-05**
	$\mathbf{R^{N,t}}$	**1.9989**	**1.9998**	**1.9999**	**1.6902**	
[12]	$\mathbf{E^{N,t}}$	2.39e-01	1.28e-01	6.60e-02	3.35e-02	1.71e-02
[13]	$\mathbf{E^{N,t}}$	2.78e-01	1.42e-03	7.13e-02	3.58e-02	1.79e-02

Example 3. Consider the following reaction-diffusion problem [13, 12]:

Table 4: Computed $E^{N,\Delta t}$ and $R^{N,\Delta t}$ for Example 2, to show the order of convergence of space variable

$\varepsilon\downarrow$	$N\rightarrow$**32**	**64**	**128**	**256**	**512**
	t $\rightarrow$ **0.25**	**0.25/4**	**0.25/4^2**	**0.25/4^3**	**0.25/4^4**
10^0	3.3960e-04	2.1010e-05	1.3131e-06	8.2068e-08	5.1299e-09
	4.0147	4.0001	4.0000	3.9998	
10^{-2}	2.1214e-02	1.3267e-03	8.2918e-05	5.1824e-06	3.2390e-07
	3.9991	4.0000	4.0000	4.0000	
10^{-4}	2.4127e-02	1.5092e-03	9.4330e-05	5.8957e-06	3.6848e-07
	3.9988	3.9999	4.0000	4.0000	
10^{-6}	2.4192e-02	1.5136e-03	9.4608e-05	5.9131e-06	3.6957e-07
	3.9984	3.9999	4.0000	4.0000	
10^{-8}	2.4192e-02	1.5137e-03	9.4615e-05	5.9136e-06	3.6960e-07
	3.9984	3.9999	4.0000	4.0000	
10^{-10}	2.4189e-02	1.5137e-03	9.4615e-05	5.9135e-06	3.6970e-07
	3.9982	3.9999	4.0000	3.9996	
E^{N,t}	**2.4192e-02**	**1.5137e-03**	**9.4615e-05**	**5.9136e-06**	**3.6970e-07**
R^{N,t}	**3.9982**	**3.9999**	**4.0000**	**3.9996**	



(a) $\varepsilon = 10^{-2}, N = 64, \Delta t = 0.25$ (b) $\varepsilon = 10^{-4}, N = 64, \Delta t = 0.25$

Figure 3: Surface plot for numerical solution for Example 2

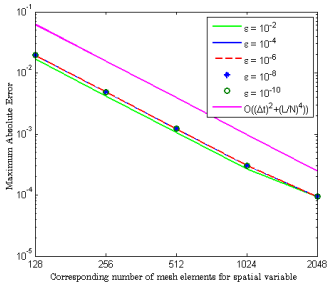


Figure 4: Log-log plot of Example 2

$$u_t - \varepsilon u_{xx} = -2e^{-1}u(x,t-1), \quad (x,t) \in (0,1)\times(0,2],$$

with

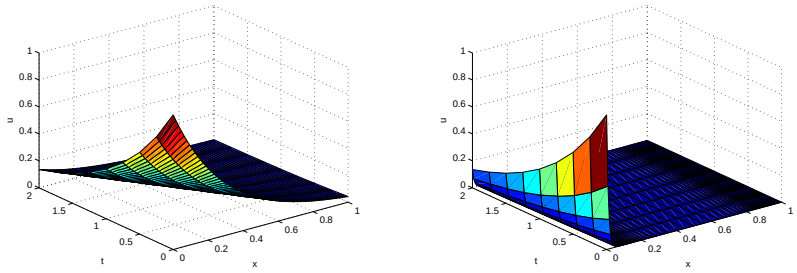
$$\begin{aligned} &u(0,t) = e^{-t}, \quad u(1,t) = e^{-(t+1/\sqrt{\varepsilon})}, \quad t \in (0,2], \\ &u(x,t) = e^{-(t+1/\sqrt{\varepsilon})}, \quad (x,t) \in [0,1]\times[-1,0]. \end{aligned} \tag{42}$$

$u(x,t) = e^{-(t+x/\sqrt{\varepsilon})}$ is the exact solution of the problem.

Table 5: Computed $E^{N,\Delta t}$ and $R^{N,\Delta t}$ for Example 3

Method	$\varepsilon\downarrow$	$N\to$128	256	512	1024	2048
		$\Delta t\to 0.25$	$0.25/2$	$0.25/2^2$	$0.25/2^3$	$0.25/2^4$
present	10^0	3.9168e-05	8.2891e-05	1.9790e-05	4.9090e-06	1.2254e-06
		2.2404	2.0665	2.0112	2.0022	
	10^{-2}	6.7565e-04	1.6837e-04	4.2083e-05	1.0517e-05	2.6290e-06
		2.0047	2.0003	2.0005	2.0001	
	10^{-4}	6.8750e-04	1.6903e-04	4.2110e-05	1.0518e-05	2.6291e-06
		2.0241	2.0051	2.0013	2.0002	
	10^{-6}	6.8750e-04	1.6903e-04	4.2110e-05	1.0518e-05	2.6291e-06
		2.0241	2.0051	2.0013	2.0002	
	10^{-8}	6.8750e-04	1.6903e-04	4.2110e-05	1.0518e-05	2.6291e-06
		2.0241	2.0051	2.0013	2.0002	
	10^{-10}	6.8750e-04	1.6903e-04	4.2110e-05	1.0518e-05	2.6291e-06
		2.0241	2.0051	2.0013	2.0002	
	10^{-12}	6.8750e-04	1.6903e-04	4.2110e-05	1.0518e-05	2.6291e-06
		2.0241	2.0051	2.0013	2.0002	
	$\mathbf{E^{N,t}}$	**6.8750e-04**	**1.6903e-04**	**4.2110e-05**	**1.0518e-05**	**2.6291e-06**
	$\mathbf{R^{N,t}}$	**2.0241**	**2.0051**	**2.0013**	**2.0002**	
[12]	$\mathbf{E^{N,t}}$	9.29e-03	4.43e-03	2.17e-03	1.07e-03	5.32e-04
[13]	$\mathbf{E^{N,t}}$	1.49e-02	7.75e-03	3.95e-03	2.00e-03	1.00e-03

The fitted mesh technique applied in these computations, described in subsection (3.2.1), is uniform in the inner region, whereas in the outer region the mesh gradually condenses more and more from the center to both left and right ends of the interval. In Example 3 because of the boundary values there is no boundary layer on Γ_r, even though we use the same mesh technique for this computation, it is possible to reshuffle the mesh appropriate to each specific example, for further description of this special characteristic, see [1]. In Tables 1 and 2, 3 and 5, for examples 1,2 and 3 respectively, the maximum absolute error and rate of convergence for different values of perturbation parameter ε and mesh numbers are presented, in all the case we observed that the computed $\varepsilon-$ uniform errors $E^{N,\Delta t}$ decreases monotonically as the number of mesh points increases, this ensures that the proposed scheme converges $\varepsilon-$ uniformly. The orders of convergence $R^{N,\Delta t}$ is also independent of the perturbation parameter ε. To visualize the numerical order



(a) $\varepsilon = 10^{-2}, N = 64, \Delta t = 0.25$ (b) $\varepsilon = 10^{-4}, N = 64, \Delta t = 0.25$

Figure 5: Solution graph for Example 3

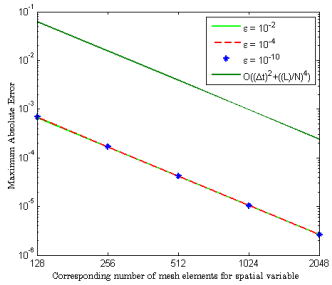


Figure 6: Log-log plot of Example 3

of convergence, the maximum point-wise errors are plotted in log-log scale in figures 2, 4 and 6. The effect of the perturbation parameter on the boundary layer behavior of the solution is also shown in figures 1, 3 and 5. The numerical result displayed in tables 2 and 4 reflects the actual theoretical order of convergence of the spatial variability of the proposed scheme.

6 Conclusion

Singularly perturbed time delay parabolic partial differential equation of reaction-diffusion type of the form (1) is considered. To get an approximate solution to this problem, we used a hybrid scheme which is a combination of a fourth-order compact difference scheme and a central difference scheme based on a special type of Shishkin mesh in space variable, and the Crank-Nicolson method on uniform mesh in the time variable. The proposed scheme is $\varepsilon-$ uniform convergent of order $O((\Delta t)^2 + (L/N)^4)$. Further, numerical experiments are carried out to demonstrate the performance of the proposed scheme. The numerical results clearly show the high accuracy and order of

convergence of the proposed scheme as compared to results available in the literature.

References

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

[2] Babu, G. and Bansal, K. *A high order robust numerical scheme for singularly perturbed delay parabolic convection diffusion problems*, J. Appl. Math. Comput. 68 (2021), 363–389.

[3] Clavero, C. and Gracia, J.L. *High order methods for elliptic and time dependent reaction–diffusion singularly perturbed problems*, Appl. Math. Comput. 168 (2005), 1109–1127.

[4] Clavero, C. and Gracia, J.L. *On the uniform convergence of a finite difference scheme for time dependent singularly perturbed reaction-diffusion problems*, Appl. Math. Comput. 216 (2010), 1478–1488.

[5] Clavero, C. and Gracia, J.L. *A high order hodie finite difference scheme for 1d parabolic singularly perturbed reaction–diffusion problems*, Appl. Math. Comput. 218 (2012), 5067–5080.

[6] Clavero, C., Gracia, J. and Jorge, J. *High-order numerical methods for one-dimensional parabolic singularly perturbed problems with regular layers*, Numer. Methods Partial Differential Equations 21 (2005), 149–169.

[7] Gelu, F.W. and Duressa, G.F. *A uniformly convergent collocation method for singularly perturbed delay parabolic reaction-diffusion problem*, Abstr. Appl. Anal. 2021 (2021).

[8] Govindarao, L. and Mohapatra, J. *A second order numerical method for singularly perturbed delay parabolic partial differential equation*, Eng. Comput.

[9] Govindarao, L., Mohapatra, J. and Das, A. *A fourth-order numerical scheme for singularly perturbed delay parabolic problem arising in population dynamics*, J. Appl. Math. Comput. 63 (2020), 171–195.

[10] Govindarao, L. and Mohapatra, J. *Numerical analysis and simulation of delay parabolic partial differential equation involving a small parameter*, Eng Comput 37 (2019), 289–312.

[11] Kadalbajoo, M.K. and Awasthi, A. *Crank–nicolson finite difference method based on a midpoint upwind scheme on a non-uniform mesh*

for time-dependent singularly perturbed convection–diffusion equations, Int. J. Comput. Math. 85 (2008), 771–790.

[12] Kumar, P.M.M. and Kanth, A.R. *Computational study for a class of time-dependent singularly perturbed parabolic partial differential equation through tension spline*, Comput. Appl. Math. 39 (2020), 1–19.

[13] Kumar, S. and Kumar, M. *High order parameter-uniform discretization for singularly perturbed parabolic partial differential equations with time delay*, Comput. Math. Appl. 68 (2014), 1355–1367.

[14] Kumar, M. and Rao, S.C.S. *High order parameter-robust numerical methodfor time dependent singularly perturbed reaction–diffusion problems*, Computing 90 (2010), 15–38.

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

[16] Mackey, M.C. and Glass, L. *Oscillation and chaos in physiological control systems*, Science 197 (1977), 287–289.

[17] Mallet-Paret, J. and Nussbaum, R.D. *A differential-delay equation arising in optics and physiology*, SIAM J. Math. Anal. 20 (1989), 249–292.

[18] Miller, J. ORiordan, E. and Shishkin, G. *Fitted numerical methods for singular perturbation problems*, rev. edn, 2012.

[19] Mohapatra, J. and Govindarao, L. *A fourth order optimal numerical approximation and its convergence for singularly perturbed time delayed parabolic problems*, Iranian Journal of Numerical Analysis and Optimization (2021).

[20] Negero, N.T. and Duressa, G.F. *Uniform convergent solution of singularly perturbed parabolic differential equations with general temporal-lag*, Iran. J. Sci. Technol. Trans. A Sci. 46 (2022), 507–524.

[21] Priyadarshana, S., Mohapatra, J. and Govindrao, L. *An efficient uniformly convergent numerical scheme for singularly perturbed semilinear parabolic problems with large delay in time*, J. Appl. Math. Comput. 68 (2021), 2617–2639.

[22] Priyadarshana, S., Mohapatra, J. and Pattanaik, S.R.*Parameter uniform optimal order numerical approximations for time-delayed parabolic convection diffusion problems involving two small parameters*, J. Comput. Appl. Math. 41 (2022).

[23] Sahu, S.R. and Mohapatra, J. *Numerical investigation of time delay parabolic differential equation involving two small parameters*, Eng. Comput. 38 (2021), 2882–2899.

[24] Salama, A. and Al-Amery, D. *Asymptotic-numerical method for singularly perturbed differential difference equations of mixed-type*, J. Appl. Math. Inform. 33 (2015), 485–502.

[25] Salama, A. and Al-Amery, D. *A higher order uniformly convergent method for singularly perturbed delay parabolic partial differential equations*, Int. J. Comput. Math. 94 (2017), 2520–2546.

[26] Vulanović, R. *A higher-order scheme for quasilinear boundary value problems with two small parameters*, Computing 67 (2001), 287–303.

[27] Wazewska-Czyzewska, M. and Lasota, A. *Mathematical models of the red cell system*, Matematyta Stosowana 6 (1976), 25–40.

[28] Yadav, S. and Rai, P. *A higher order scheme for singularly perturbed delay parabolic turning point problem*, Eng. Comput. 38 (2020), 819–851.

How to cite this article
Amsalu Ayele, M., Andargie Tiruneh, A. and Adamu Derese, G., Fitted scheme for singularly perturbed time delay reaction-diffusion problems. *Iran. j. numer. anal. optim.*, 2023; 13(2): 262-284. https://doi.org/10.22067/ijnao.2022.76652.1161

Iranian Journal of Numerical Analysis and Optimization
Vol. 13, No. 2, 2023, pp 285–316
https://doi.org/10.22067/ijnao.2022.78723.1181
https://ijnao.um.ac.ir/



How to cite this article
Research Article



Modified Runge–Kutta method with convergence analysis for nonlinear stochastic differential equations with Hölder continuous diffusion coefficient

A. Haghighi

Abstract

The main goal of this work is to develop and analyze an accurate truncated stochastic Runge–Kutta (TSRK2) method to obtain strong numerical solutions of nonlinear one-dimensional stochastic differential equations (SDEs) with continuous Hölder diffusion coefficients. We will establish the strong L^1-convergence theory to the TSRK2 method under the local Lipschitz condition plus the one-sided Lipschitz condition for the drift coefficient and the continuous Hölder condition for the diffusion coefficient at a time T and over a finite time interval $[0, T]$, respectively. We show that the new method can achieve the optimal convergence order at a finite time T compared to the classical Euler–Maruyama method. Finally, numerical examples are given to support the theoretical results and illustrate the validity of the method.

AMS subject classifications (2020): 65C30; 65L07; 65L04

Keywords: Stochastic differential equation; Strong convergence; Truncated methods; Hölder continuous coefficient.

1 Introduction

Consider the scalar stochastic differential equation of Itô type

$$\begin{aligned} dx(\xi) &= a(x(\xi))d\xi + b(x(\xi))dW(\xi), \qquad \xi \in [0, T], \\ x(0) &= x_0 \in \mathbb{R}, \end{aligned} \tag{1}$$

Received 10 September 2022; revised 13 November 2022; accepted 14 November 2022

Amir Haghighi
Department of Mathematics, Faculty of Science, Razi University, Kermanshah 67149, Iran. e-mail: a.haghighi@razi.ac.ir

where $T > 0$ and $a, b : \mathbb{R} \to \mathbb{R}$ are Borel measurable functions and $W(\xi)$ is a one-dimensional Wiener process. We refer the reader to [5, 16, 24, 26] for an overview of stochastic differential equations (SDEs) and their applications. We assume that the drift coefficient a satisfies the one-sided Lipschitz and local Lipschitz conditions and that the diffusion coefficient b satisfies the Hölder continuity condition. Such applications exist in finance, for example, the Aït–Sahalia-type interest rate model [1, 4] and the Cox–Ingersoll–Ross model [6], and in biology, for example, the stochastic SIS epidemic model [7] and laser emission models in chemical physics [16]. Since these equations can be very complicated and analytical solutions are not always possible, numerical methods have become an efficient tool for computing approximate solutions for SDEs. Many numerical methods have been developed for SDEs under the global Lipschitz and linear growth conditions; see, for example, [24, 16, 29]. In particular, derivative-free stochastic Runge–Kutta (SRK) methods for strong approximations have been proposed [3, 27, 8]. However, Hutzenthaler, Jentzen, and Kloeden [13] have shown that the classical Euler–Maruyama (EM) and Milstein methods do not converge strongly to the solution of (1) when the global Lipschitz and linear growth conditions of the drift or diffusion coefficients are perturbed.

In recent years, an increasing number of numerical methods have been developed for solving nonlinear SDEs without global Lipschitz conditions. These methods include implicit methods [25, 23, 9], the tamed numerical methods, the first of which was presented in Hutzenthaler et al. [14, 28, 12], and the stopped EM method [19]. However, the use of implicit or drift-implicit numerical methods requires the solution of a nonlinear algebraic equation at each time step and thus can be very inefficient. Moreover, the tamed methods may lead to inaccurate results due to the perturbation of the flow by changing the drift and diffusion coefficients even at moderately small step sizes [30]. Recently, Mao [21] proposed the truncated EM method for a strong approximation of the nonlinear SDEs under the local Lipschitz condition and the Khasminskii-type condition. After that, the L^q-convergence rates and stability properties of the truncated EM method have been studied by some researchers [22, 11]. Yang et al. [32] investigated the strong convergence of the truncated EM method for one-dimensional SDEs with superlinearly growing drifts and the Hölder continuous diffusion conditions. Then, several new techniques of the partially truncated EM method were proposed in [33] to determine the optimal convergence rate. The authors also investigated the stability of these methods.

Despite the strength of explicit methods in terms of computational cost, there is still a drawback. As mentioned in [16, 2], explicit methods may have to use very small step sizes if the SDEs to be solved are stiff. Although under the classical global Lipschitz condition, there are some classes of explicit methods with extended stability regions that are well suited for solving stiff problems, especially those whose eigenvalues are close to the negative real axis. For example, Komori and Burrage [17] have developed strong first-

order SROCK methods for Itô and Stratonovich SDEs. However, they are still far from accurate numerical algorithms suitable for highly nonlinear stiff SDEs with explicit methods.

In this paper, we will bring all these ideas together. Based on Mao's truncation strategies [21, 22, 18], we derive a stochastic two-stage truncated Runge–Kutta method for nonlinear SDEs with the superlinearly growing drift coefficient and the continuous Hölder diffusion coefficient. The proposed scheme is explicit and includes some free parameters that can extend the accuracy of the results and stability regions. To the best of the author's knowledge, truncated SRK is the only truncated Runge–Kutta method with the relevant results for a strong approximation of solutions of SDEs with Hölder diffusion coefficients. We will study the strong convergence of the proposed method under the local Lipschitz condition plus the one-sided Lipschitz condition for the drift term and the continuous Hölder condition for the diffusion term at a time T and over a finite time interval $[0, T]$, respectively. We show that the new method can achieve the optimal order of convergence compared to the classical EM method at a finite time T without any restriction on the step size. To show the effectiveness of our methods, we simulate some stiff SDEs with several Hölder parameters $\alpha \in [0, \frac{1}{2})$.

The rest of the paper is organized as follows. In Section 2, we describe some relevant assumptions that must be satisfied for the drift and diffusion coefficients, as well as results for solving the original SDEs. In Section 3, we first present mathematical notations and preliminary results for truncated methods. We then develop the two-stage truncated stochastic Runge–Kutta (TSRK2) method, which is the main goal of the paper. In this section, we present convergence results of the new method at time T and give several technical lemmas. In Section 4, we study the convergence rate over a finite interval $[0, T]$. Numerical results and conclusions are given in Sections 5 and 6, respectively.

2 Mathematical preliminaries

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space with right continuous and increasing filtration $\{\mathcal{F}_\xi\}_{0\le\xi\le T}$, where $\mathcal{F}_0$ contains all $\mathbb{P}$-null sets. Let $\{W(\xi)\}_{0\le\xi\le T}$ be a one-dimensional standard $\{\mathcal{F}_\xi\}_{0\le\xi\le T}$-adapted Wiener process on the probability space. For $z_1, z_2 \in \mathbb{R}$, we use $z_1 \vee z_2 = \max\{z_1, z_2\}$ and $z_1 \wedge z_2 = \min\{z_1, z_2\}$. If B is a set, then its indicator function is denoted by I_B, namely, $I_B(z) = 1$ if $z \in B$ and 0 otherwise. For any fixed $p \in [1, \infty)$, we frequently make use of the Young inequality

$$z_1^p z_2 \le \frac{p\delta}{q} z_1^q + \frac{q-p}{q\delta^{p/q-p}} z_2^{p/q-p}, \qquad \text{for all } z_1, z_2 \in \mathbb{R}^+, \tag{2}$$

for any $\delta > 0$ and $q \in [p, \infty)$. To construct the new method, we now make assumptions about a and b.

Assumption 1. Suppose that there exist real positive constants K_1 and ρ such that

$$|a(z_1) - a(z_2)|^2 \leq K_1(1 + |z_1|^\rho + |z_2|^\rho)|z_1 - z_2|^2, \tag{3}$$

for all $z_1, z_2 \in \mathbb{R}^d$.

From (3) we can conclude that the drift coefficient a satisfies the local Lipschitz condition: For any $u > 0$, there exists $K_u > 0$ such that

$$|a(z_1) - a(z_2)| \leq K_u|z_1 - z_2|, \tag{4}$$

for all $z_1, z_2 \in \mathbb{R}$ with $|z_1| \vee |z_2| \leq u$.

Assumption 2. We assume that the drift coefficient a satisfies the one-sided Lipschitz condition and the diffusion coefficient b satisfies the Hölder continuity condition: Real constants $H_1, H_2 \in \mathbb{R}^+$ and $0 \leq \alpha < 0.5$ exist such that

$$(z_1 - z_2)(a(z_1) - a(z_2)) \leq H_1|z_1 - z_2|^2, \tag{5}$$

$$|b(z_1) - b(z_2)| \leq H_2|z_1 - z_2|^{\frac{1}{2}+\alpha}, \tag{6}$$

for all $z_1, z_2 \in \mathbb{R}$.

Clearly, from Assumption 2, it is easy to verify that

$$za(z) \leq H_1|z|^2 + |z||a(0)| \leq M_1(1 + |z|^2), \tag{7}$$

$$|b(z)| \leq H_2|z|^{0.5+\alpha} + |b(0)| \leq M_2(1 + |z|), \tag{8}$$

for all $z \in \mathbb{R}$, where $M_1 = 0.5\big(|a(0)|^2 \vee [2H_1 + 1]\big)$ and $M_2 = H_2(0.5 + \alpha) \vee \big[H_2(0.5 - \alpha) + |b(0)|\big]$. The relation (8) shows that the coefficient function b satisfies the linear growth condition.

In this work, we use C for the generic positive real constant that depends on p, T, α, x_0, and so on but are independent of the time step size Δ and R, and whose values can change between occurrences.

Remark 1. Under Assumption 2, for all $p \in (2, \infty)$, there is a positive constant C such that

$$za(z) + \frac{p-1}{2}|b(z)|^2 \leq C(1 + |z|^2), \quad \text{for all } z \in \mathbb{R}. \tag{9}$$

To construct the numerical method, we first estimate the growth rate of a as follows. We choose a strictly increasing continuous function $\nu : \mathbb{R}^+ \to \mathbb{R}^+$ such that $\nu(r) \to \infty$ as $r \to \infty$ and

$$\sup_{0<|z_1|\vee|z_2|\le r}\frac{|a(z_1)-a(z_2)|}{|z_1-z_2|}\vee\sup_{|z|\le r}|a(z)|\le\nu(r),\qquad\text{for all } r\ge 1. \tag{10}$$

We see that $\nu^{-1}:[\nu(0),\infty)\to(0,\infty)$ as the inverse function of ν is also a strictly increasing continuous function. From Assumption 1, we observe that

$$\begin{aligned}\sup_{0<|z_1|\vee|z_2|\le r}\frac{|a(z_1)-a(z_2)|}{|z_1-z_2|}&\le\sqrt{K_1}(1+\sqrt{2}r^{\rho/2}),\\ \sup_{|z|\le r}|a(z)|&\le|a(0)|+\sqrt{K_1}r(1+r^{\rho/2}).\end{aligned} \tag{11}$$

Therefore, we can set $\nu(r)=\eta_\nu r^{1+\rho/2}$ for all $r\ge 1$ with $\eta_\nu=\sqrt{2K_1}+|a(0)|$. Yang et al. [32] proved the existence and uniqueness of the strong solution of the scalar SDE (1) with Hölder continuous diffusion coefficients presented in the following theorem. They proved the theorem based on [31, Yamada–Watanabe theorem] and [10, Lemma 3.2].

Theorem 1. Let Assumptions 1 and 2 be satisfied. Then the SDE (1) with initial value $x(0)=x_0\in\mathbb{R}$ has a unique global solution $x(t)$. Moreover, for all $p>0$, there is a positive constant C that depends on T, p, and x_0 such that

$$\mathbb{E}\Big(\sup_{0\le\xi\le T}|x(\xi)|^p\Big)\le C, \tag{12}$$

where $\mathbb{E}$ denotes the probability expectation under the probability measure $\mathbb{P}$.

For any real number $R>|x(0)|$, we define a stopping time

$$\tau_R=\inf\{t\ge 0:|x(t)|\ge R\}. \tag{13}$$

Lemma 1. Let Assumptions 1 and 2 hold. Fix any $p\in(0,\infty)$. Then, for any real number $R>|x(0)|$, we have

$$\mathbb{P}(\tau_R\le T)\le\frac{C}{R^p}, \tag{14}$$

where C stands for the generic positive real constant here in independent of R.

Proof. By replacing ξ by $\tau_R\wedge T$ in (12), we see

$$\mathbb{E}|x(\tau_R\wedge T)|^p\le C. \tag{15}$$

Then, by Markov's inequality, we have

$$R^p\mathbb{P}(\tau_R\le T)\le\mathbb{E}(|x(\tau_R)|^pI_{\tau_R\le T})\le\mathbb{E}(|x(\tau_R\wedge T)|^p)\le C, \tag{16}$$

which completes the proof. □

In the following sections, we consider numerical methods on a uniform mesh $t_n = n\Delta$ for $n = 1, \dots, N$ with $\Delta = T/N$ for some $N \in \mathbb{N}$.

3 An explicit two-stage truncated Runge–Kutta method

In this section, we develop an explicit two-stage truncated Runge–Kutta scheme for the nonlinear SDE (1) with a superlinearly growing drift coefficient and a continuous Hölder diffusion coefficient. It is worth mentioning that in this regard, we adopt the idea of constructing the truncating functions of Mao [22] and Li, Mao, and Yin [18] to construct the new method. First, we outline some notations and preliminary results of the truncated methods, which will be used in the following sections. Further details can be found in the literature [21, 22, 11]. Let $h : (0, 1] \to (0, \infty)$ be a strictly decreasing function such that for a constant $\hat{h} \geq 1$

$$h(1) \geq \nu(1), \quad \Delta^{1/4} h(\Delta) \leq \hat{h}, \quad \lim_{\Delta \to 0^+} h(\Delta) = \infty, \quad \text{for all } \Delta \in (0, 1]. \tag{17}$$

For example, we can consider $h(\Delta) = \eta_h \Delta^{-\epsilon\omega}$ with $\eta_h \geq \eta_\nu$ for any $\epsilon \in (0, 1/4\omega)$, where $\omega > 0$. For a given step size $\Delta \in (0, 1]$, let $\kappa_\Delta : \mathbb{R} \to \mathbb{R}$ denote the truncation mapping defined by $\kappa_\Delta(z) := (\nu^{-1}(h(\Delta)) \wedge |z|)\frac{z}{|z|}$, for all $z \in \mathbb{R}$. We set $\frac{z}{|z|} = 0$ if $z = 0$. It can be easily deduced

$$|\kappa_\Delta(z_1)| \leq |z_1|, \qquad |\kappa_\Delta(z_1) - \kappa_\Delta(z_2)| \leq 2|z_1 - z_2|, \quad \text{for all } z_1, z_2 \in \mathbb{R}^d. \tag{18}$$

Accordingly, we define the truncated coefficient by

$$a_\Delta(z) := a(\kappa_\Delta(z)), \tag{19}$$

for all $z \in \mathbb{R}$. It is obvious from (3), (10) and (18) that

$$|a_\Delta(z)| \leq \nu\big(\nu^{-1}(h(\Delta))\big) = h(\Delta), \qquad \text{for all } z \in \mathbb{R}^d, \tag{20}$$

and

$$|a_\Delta(z_1) - a_\Delta(z_2)|^2 \leq 4K_1\big(1 + |z_1|^\rho + |z_2|^\rho\big)|z_1 - z_2|^2, \tag{21}$$

for all $z_1, z_2 \in \mathbb{R}$.

Remark 2. From (10) and (18) it can be deduced for all $z \in \mathbb{R}$ and $\Delta \in (0, 1]$ that

$$\begin{aligned} |a_\Delta(z)| &\leq |a_\Delta(z) - a(0)| + |a(0)| \leq \nu(\nu^{-1}(h(\Delta))|\kappa_\Delta(z)| + |a(0)| \\ &\leq Ch(\Delta)(1 + |z|). \end{aligned} \tag{22}$$

Also, from (8) we can easily write

$$|b(z)|^2 \leq C(1+|z|^2). \tag{23}$$

In fact, relations (22) and (23) are the same inequalities (2.5) and (2.6) presented in Remark 2.1 in [33], which play a fundamental role in determining the optimal convergence rate of the partially truncated EM methods.

The truncated function a_Δ preserves the relation (9) in Remark 1 for all $\Delta \in (0,1]$. We describe this in the following lemma.

Lemma 2. Let Assumption 2 holds. Then for all $p>2$, there is a constant K_2 such that for all $\Delta \in (0,1]$

$$za_\Delta(z) + \frac{p-1}{2}|b(z)|^2 \leq K_2(1+|z|^2), \quad \text{for all } z \in \mathbb{R}, \tag{24}$$

where $K_2 = \frac{5}{2}\Big(M_1 \vee \big[\frac{M_1}{\nu^{-1}(h(1))}\big] \vee \big[(p-1)M_2^2\big]\Big)$.

Proof. For any $z \in \mathbb{R}$ with $|z| \leq \nu^{-1}(h(\Delta))$ from (7) and (8), we have

$$\begin{aligned} za_\Delta(z) + \frac{p-1}{2}|b(z)|^2 &= za(z) + \frac{p-1}{2}|b(z)|^2 \\ &\leq \big(M_1 \vee [(p-1)M_2^2]\big)(1+|z|^2). \end{aligned} \tag{25}$$

Furthermore, for any $z \in \mathbb{R}$ with $|z| > \nu^{-1}(h(\Delta))$, we can write

$$\begin{aligned} za_\Delta(z) + \frac{p-1}{2}|b(z)|^2 \leq& \nu^{-1}(h(\Delta))\frac{z}{|z|}a\Big(\nu^{-1}(h(\Delta))\frac{z}{|z|}\Big) + (p-1)M_2^2(1+|z|^2) \\ &+ \big(\frac{|z|}{\nu^{-1}(h(\Delta))} - 1\big)\nu^{-1}(h(\Delta))\frac{z}{|z|}a\Big(\nu^{-1}(h(\Delta))\frac{z}{|z|}\Big) \\ \leq& M_1\frac{|z|}{\nu^{-1}(h(\Delta))}\Big(1+[\nu^{-1}(h(\Delta))]^2)\Big) + (p-1)M_2^2(1+|z|^2) \\ \leq& M_1|z|\Big(\frac{1}{\nu^{-1}(h(1))} + |z|)\Big) + (p-1)M_2^2(1+|z|^2) \\ \leq& \frac{5}{2}\Big(M_1 \vee \big[\frac{M_1}{\nu^{-1}(h(1))}\big] \vee \big[(p-1)M_2^2\big]\Big)(1+|z|^2) \end{aligned} \tag{26}$$

The inequalities (25) and (26) imply the required assertion (24) easily. □

Next, we construct our numerical algorithm TSRK2 to approximate the exact solution (1). For any given step size $\Delta \in (0,1]$, define

$$\begin{cases} Y(t_0) = x_0 \\ Z_\Delta(t_n) = Y_\Delta(t_n) + \Delta\theta a_\Delta(Y_\Delta(t_n)), \\ Y_\Delta(t_{n+1}) = Y_\Delta(t_n) + \Delta\Big(\alpha_1 a_\Delta(Y_\Delta(t_n)) + \alpha_2 a_\Delta(Z_\Delta(t_n))\Big) + b(Z_\Delta(t_n))\Delta W_n, \end{cases} \tag{27}$$

for $n = 0, 1, \ldots, N$, where $t_n = n\Delta$ and $\Delta W_n := W(t_{n+1}) - W(t_n)$. Here $\theta \in \mathbb{R}$ and $\alpha_1, \alpha_2 \in [0, 1]$ with $\alpha_1 + \alpha_2 = 1$ are free parameters of the TSRK2 procedure. As an example, for $\theta = 0$, we obtain the explicit truncated EM method [21]. Thus, the presented class of TSRK2 methods turns out to be a generalization of the truncated EM method. It is worth noting that when the truncated function a_Δ is replaced by a, the TSRK2 method with $\alpha_1 = 0$, $\alpha_2 = 1$, and $\theta = 0.5$ is the SRK scheme DDIRDI1 [3] and reduces to the midpoint rule when applied to an ordinary differential equation.

We now form a continuous-time version of the TSRK2 procedure (27). To do this, we first for any fixed step size $\Delta \in (0, 1]$ set

$$Y_\Delta(\xi) = \sum_{n=0}^{\infty} Y_\Delta(t_n) I_{[t_n, t_{n+1})}(\xi), \quad Z_\Delta(\xi) = \sum_{n=0}^{\infty} Z_\Delta(t_n) I_{[t_n, t_{n+1})}(\xi), \qquad \text{for all } \xi \geq 0.$$

$$y_\Delta(\xi) = Y_\Delta(\xi) + \int_{t_n}^{\xi} \Big(\alpha_1 a_\Delta(Y_\Delta(\zeta)) + \alpha_2 a_\Delta(Z_\Delta(\zeta))\Big) d\zeta + \int_{t_n}^{\xi} b(Z_\Delta(\zeta)) dW(\zeta), \tag{28}$$

or equivalently

$$y_\Delta(\xi) = x_0 + \int_{0}^{\xi} \Big(\alpha_1 a_\Delta(Y_\Delta(\zeta)) + \alpha_2 a_\Delta(Z_\Delta(\zeta))\Big) d\zeta + \int_{0}^{\xi} b(Z_\Delta(\zeta)) dW(\zeta). \tag{29}$$

3.1 Moment bound of the TSRK2 method

In this subsection, we will state a new result showing the uniform boundedness of the solutions of the TSRK2 method (27). First, the following lemma shows how to conclude that the values $Y_\Delta(\xi)$ and $Z_\Delta(\xi)$ are close to $y_\Delta(\xi)$ with respect to the L^p-norm.

Lemma 3. For any $\tilde{p} > 0$ and any $T > 0$, there is a positive constant number $C_{\tilde{p}}$ such that for every step size $\Delta \in (0, 1]$

$$\mathbb{E}|y_\Delta(\xi) - Y_\Delta(\xi)|^{\tilde{p}} \vee \mathbb{E}|y_\Delta(\xi) - Z_\Delta(\xi)|^{\tilde{p}} \leq C_{\tilde{p}} \Delta^{\tilde{p}/2} (h(\Delta))^{\tilde{p}}, \qquad \text{for all } \xi \in [0, T]. \tag{30}$$

Proof. We first prove the lemma for any $\tilde{p} \geq 2$. In this case, for a given $\xi \in [0, T]$, there is an unique $n \geq 0$ such that $t_n \leq \xi < t_{n+1}$. By (28) and from the Hölder inequality, [20, Theorem 7.1], (8), and (20), one can conclude

$$\mathbb{E}|y_\Delta(\xi) - Y_\Delta(\xi)|^{\tilde{p}} \leq 3^{\tilde{p}-1} \Big(\alpha_1 \Delta^{\tilde{p}-1} \mathbb{E} \int_{t_n}^{\xi} \big|a_\Delta(Y_\Delta(\zeta))\big|^{\tilde{p}} d\zeta \tag{31}$$

$$
\begin{aligned}
&+\alpha_2\Delta^{\tilde{p}-1}\mathbb{E}\int_{t_n}^{\xi}\left|a_\Delta(Z_\Delta(\zeta))\right|^{\tilde{p}}d\zeta+\Delta^{\frac{\tilde{p}-2}{2}}\mathbb{E}\int_{t_n}^{\xi}\left|b(Z_\Delta(\zeta))\right|^{\tilde{p}}d\zeta\Big)\\
\leq& 3^{\tilde{p}-1}\Big(\alpha_1\Delta^{\tilde{p}}(h(\Delta))^{\tilde{p}}+\alpha_2\Delta^{\tilde{p}}(h(\Delta))^{\tilde{p}}\\
&+2^{\tilde{p}-1}M_2\Delta^{\tilde{p}/2}\big(1+\mathbb{E}|Z_\Delta(t_n)|^{\tilde{p}}\big)\Big)
\end{aligned}
$$

Moreover, for every $\zeta\in[0,T]$, we can find a unique positive integer k such that $t_k\leq\xi<t_{k+1}$. So, by (20) and (27), we have

$$
\mathbb{E}|Z_\Delta(\zeta)-Y_\Delta(\zeta)|^{\tilde{p}}=\mathbb{E}|Z_\Delta(t_k)-Y_\Delta(t_k)|^{\tilde{p}}=|\theta|^{\tilde{p}}\Delta^{\tilde{p}}\mathbb{E}|a_\Delta(Y_\Delta(t_k))|^{\tilde{p}}\leq|\theta|^{\tilde{p}}\Delta^{\tilde{p}}(h(\Delta))^{\tilde{p}}. \tag{32}
$$

On the other hand, by relations (8), (29), and (32) and the Doob martingale inequality, we have

$$
\begin{aligned}
\mathbb{E}|y_\Delta(\xi)|^{\tilde{p}}\leq&3^{\tilde{p}-1}\Big(|x_0|^{\tilde{p}}+\mathbb{E}\int_0^{\xi}\left|a_\Delta(Y_\Delta(\zeta))\right|^{\tilde{p}}d\zeta+\alpha_2\mathbb{E}\int_0^{\xi}\left|a_\Delta(Z_\Delta(\zeta))\right|^{\tilde{p}}d\zeta\\
&+T^{\frac{\tilde{p}-2}{2}}\mathbb{E}\int_0^{\xi}\left|b(Z_\Delta(\zeta))\right|^{\tilde{p}}d\zeta\Big)\leq C\Big(1+(h(\Delta))^{\tilde{p}}+\int_0^{\xi}\mathbb{E}|Z_\Delta(\zeta)|^{\tilde{p}}d\zeta\Big)\\
\leq&C\Big(1+(h(\Delta))^{\tilde{p}}+\int_0^{\xi}\mathbb{E}|Z_\Delta(\zeta)-Y_\Delta(\zeta)|^{\tilde{p}}d\zeta+\int_0^{\xi}\mathbb{E}|Y_\Delta(\zeta)|^{\tilde{p}}d\zeta\Big).
\end{aligned} \tag{33}
$$

From (32) and (33), we can conclude

$$
\sup_{0\leq u\leq\xi}\mathbb{E}|Y_\Delta(u)|^{\tilde{p}}\leq\sup_{0\leq u\leq\xi}\mathbb{E}|y_\Delta(u)|^{\tilde{p}}\leq C\Big(1+(h(\Delta))^{\tilde{p}}+\int_0^{\xi}\sup_{0\leq u\leq\zeta}\mathbb{E}|Y_\Delta(u)|^{\tilde{p}}d\zeta\Big).
$$

By the Gronwall's inequality, we can deduce

$$
\sup_{0\leq\xi\leq T}\mathbb{E}|Y_\Delta(\xi)|^{\tilde{p}}\leq C\Big(1+(h(\Delta))^{\tilde{p}}\Big). \tag{34}
$$

So, from (32), (34), and (17), we can write

$$
\sup_{0\leq\xi\leq T}\mathbb{E}|Z_\Delta(\xi)|^{\tilde{p}}\leq 2^{\tilde{p}-1}|\theta|^{\tilde{p}}\Delta^{\tilde{p}}(h(\Delta))^{\tilde{p}}+2^{\tilde{p}-1}\sup_{0\leq\xi\leq T}\mathbb{E}|Y_\Delta(\xi)|^{\tilde{p}}\leq C\Big(1+(h(\Delta))^{\tilde{p}}\Big). \tag{35}
$$

By inserting (35) in (31), we have

$$
\mathbb{E}|y_\Delta(\xi)-Y_\Delta(\xi)|^{\tilde{p}}\leq C\Big(\Delta^{\tilde{p}}(h(\Delta))^{\tilde{p}}+\Delta^{\tilde{p}/2}+\Delta^{\tilde{p}/2}(h(\Delta))^{\tilde{p}}\Big)\leq C\Delta^{\tilde{p}/2}(h(\Delta))^{\tilde{p}}. \tag{36}
$$

To prove the lemma for any $\bar{p}\in(0,2)$, fix a number $\tilde{p}\geq 2$. Then by the Hölder inequality, we can write

$$\mathbb{E}|y_\Delta(\xi)-Y_\Delta(\xi)|^{\bar{p}} \le \Big(\mathbb{E}|y_\Delta(\xi)-Y_\Delta(\xi)|^{p}\Big)^{\bar{p}/\tilde{p}} \le \Big(C\Delta^{\tilde{p}/2}(h(\Delta))^{\tilde{p}}\Big)^{\bar{p}/\tilde{p}} = C_{\bar{p}}\Delta^{\bar{p}/2}(h(\Delta))^{\bar{p}}. \tag{37}$$

So, from (32), (36), and (37) the proof of theorem is complete. □

Lemma 4. Let Assumption 1 and 2 hold. Then, we have

$$\sup_{0\le\Delta\le 1}\sup_{0\le\xi\le T}\mathbb{E}|y_\Delta(\xi)|^p \le C, \qquad \text{for all } p>0. \tag{38}$$

Proof. Let us fix $\Delta\in(0,1]$ and $\xi\in[0,T]$. Using the Itô formula from (29) for any $p>2$, we can write

$$\begin{aligned}|y_\Delta(\xi)|^p \le& |x_0|^p + \int_0^\xi p|y_\Delta(\zeta)|^{p-2}y_\Delta(\zeta)b(Z_\Delta(\zeta))dW(\zeta)\\ &+\int_0^\xi p|y_\Delta(\zeta)|^{p-2}\bigg(y_\Delta(\zeta)\Big(\alpha_1 a_\Delta(Y(\zeta))+\alpha_2 a_\Delta(Z_\Delta(\zeta))\Big)\\ &+\frac{p-1}{2}|b(Z_\Delta(\zeta)|^2\bigg)d\zeta,\end{aligned}$$

for all $\xi\in[0,T]$. Since $\alpha_1+\alpha_2=1$, we have

$$\begin{aligned}|y_\Delta(\xi)|^p \le& |x_0|^p+\Gamma_1(\xi)+\Gamma_2(\xi)+\Gamma_3(\xi)\\ &+\int_0^\xi p|y_\Delta(\zeta)|^{p-2}y_\Delta(\zeta)b(Z_\Delta(\zeta))dW(\zeta),\end{aligned} \tag{39}$$

where

$$\begin{aligned}\Gamma_1(\xi) &:= \int_0^\xi p|y_\Delta(\zeta)|^{p-2}\Big(Z_\Delta(\zeta)a_\Delta(Z_\Delta(\zeta))+\frac{p-1}{2}|b(Z_\Delta(\zeta))|^2\Big)d\zeta,\\ \Gamma_2(\xi) &:= \int_0^\xi p|y_\Delta(\zeta)|^{p-2}\big(y_\Delta(\zeta)-Z_\Delta(\zeta)\big)a_\Delta(Z_\Delta(\zeta))d\zeta,\\ \Gamma_3(\xi) &:= \alpha_1\int_0^\xi p|y_\Delta(\zeta)|^{p-2}y_\Delta(\zeta)\Big(a_\Delta(Y_\Delta(\zeta))-a_\Delta(Z_\Delta(\zeta))\Big)d\zeta.\end{aligned} \tag{40}$$

By the facts that $p|y_\Delta(\zeta)|^{p-2}y_\Delta(\zeta)b(Z_\Delta(\zeta))$ is $\mathcal{F}_\zeta-$measurable, we have

$$\mathbb{E}|y_\Delta(\xi)|^p \le \mathbb{E}|x_0|^p+\mathbb{E}(\Gamma_1(\xi))+\mathbb{E}(\Gamma_2(\xi))+\mathbb{E}(\Gamma_3(\xi)). \tag{41}$$

Next, we try to estimate the values $\mathbb{E}(\Gamma_1(\xi))$, $\mathbb{E}(\Gamma_2(\xi))$, and $\mathbb{E}(\Gamma_3(\xi))$ in (41). By the relations (20), (24), and (27) and the special form of Young's inequality (2), which reads

$$z_1^{p-2}z_2 \leq \frac{p-2}{p}z_1^p + \frac{2}{p}z_2^{p/2}, \quad \text{for all } z_1, z_2 \geq 0, \tag{42}$$

we can approximate $\Gamma_1(\xi)$ as follows:

$$\begin{aligned}\mathbb{E}(\Gamma_1(\xi)) \leq & \mathbb{E}\int_0^\xi \big((p-2)|y(\zeta)|^p + 2K_2^{p/2}(1+|Z_\Delta(\zeta)|^2)^{p/2}\big)d\zeta \\ \leq & (2K_2)^{p/2}T + 2^{3p/2-1}K_2^{p/2}T|\theta|^p\Delta^p(h(\Delta))^p \\ & + (p-2)\int_0^\xi \mathbb{E}|y_\Delta(\zeta)|^p d\zeta + 2^{3p/2-1}K_2^{p/2}\int_0^\xi \mathbb{E}|Y_\Delta(\zeta)|^p d\zeta. \end{aligned} \tag{43}$$

For $\Gamma_2(\xi)$, with the help of Young's inequality (42), relations (10) and (20) and Lemma 3, we obtain

$$\begin{aligned}\mathbb{E}(\Gamma_2(\xi)) &\leq (p-2)\int_0^\xi \mathbb{E}|y_\Delta(\zeta)|^p d\zeta + 2(h(\Delta))^{p/2}\int_0^T \mathbb{E}|y_\Delta(\zeta) - Z_\Delta(\zeta)|^{p/2}d\zeta \\ &\leq (p-2)\int_0^\xi \mathbb{E}|y_\Delta(\zeta)|^p d\zeta + 2TC_{p/2}(h(\Delta))^p\Delta^{p/4}. \end{aligned} \tag{44}$$

For $\Gamma_3(\xi)$, the Young inequality as well as (10), (18), and (20) yields

$$\begin{aligned}\mathbb{E}(\Gamma_3(\xi)) &\leq \alpha_1\mathbb{E}\int_0^\xi \Big((p-2)|y_\Delta(\zeta)|^p + 2|y_\Delta(\zeta)|^{p/2}\big|a_\Delta(Y(\zeta)) - a_\Delta(Z_\Delta(\zeta))\big|^{p/2}\Big)d\zeta \\ &\leq \alpha_1(p-1)\int_0^\xi \mathbb{E}|y_\Delta(\zeta)|^p d\zeta + \alpha_1(h(\Delta))^p\int_0^T \big|\kappa_\Delta(Y(\zeta)) - \kappa_\Delta(Z_\Delta(\zeta))\big|^p d\zeta \\ &\leq \alpha_1(p-1)\int_0^\xi \mathbb{E}|y_\Delta(\zeta)|^p d\zeta + 2\alpha_1|\theta|^p T(h(\Delta))^{2p}\Delta^p \\ &\leq (p-1)\alpha_1\int_0^\xi \sup_{0\leq u\leq\zeta}\mathbb{E}|y_\Delta(u)|^p du + 2|\theta|^p\alpha_1 T\hat{h}^2. \end{aligned} \tag{45}$$

Inserting (43)–(45) into (41), we have

$$\mathbb{E}|y_\Delta(\xi)|^p \leq C\Big(1 + \int_0^\xi \sup_{0\leq u\leq\zeta}\mathbb{E}|y_\Delta(u)|^p d\zeta\Big).$$

Therefore, we can write

$$\sup_{0\leq u\leq\xi}\mathbb{E}|y(u)|^p \leq C\Big(1 + \int_0^\xi \sup_{0\leq u\leq\zeta}\mathbb{E}|y(u)|^p d\zeta\Big).$$

It can be deduced from the Gronwall's inequality that

$$\sup_{0\leq u\leq\xi}\mathbb{E}|y(u)|^p \leq C,$$

for a positive constant number C independent of Δ. For $0 < \hat{p} \leq p$, according to the Lyapunov inequality, (38) still holds, which completes the proof of the lemma. □

Remark 3. For every $\xi \in [0, T]$, there is a single integer $n \geq 0$ such that $t_n \leq \xi < t_{n+1}$. From (17) and (27) we can deduce that

$$\begin{aligned}\mathbb{E}|Z_\Delta(\xi)|^p = \mathbb{E}|Z_\Delta(t_n)|^p &\leq 2^{p-1}\Big(\mathbb{E}|Y_\Delta(t_n))|^p + \Delta^p|\theta|^p\mathbb{E}\big|a_\Delta(Y_\Delta(t_n))\big|^p\Big)\\ &\leq 2^{p-1}\Big(\mathbb{E}|Y_\Delta(t_n)|^p + \Delta^{3p/4}|\theta|^p\hat{h}^p\Big).\end{aligned}$$

Therefore, by Lemma 3 we can write

$$\sup_{0\leq\Delta\leq 1}\sup_{0\leq\xi\leq T}\mathbb{E}|Z_\Delta(\xi)|^p \leq C. \tag{46}$$

In addition to (13), for any real number $R > |x(0)|$, we define two other stopping times

$$\theta_R^{(1)} := \inf\{t \geq 0 : |Y_\Delta(t)| \geq R\}, \quad \theta_R^{(2)} := \inf\{t \geq 0 : |Z_\Delta(t)| \geq R\}. \tag{47}$$

Lemma 5. Let Assumption 1 and 2 hold. Fix any $p \in (0, \infty)$. Then, for any real number $R > |x(0)|$, we have

$$\mathbb{P}(\theta_R^{(1)} \leq T) \vee \mathbb{P}(\theta_R^{(2)} \leq T) \leq \frac{C}{R^p},$$

where C stands for the generic positive real constant here in independent of R.

Proof. The proof of this lemma is similar to that of Lemma 1. Namely, replacing ξ by $\theta_R^{(1)} \wedge T$ in (38), we see

$$\mathbb{E}|y_\Delta(\theta_R^{(1)} \wedge T)|^p \leq C.$$

Then, by Markov's inequality we have

$$R^p\mathbb{P}(\theta_R^{(1)} \leq T) \leq \mathbb{E}(|y_\Delta(\theta_R^{(1)})|^p I_{\theta_R^{(1)}\leq T}) \leq \mathbb{E}(|y_\Delta(\theta_R^{(1)} \wedge T)|^p) \leq C.$$

Moreover, it follows from Remark 3 that $\mathbb{E}|Z_\Delta(\theta_R^{(2)} \wedge T)|^p \leq C$. Therefore, as above

$$R^p\mathbb{P}(\theta_R^{(2)} \leq T) \leq \mathbb{E}(|Z_\Delta(\theta_R^{(2)})|^p I_{\theta_R^{(2)}\leq T}) \leq \mathbb{E}(|Z_\Delta(\theta_R^{(2)} \wedge T)|^p) \leq C,$$

which completes the proof. □

3.2 Convergence of the new method at time T

In this section, we study the effectiveness of the method TASK2 in solving problem (1) and obtain the corresponding convergence results at finite time T. Using the method of Yamada and Watanabe, for each $\delta \in [1,\infty)$ and $\epsilon \in (0,\infty)$, we choose a nonnegative continuous function $\chi_{\delta\epsilon} : [0,\infty) \to \mathbb{R}$ such that [15, 31]:

- $\chi_{\delta\epsilon}(\xi) \leq \frac{2}{x\ln(\delta)}$, if $\xi \in [\epsilon/\delta, \epsilon]$, $\quad \chi_{\delta\epsilon}(\xi) = 0$ otherwise;
- $\int_{\epsilon/\delta}^{\epsilon} \chi_{\delta\epsilon}(\xi)d\xi = 1$.

Moreover, we approximate the function $\xi :\to |\xi|$ by the function $\varpi_{\delta\epsilon}$ defined by

$$\varpi_{\delta\epsilon}(\xi) := \int_0^{|\xi|}\int_0^{\zeta_1} \chi_{\delta\epsilon}(\zeta_2)d\zeta_2 d\zeta_1, \qquad \text{for all } \xi \in \mathbb{R}. \tag{48}$$

Below we outline some properties of the $\varpi_{\delta\epsilon}$ function that will be used in the following sections. Further details can be found in the literature [32, 33].

Lemma 6. Let $\delta \in [1,\infty)$ and let $\epsilon \in (0,\infty)$. Then for all $\xi \in \mathbb{R}$

1. $|\xi| \leq \varpi_{\delta\epsilon}(\xi) + \epsilon$,
2. $0 \leq |\varpi'_{\delta\epsilon}(\xi)| \leq 1$,
3. $\varpi''_{\delta\epsilon}(\xi) = \chi_{\delta\epsilon}(|\xi|) \leq \frac{2}{|\xi|\ln\delta} I_{\{\epsilon/\delta \leq |\xi| \leq \epsilon\}}$,
4. $\frac{\varpi'_{\delta\epsilon}(\xi)}{\xi} > 0$, for all $\xi \in \mathbb{R}\setminus\{0\}$.

Remark 4. Since $\sup_{\xi\in\mathbb{R}}|\varpi'_{\delta\epsilon}(\xi)| \leq 1$, under Assumption 2, we can easily conclude

$$\varpi'_{\delta\epsilon}(z_1 - z_2)(a(z_1) - a(z_2)) \leq H_1|z_1 - z_2|, \tag{49}$$

for all $z_1, z_2 \in \mathbb{R}$; see [32].

Define the stopping time $\beta_{\Delta,R} := \tau_R \wedge \theta_R^{(1)} \wedge \theta_R^{(2)}$, where τ_R and $\theta_R^{(i)}$ for $i = 1, 2$, are defined by (13) and (47), respectively. Then, the moment deviation between $x(\xi \wedge \beta_{\Delta,R})$ and $y_\Delta(\xi \wedge \beta_{\Delta,R})$ is estimated as follows.

Lemma 7. Consider the initial problem (1), which satisfies Assumptions 1 and 2. Suppose that $R > |x_0|$ is a real number and that $\Delta \in (0,1]$ is sufficiently small such that $\nu^{(-1)}(h(\Delta)) \geq R$. Then there exists a constant C such that

$$\mathbb{E}|e_\Delta(\xi \wedge \beta_{\Delta,R})| \leq \begin{cases} C\Big(-\frac{1}{\ln(\Delta)} + \Delta^{\frac{1}{4}}(h(\Delta))^{\frac{1}{2}}\Big), & \text{if } \alpha = 0, \\ C\big(\Delta^{\frac{1}{2}}h(\Delta)\big)^{2\alpha}, & \text{if } 0 < \alpha < 0.5, \end{cases} \tag{50}$$

where

$$e_\Delta(\xi) := x(\xi) - y_\Delta(\xi), \qquad \text{for } 0 \leq \xi \leq T.$$

Proof. We note for $0 \leq \zeta \leq \xi \wedge \beta_{\Delta,R}$ that $|Y_\Delta(\zeta)| \leq R$ and $|Z_\Delta(\zeta)| \leq R$. Therefore, we conclude by the condition $\nu^{(-1)}(h(\Delta)) \geq R$ and the definition of the truncation function a_Δ in (19) that

$$a_\Delta(Y_\Delta(\zeta)) = a(Y_\Delta(\zeta)) \qquad \text{and} \qquad a_\Delta(Z_\Delta(\zeta)) = a(Z_\Delta(\zeta)).$$

It therefore follows from the Itô formula that

$$\begin{aligned}\mathbb{E}|e_\Delta(\xi \wedge \beta_{\Delta,R})| \leq & \epsilon + \mathbb{E}\big(\varpi_{\delta\epsilon}(e_\Delta(\xi \wedge \beta_{\Delta,R}))\big) \\ = & \epsilon + \mathbb{E}\int_0^{\xi\wedge\beta_{\Delta,R}} \varpi'_{\delta\epsilon}(e_\Delta(\zeta))\Big[a(x(\zeta)) - \alpha_1 a(Y_\Delta(\zeta)) - \alpha_2 a(Z_\Delta(\zeta))\Big]d\zeta \\ & + \frac{1}{2}\mathbb{E}\int_0^{\xi\wedge\beta_{\Delta,R}} \varpi''_{\delta\epsilon}(e_\Delta(\zeta))\big[b(x(\zeta)) - b(Z_\Delta(\zeta))\big]^2 d\zeta := \epsilon + \Pi_1 + \Pi 2.\end{aligned} \tag{51}$$

In what follows, we attempt to estimate the values Π_1 and Π_2 in (51). Since $\alpha_1 + \alpha_2 = 1$, by re-arranging we get that

$$\begin{aligned}\Pi_1 \leq & \mathbb{E}\int_0^{\xi\wedge\beta_{\Delta,R}} \varpi'_{\delta\epsilon}(e_\Delta(\zeta))\big[a(x(\zeta)) - a(y_\Delta(\zeta))\big]d\zeta \\ & + \mathbb{E}\int_0^{\xi\wedge\beta_{\Delta,R}} \big|\varpi'_{\delta\epsilon}(e_\Delta(\zeta))\big|\big|a(y_\Delta(\zeta)) - a(Y_\Delta(\zeta))\big|d\zeta \\ & + \alpha_2\mathbb{E}\int_0^{\xi\wedge\beta_{\Delta,R}} \big|\varpi'_{\delta\epsilon}(e_\Delta(\zeta))\big|\big|a(Y_\Delta(\zeta)) - a(Z_\Delta(\zeta))\big|d\zeta.\end{aligned}$$

By Assumption 1, Lemma 6, and Remark 4 we can write

$$\begin{aligned}\Pi_1 \leq & H_1\mathbb{E}\int_0^{\xi\wedge\beta_{\Delta,R}} |e_\Delta(\zeta))|d\zeta \\ & + \sqrt{K_1}\mathbb{E}\int_0^{\xi\wedge\beta_{\Delta,R}} \big(1 + |y_\Delta(\zeta)|^\rho + |Y_\Delta(\zeta)|^\rho\big)^{\frac{1}{2}}\big|y_\Delta(\zeta)) - Y_\Delta(\zeta)\big|d\zeta \\ & + \alpha_2\sqrt{K_1}\mathbb{E}\int_0^{\xi\wedge\beta_{\Delta,R}} \big(1 + |Y_\Delta(\zeta)|^\rho + |Z_\Delta(\zeta)|^\rho\big)^{\frac{1}{2}}\big|Y_\Delta(\zeta)) - Z_\Delta(\zeta)\big|d\zeta.\end{aligned}$$

So, by the Hölder inequality, Lemmas 3 and 4, we have

$$\Pi_1 \leq H_1\mathbb{E}\int_0^{\xi\wedge\beta_{\Delta,R}} |e_\Delta(\zeta))|d\zeta + C\Delta^{\frac{1}{2}}h(\Delta). \tag{52}$$

As for Π_2, it follows from Assumption 2 and Lemmas 3 Lemma 6 that

$$\begin{aligned}\Pi_2 \leq & H_2^2\mathbb{E}\int_0^{\xi\wedge\beta_{\Delta,R}} \big(x(\zeta) - Z_\Delta(\zeta)\big)^{1+2\alpha}\frac{1}{|e_\Delta(\zeta)|\ln\delta} I_{\{\epsilon/\delta \leq |e_\Delta(\zeta)| \leq \epsilon\}}d\zeta \\ \leq & \frac{2^{2\alpha}H_2^2\epsilon^{2\alpha}T}{\ln(\delta)} + \frac{2^{2\alpha}H_2^2\delta}{\epsilon\ln(\delta)}\mathbb{E}\int_0^{\xi}\big|y_\Delta(\zeta) - Z_\Delta(\zeta)\big|^{1+2\alpha}d\zeta\end{aligned}$$

$$\leq 2^{2\alpha} H_2^2 T\Big[\frac{\epsilon^{2\alpha}}{\ln(\delta)} + \frac{\delta}{\epsilon \ln(\delta)}\Big]\big(\Delta^{\frac{1}{2}} h(\Delta)\big)^{1+2\alpha}. \tag{53}$$

Substituting (52) and (53) in (51) and by the Gronwall inequality, we have

$$\mathbb{E}|e_\Delta(\xi \wedge \beta_{\Delta,R})| \leq C\Big(\epsilon + \Delta^{\frac{1}{2}} h(\Delta) + \frac{\epsilon^{2\alpha}}{\ln(\delta)} + \frac{\delta}{\epsilon \ln(\delta)}\big(\Delta^{\frac{1}{2}} h(\Delta)\big)^{1+2\alpha}\Big).$$

In the case when $\alpha = 0$, we set $\delta = \Delta^{-\frac{1}{8}}$ and $\epsilon = -\frac{1}{\ln(\Delta)}$, which implies

$$\mathbb{E}|e_\Delta(\xi \wedge \beta_{\Delta,R})| \leq C\Big(-\frac{1}{\ln(\Delta)} + \Delta^{\frac{1}{4}} (h(\Delta))^{\frac{1}{2}}\Big). \tag{54}$$

If we take $\delta = 2$ and $\epsilon = \Delta^{\frac{1}{2}} h(\Delta)$ for the case where $\alpha \in (0, 0.5)$, then we have

$$\mathbb{E}|e_\Delta(\xi \wedge \beta_{\Delta,R})| \leq C\big(\Delta^{\frac{1}{2}} h(\Delta)\big)^{2\alpha}. \tag{55}$$

The inequalities (54) and (55) prove the desired. □

Theorem 2. Let conditions in Assumptions 1 and 2 be fulfilled and let $p > 1$. Let $R_\Delta := \big(\Delta^{\frac{1}{2}} h(\Delta)\big)^{-1/(p-1)}$ for any $\Delta \in (0, 1]$. If there is a positive real number $\Delta^* \in (0, 1]$ such that

$$\nu^{-1}(h(\Delta)) \geq R_\Delta, \quad \text{for all } \Delta \in (0, \Delta^*]. \tag{56}$$

Then, there exists a positive constant C independent of Δ such that

$$\mathbb{E}|e_\Delta(T)| \leq \begin{cases} C\Big(-\frac{1}{\ln(\Delta)} + \Delta^{\frac{1}{4}} (h(\Delta))^{\frac{1}{2}}\Big), & \text{if } \alpha = 0, \\ C\big(\Delta^{\frac{1}{2}} h(\Delta)\big)^{2\alpha}, & \text{if } \alpha \in (0, \frac{1}{2}). \end{cases} \tag{57}$$

Proof. We first divide the left side of (57) as below:

$$\mathbb{E}|e_\Delta(T)| = \mathbb{E}|e_\Delta(T) I_{\{\beta_{\Delta,R} > T\}}| + \mathbb{E}|e_\Delta(T) I_{\{\beta_{\Delta,R} \leq T\}}| \tag{58}$$

By Young's inequality (2), Theorem 1, and Lemmas 4 and 5, we obtain

$$\begin{aligned} \mathbb{E}|e_\Delta(T) I_{\{\beta_{\Delta,R} \leq T\}}| &\leq \frac{\Delta^{\frac{1}{2}} h(\Delta)}{p} \mathbb{E}|e_\Delta(T)|^p + \frac{p-1}{p\big(\Delta^{\frac{1}{2}} h(\Delta)\big)^{1/(p-1)}} \mathbb{P}(\beta_{\Delta,R} \leq T) \\ &\leq C\Delta^{\frac{1}{2}} h(\Delta). \end{aligned} \tag{59}$$

Since $\nu^{-1}(h(\Delta)) \geq R_\Delta$, substituting (59) in (58) by Lemma 7, we obtain the result of the theorem. □

Remark 5. This theorem shows that the TSRK2 method has an order of convergence close to α for $\alpha \in (0, 0.5)$. This is theoretically almost optimal if we recall that the classical EM method has a convergence order of

α. However, the condition (56) could sometimes make the TSRK2 method impractical; see [33] for more details.

In what follows, we use the mathematical techniques developed by Yang and Huang [33] to remove the imposed condition (56). In this context, we replace the condition $\Delta^{1/4}h(\Delta) \leq \hat{h}$ by the more general condition $\Delta^{1/2}h(\Delta) \leq \hat{h}$ for all $\Delta \in [0,1]$. Without limiting generality, for simplicity, we use here $h(\Delta) = \eta_h \Delta^{-\frac{1}{2}}$ with $\eta_h > 0$.

Lemma 8. Let Assumptions 1 and 2 be satisfied. If $h(\Delta) = \eta_h \Delta^{-\frac{1}{2}}$ with $\eta_h > 0$, then for every $\Delta \in (0,1]$ and every $p > 0$, it holds that

$$\sup_{0<\Delta\leq 1} \mathbb{E}\Big[\sup_{0\leq\xi\leq T} |y_\Delta(\xi)|^p\Big] \leq C, \tag{60}$$

and

$$\Big(\sup_{0\leq\xi\leq T} \mathbb{E}|y_\Delta(\xi) - Y_\Delta(\xi)|^p\Big) \vee \Big(\sup_{0\leq\xi\leq T} \mathbb{E}|y_\Delta(\xi) - Z_\Delta(\xi)|^p\Big) \leq C\Delta^{\frac{p}{2}}, \tag{61}$$

for all real positive number T.

Proof. We fix $\Delta \in (0,1]$ and use the same notation as in the proof of Lemma 4. By using the Itô formula, we can write that for any $p > 2$,

$$\begin{aligned}\sup_{0\leq\xi\leq T} |y_\Delta(\xi)|^p \leq & |x_0|^p + \Gamma_1(T) + \Gamma_2(T) + \Gamma_3(T) \\ & + \sup_{0\leq\xi\leq T}\Big|\int_0^\xi p|y_\Delta(\zeta)|^{p-2} y_\Delta(\zeta) b(Z_\Delta(\zeta))dW(\zeta)\Big|. \end{aligned} \tag{62}$$

So, by the Burkholder–Davis-Gundy inequality and the linear growth condition property of b in (8), we have

$$\begin{aligned}\mathbb{E}\Big(\sup_{0\leq\xi\leq T} |y_\Delta(\xi)|^p\Big) \leq & |x_0|^p + \Gamma_1(T) + \Gamma_2(T) + \Gamma_3(T) \\ & + 8pM_2\mathbb{E}\Big(\Big[\sup_{0\leq\xi\leq T} |y_\Delta(\xi)|^p \int_0^T |y_\Delta(\xi)|^{p-2}\big(1 + |Z_\Delta(\xi)|^2\big)d\xi\Big]^{1/2}\Big). \end{aligned} \tag{63}$$

In what follows, we estimate the values $\mathbb{E}(\Gamma_1(T))$, $\mathbb{E}(\Gamma_2(T))$, and $\mathbb{E}(\Gamma_3(T))$ in (63) with the more general condition $h(\Delta) = \eta_h \Delta^{-\frac{1}{2}}$. For $\mathbb{E}(\Gamma_1(\xi))$, we obtain from (43)

$$\mathbb{E}(\Gamma_1(\xi)) \leq C\Big(1 + \int_0^T \mathbb{E}|y_\Delta(\zeta)|^p d\zeta + \int_0^T \mathbb{E}|Y_\Delta(\zeta)|^p d\zeta\Big). \tag{64}$$

As for $\Gamma_2(\xi)$, we obtain by Young's inequality (42) that

$$\begin{aligned}\mathbb{E}(\Gamma_2(\xi)) \leq & (p-2)\int_0^T \mathbb{E}|y_\Delta(\zeta)|^p d\zeta + 2\mathbb{E}\int_0^T |y_\Delta(\zeta)-Z_\Delta(\zeta)|^{\frac{p}{2}}|a_\Delta(Y_\Delta(\zeta))|^{\frac{p}{2}}d\zeta \\ \leq & C\Delta^{-\frac{p}{4}}\int_0^T (\mathbb{E}|y_\Delta(\zeta)-Z_\Delta(\zeta)|^p)^{\frac{1}{2}}(1+\mathbb{E}|Y_\Delta(\zeta)|^p)^{\frac{1}{2}}d\zeta \\ & + (p-2)\int_0^T \mathbb{E}|y_\Delta(\zeta)|^p d\zeta. \end{aligned} \tag{65}$$

Since $(h(\Delta))^{2p}\Delta^p \leq \eta_h^{2p}\hat{h}^p$, following an approach very similar to that used in (45), we can show that

$$\mathbb{E}(\Gamma_3(\xi)) \leq C\big(1+\int_0^T \mathbb{E}|y_\Delta(\zeta)|^p du\big). \tag{66}$$

For any $\zeta \in [0,T]$, there is a unique integer $n \geq 0$ such that $t_n \leq \zeta \leq t_{n+1}$. Then we can write

$$\begin{aligned}\mathbb{E}|Z_\Delta(\zeta)-Y_\Delta(\zeta)|^p &= \mathbb{E}|Z_\Delta(t_n)-Y_\Delta(t_n)|^p = \Delta^p|\theta|^p\mathbb{E}|a_\Delta(Y_\Delta(t_n))|^p \\ &\leq C\Delta^{\frac{p}{2}}(1+\mathbb{E}|Y_\Delta(t_n)|^p) = C\Delta^{\frac{p}{2}}(1+\mathbb{E}|Y_\Delta(\zeta)|^p), \end{aligned} \tag{67}$$

and

$$\begin{aligned}\mathbb{E}|y_\Delta(\zeta)-Y_\Delta(\zeta)|^p =& \mathbb{E}|y_\Delta(\zeta)-Y_\Delta(t_n)|^p \leq C\Delta^{\frac{p}{2}}\Big(\alpha_1\Delta^{\frac{p}{2}}\mathbb{E}\big[|a_\Delta(Y_\Delta(t_n))|^p\big] \\ & + \alpha_2\Delta^{\frac{p}{2}}\mathbb{E}\big[|a_\Delta(Z_\Delta(t_n))|^p\big] + \mathbb{E}\big[|b(Z_\Delta(t_n))|^p\big]\Big). \end{aligned} \tag{68}$$

Therefore, by Remark 2 and relations (67) and (68) we can obtain

$$\begin{aligned}\mathbb{E}|y_\Delta(\zeta)-Y_\Delta(\zeta)|^p &\leq C\Delta^{\frac{p}{2}}\big(1+\mathbb{E}|Y_\Delta(t_n)|^p+\mathbb{E}|Z_\Delta(t_n)|^p\big) \\ &\leq C\Delta^{\frac{p}{2}}(1+\mathbb{E}|Y_\Delta(\zeta)|^p). \end{aligned} \tag{69}$$

From (67) and (69) we can simply conclude

$$\mathbb{E}|y_\Delta(\zeta)-Z_\Delta(\zeta)|^p \leq C\Delta^{\frac{p}{2}}(1+\mathbb{E}|Y_\Delta(\zeta)|^p). \tag{70}$$

Substituting (64)–(66) and (70) into (63) and applying Young's inequality that is

$$z_1 z_2 \leq \frac{z_1^2}{2\delta}+\frac{\delta z_2^2}{2}, \qquad \text{for all } z_1, z_2 \in \mathbb{R}, \quad \text{for all } \delta > 0, \tag{71}$$

with $\delta = 8pM_2$, we can write

$$\mathbb{E}\Big(\sup_{0\leq\xi\leq T}|y_\Delta(\xi)|^p\Big) \leq C\Big(1+\int_0^T \mathbb{E}|y_\Delta(\zeta)|^p d\zeta + \int_0^T \mathbb{E}|Y_\Delta(\zeta)|^p d\zeta\Big)$$

$$
\begin{aligned}
&+\frac{1}{2}\mathbb{E}\Big(\sup_{0\le\xi\le T}|y_\Delta(\xi)|^p\Big)\\
&+32p^2M_2^2\mathbb{E}\Big(\int_0^T|y_\Delta(\xi)|^{p-2}\big(1+|Z_\Delta(\xi)|^2\big)d\xi\Big)\\
\le&\frac{1}{2}\mathbb{E}\Big(\sup_{0\le\xi\le T}|y_\Delta(\xi)|^p\Big)\\
&+C\Big(1+\int_0^T\mathbb{E}|y_\Delta(\zeta)|^p d\zeta+\int_0^T\mathbb{E}|Y_\Delta(\zeta)|^p d\zeta\Big),
\end{aligned}
$$

which implies

$$
\mathbb{E}\Big(\sup_{0\le\xi\le T}|y_\Delta(\xi)|^p\Big)\le C\Big(1+\int_0^T\mathbb{E}\big[\sup_{0\le s\le\xi}|y_\Delta(s)|^p\big]ds\Big).
$$

Then, the Gronwall inequality implies the relation (60) for $p>2$. For $\bar{p}\in(0,2]$, (60) still holds as desired due to the Lyapunov inequality. Moreover, (61) is obtained directly by substituting (60) into (69) and (70) , which completes the proof. □

Theorem 3. Consider the initial problem (1) that satisfies Assumptions 1 and 2. Let $\nu(r)=\eta_h r^{1+\bar{\rho}}$ and $h(\Delta)=\eta_h\Delta^{-\frac{1}{2}}$ in which $\bar{\rho}\ge\frac{\rho}{2}$. Then there exists a constant C such that

$$
\mathbb{E}|e_\Delta(T)|\le\begin{cases}C\Big(-\frac{1}{\ln(\Delta)}+\Delta^{\frac{1}{4}}\Big), & \text{if } \alpha=0,\\ C\Delta^{\alpha}, & \text{if } 0<\alpha<0.5.\end{cases}\tag{72}
$$

Proof. We split the left side of (72) into two parts

$$
\mathbb{E}|e_\Delta(T)|=\mathbb{E}|e_\Delta(T)I_{\{\beta_{\Delta,R}\le T\}}|+\mathbb{E}|e_\Delta(T)I_{\{\beta_{\Delta,R}>T\}}|.\tag{73}
$$

Using the Young inequality, we have

$$
\mathbb{E}|e_\Delta(T)I_{\{\beta_{\Delta,R}\le T\}}|\le\frac{\Delta^{\frac{1}{2}}}{\bar{\rho}+2}\mathbb{E}|e_\Delta(T)|^{\bar{\rho}+2}+\frac{\bar{\rho}+1}{\bar{\rho}+2}\Delta^{-\frac{1}{2(\bar{\rho}+1)}}\mathbb{P}(\beta_{\Delta,R}\le T).\tag{74}
$$

For any $R>|x_0|$, by Lemmas 1 and 5, we can write

$$
R^{\bar{\rho}+2}\mathbb{P}(\beta_{\Delta,R}\le T)\le C,
$$

which follows that

$$
\mathbb{P}(\beta_{\Delta,R}\le T)\le\frac{C}{R^{\bar{\rho}+2}}.\tag{75}
$$

If we substitute (75) into (74) and use Theorem 1 and Lemma 8, we get

$$
\mathbb{E}|e_\Delta(T)I_{\{\beta_{\Delta,R}\le T\}}|\le C\Big(\Delta^{\frac{1}{2}}+\frac{\Delta^{-\frac{1}{2(\bar{\rho}+1)}}}{R^{\bar{\rho}+2}}\Big).
$$

By choosing $R = \nu^{-1}(h(\Delta)) = (\frac{\eta_h}{\eta_\nu}\Delta^{-\frac{1}{2}})^{\frac{1}{1+\rho}}$, we have

$$\mathbb{E}|e_\Delta(T)I_{\{\beta_{\Delta,R}\leq T\}}| \leq C\Delta^{\frac{1}{2}}. \tag{76}$$

In what follows, we try to estimate the second terms in (73). Since $R = \nu^{-1}(h(\Delta))$, if $0 \leq \zeta \leq \beta_{\Delta,R}$, then $|Y_\Delta(\zeta)| \vee |Z_\Delta(\zeta)| \leq \nu^{-1}(h(\Delta))$ for all $\Delta \in [0,1]$. By the definition of the truncated function (19), $a_\Delta(|Y_\Delta(\zeta)|) = a(|Y_\Delta(\zeta)|)$ and $a_\Delta(|Z_\Delta(\zeta)|) = a(|Z_\Delta(\zeta)|)$. Using the same notation as in the proof of Lemma 7, according to the Itô formula we get

$$\mathbb{E}|e_\Delta(\xi \wedge \beta_{\Delta,R})| \leq \epsilon + \Pi_1 + \Pi 2, \tag{77}$$

where Π_1 and $\Pi 2$ are defined in (51). According to Assumption 1, Lemma 6, and Remark 4 as well as the Hölder inequality the following equation holds for Π_1

$$\begin{aligned}\Pi_1 \leq & H_1 \int_0^\xi \mathbb{E}|e_\Delta(\zeta \wedge \beta_{\Delta,R}))|d\zeta \\ & + \sqrt{K_1}\int_0^{\xi\wedge\beta_{\Delta,R}} \big(1 + \mathbb{E}|y_\Delta(\zeta)|^\rho + \mathbb{E}|Y_\Delta(\zeta)|^\rho\big)^{\frac{1}{2}} \big(\mathbb{E}\big|y_\Delta(\zeta)) - Y_\Delta(\zeta)\big|^2\big)^{\frac{1}{2}} d\zeta \\ & + \alpha_2\sqrt{K_1}\mathbb{E}\int_0^{\xi\wedge\beta_{\Delta,R}} \big(1 + \mathbb{E}|Y_\Delta(\zeta)|^\rho + \mathbb{E}|Z_\Delta(\zeta)|^\rho\big)^{\frac{1}{2}} \big(\mathbb{E}\big|Y_\Delta(\zeta)) - Z_\Delta(\zeta)\big|^2\big)^{\frac{1}{2}} d\zeta.\end{aligned}$$

From Theorem 1 and Lemma 8 we can thus conclude

$$\Pi_1 \leq C\int_0^\xi \mathbb{E}|e_\Delta(\zeta \wedge \beta_{\Delta,R}))|d\zeta + C\Delta^{\frac{1}{2}}. \tag{78}$$

As for Π_2, it follows from (6) and Lemmas 6 and 8 that

$$\begin{aligned}\Pi_2 &\leq H_2^2\mathbb{E}\int_0^{\xi\wedge\beta_{\Delta,R}} \big(x(\zeta) - Z_\Delta(\zeta)\big)^{1+2\alpha} \frac{1}{|e_\Delta(\zeta)|\ln\delta} I_{\{\epsilon/\delta\leq|e_\Delta(\zeta)|\leq\epsilon\}} d\zeta \\ &\leq \frac{2^{2\alpha}H_2^2\epsilon^{2\alpha}T}{\ln(\delta)} + \frac{2^{2\alpha}H_2^2\delta}{\epsilon\ln(\delta)}\mathbb{E}\int_0^\xi \big|y_\Delta(\zeta) - Z_\Delta(\zeta)\big|^{1+2\alpha} d\zeta \\ &\leq C\big(\frac{\epsilon^{2\alpha}}{\ln(\delta)} + \frac{\delta\Delta^{1/2+\alpha}}{\epsilon\ln(\delta)}\big).\end{aligned} \tag{79}$$

Substituting (78) and (79) into (77) and applying Gronwall's inequality yield

$$\mathbb{E}|e_\Delta(\xi \wedge \beta_{\Delta,R})| \leq C\big(\frac{\epsilon + \epsilon^{2\alpha}}{\ln(\delta)} + \frac{\delta\Delta^{1/2+\alpha}}{\epsilon\ln(\delta)} + \Delta^{1/2}\big).$$

In what follows we divide the proof into two cases $\alpha = 0$ and $\alpha \in (0, \frac{1}{2})$. In the case when $\alpha = 0$, we set $\delta = \Delta^{-\frac{1}{4}}$ and $\epsilon = \frac{1}{4}$, implying

$$\mathbb{E}|e_\Delta(\xi \wedge \beta_{\Delta,R})| \leq C\big(-\frac{1}{\ln(\Delta)} + \Delta^{\frac{1}{4}}\big). \tag{80}$$

In the case where $\alpha \in (0, \frac{1}{2})$, we set $\Delta = 2$ and $\epsilon = \Delta^{\frac{1}{2}}$, which implies

$$\mathbb{E}|e_\Delta(\xi \wedge \beta_{\Delta,R})| \leq C\Delta^\alpha. \tag{81}$$

From (77), (80), and (81) the proof follows directly. □

4 Convergence rate over a finite interval

Sometimes we need to approximate path-dependent quantities, for example, the value of the European barrier option value. In these cases, we need a stronger convergence result such as

$$\lim_{\Delta\to 0} \mathbb{E}\Big(\sup_{0\leq\xi\leq T} |x(\xi) - y_\Delta(\xi)|\Big) = 0.$$

Lemma 9. Consider the initial problem (1), which satisfies Assumptions 1 and 2. Suppose that $R > |x_0|$ is a real number and that $\Delta \in (0, 1]$ is sufficiently small such that $\nu^{(-1)}(h(\Delta)) \geq R$. Then for all $0 \leq \xi \leq T$, a constant C exists such that

$$\mathbb{E}\Big(\sup_{0\leq\zeta\leq\xi} |x(\zeta) - y_\Delta(\zeta)|\Big) \leq \begin{cases} C\Big(\frac{1}{\ln(\Delta^{-1})} + \Delta^{\frac{1}{4}}(h(\Delta))^{\frac{1}{2}}\Big)^{\frac{1}{2}}, & \text{if} \quad \alpha = 0, \\ C\big(\Delta^{\frac{1}{2}}h(\Delta)\big)^{4\alpha^2}, & \text{if} \quad \alpha \in (0, \frac{1}{2}). \end{cases} \tag{82}$$

Proof. By the Itô formula, we can write

$$\begin{aligned} |e_\Delta(\xi \wedge \beta_{\Delta,R})| \leq& \varepsilon + \big(\varpi_{\delta,\epsilon}(e_\Delta(\xi \wedge \beta_{\Delta,R}))\big) \\ =& \epsilon + \int_0^{\xi\wedge\beta_{\Delta,R}} \varpi'_{\delta,\epsilon}(e_\Delta(\zeta))\Big[a(x(\zeta)) - \alpha_1 a_\Delta(Y_\Delta(\zeta)) - \alpha_2 a_\Delta(Z_\Delta(\zeta))\Big]d\zeta \\ &+ \frac{1}{2}\int_0^{\xi\wedge\beta_{\Delta,R}} \varpi''_{\delta,\epsilon}(e_\Delta(\zeta))\big[b(x(\zeta)) - b(Z_\Delta(\zeta))\big]^2 d\zeta + S_{\delta,\epsilon,\Delta}(\xi), \end{aligned}$$

where, for $\xi \geq 0$

$$S_{\delta,\epsilon,\Delta}(\xi) = \int_0^{\xi\wedge\beta_{\Delta,R}} \varpi'_{\delta,\epsilon}(e_\Delta(\zeta))\big(b(x(\zeta)) - b(Z_\Delta(\zeta))\big)dW(\zeta). \tag{83}$$

Since $\nu^{(-1)}(h(\Delta)) \geq R$, if $0 \leq \zeta \leq \xi \wedge \beta_{\Delta,R}$, we have $|Y_\Delta(\zeta)| \vee |Z_\Delta(\zeta)| \leq \nu^{(-1)}(h(\Delta))$, which gives $a_\Delta(Y_\Delta(\zeta)) = a(Y_\Delta(\zeta))$ and $a_\Delta(Z_\Delta(\zeta)) = a(Z_\Delta(\zeta))$. If we use the same technique as in the proof of Lemma 7 and apply Assumption 1, then we have

$$|e_\Delta(\xi \wedge \beta_{\Delta,R})| \leq \varepsilon + H_1 \int_0^{\xi\wedge\beta_{\Delta,R}} |e_\Delta(\zeta))|d\zeta \tag{84}$$

$$
\begin{aligned}
&+\sqrt{K_1}\int_0^{\xi\wedge\beta_{\Delta,R}}\big(1+|y_\Delta(\zeta)|^\rho+|Y_\Delta(\zeta)|^\rho\big)^{\frac{1}{2}}\big|y_\Delta(\zeta))-Y_\Delta(\zeta)\big|d\zeta\\
&+\alpha_2\sqrt{K_1}\int_0^{\xi\wedge\beta_{\Delta,R}}\big(1+|Y_\Delta(\zeta)|^\rho+|Z_\Delta(\zeta)|^\rho\big)^{\frac{1}{2}}\big|Y_\Delta(\zeta))-Z_\Delta(\zeta)\big|d\zeta\\
&+\frac{2^{2\alpha}H_2^2\varepsilon^{2\alpha}}{\ln(\delta)}\xi\\
&+\frac{2^{2\alpha}H_2^2\delta}{\varepsilon\ln(\delta)}\int_0^{\xi\wedge\beta_{\Delta,R}}\big|y_\Delta(\zeta)-Z_\Delta(\zeta)\big|^{1+2\alpha}d\zeta+S_{\delta,\epsilon,\Delta}(\xi).
\end{aligned}
$$

By the Hölder's inequality and Lemmas 3 and 4, we have

$$
\begin{aligned}
\mathbb{E}\Big(\sup_{0\le u\le\xi}|e_\Delta(u\wedge\beta_{\Delta,R})|\Big)\le&\varepsilon+H_1\int_0^\xi\mathbb{E}\Big(\sup_{0\le u\le\zeta}|e_\Delta(u\wedge\beta_{\Delta,R})|\Big)d\zeta+C\Delta^{\frac{1}{2}}h(\Delta)\\
&+\frac{2^{2\alpha}H_2^2\varepsilon^{2\alpha}}{\ln(\delta)}\xi+\frac{2^{2\alpha}H_2^2\delta}{\varepsilon\ln(\delta)}\mathbb{E}\int_0^\xi\big|y_\Delta(\zeta)-Z_\Delta(\zeta)\big|^{1+2\alpha}d\zeta\\
&+\mathbb{E}\Big(\sup_{0\le u\le\xi}|S_{\delta,\epsilon,\Delta}(u)|\Big).
\end{aligned}
\tag{85}
$$

Next, we try to estimate the value $\mathbb{E}\Big(\sup_{0\le u\le\xi}|S_{\delta,\epsilon,\Delta}(u)|\Big)$ in (85). By the Burkholder–Davis–Gundy inequality and Lemma 6, we can write

$$
\begin{aligned}
\mathbb{E}\Big(\sup_{0\le u\le\xi}|S_{\delta,\epsilon,\Delta}(u)|\Big)\le&4\sqrt{2}\mathbb{E}\Big(\int_0^{\xi\wedge\beta_{\Delta,R}}|\varpi'_{\delta\epsilon}(e_\Delta(\zeta))|^2\big|b(x(\zeta))-b(Z_\Delta(\zeta))\big|^2d\zeta\Big)^{\frac{1}{2}}\\
\le&2^{\alpha+\frac{5}{2}}H_2\mathbb{E}\Big(\int_0^{\xi\wedge\beta_{\Delta,R}}|x(\zeta)-x_\Delta(\zeta)|^{1+2\alpha}d\zeta\Big)^{\frac{1}{2}}\\
&+2^{\alpha+\frac{5}{2}}H_2\mathbb{E}\Big(\int_0^{\xi\wedge\beta_{\Delta,R}}|x_\Delta(\zeta)-Z_\Delta(\zeta)|^{1+2\alpha}d\zeta\Big)^{\frac{1}{2}}.
\end{aligned}
$$

In fact, by Assumption 2 we have

$$
\begin{aligned}
\big|b(x(\zeta))-b(Z_\Delta(\zeta))\big|^2&\le H_2|x(\zeta)-Z_\Delta(\zeta)|^{1+2\alpha}\\
&\le 2^\alpha H_2|x(\zeta)-x_\Delta(\zeta)|^{1+2\alpha}+2^\alpha H_2|x_\Delta(\zeta)-Z_\Delta(\zeta)|^{1+2\alpha}.
\end{aligned}
$$

Therefore, by Lemma 3 we have

$$
\begin{aligned}
\mathbb{E}\Big(\sup_{0\le u\le\xi}|S_{\delta,\epsilon,\Delta}(u)|\Big)\le&C\Big(\Delta^{\frac{1}{4}}\sqrt{h(\Delta)}\Big)^{1+2\alpha}\\
&+2^{\alpha+\frac{5}{2}}H_2\mathbb{E}\Big(\int_0^{\xi\wedge\beta_{\Delta,R}}|x(\zeta)-x_\Delta(\zeta)|^{1+2\alpha}d\zeta\Big)^{\frac{1}{2}}
\end{aligned}
\tag{86}
$$

In the rest, we divide the proof into two cases.

Case 1: Suppose that $\alpha = 0$. In this case, according to Lemma 7 and from the relation (86), it follows that

$$\mathbb{E}\Big(\sup_{0\le u\le \xi}|S_{\delta,\epsilon,\Delta}(u)|\Big) \le C\left(\Delta^{\frac{1}{4}}\sqrt{h(\Delta)} + \Big(\frac{1}{\ln(\Delta^{-1})} + \Delta^{\frac{1}{4}}(h(\Delta))^{\frac{1}{2}}\Big)^{\frac{1}{2}}\right). \quad (87)$$

Substituting (87) into (85) and choosing $\delta = \Delta^{-\frac{1}{8}}$ and $\varepsilon = -1/\ln(\Delta)$, we get

$$\begin{aligned}\mathbb{E}\Big(\sup_{0\le u\le \xi}|e_\Delta(u\wedge\beta_{\Delta,R})|\Big) \le& H_1\int_0^\xi \mathbb{E}\Big(\sup_{0\le u\le \zeta}|e_\Delta(u\wedge\beta_{\Delta,R})|\Big)d\zeta \\ &+ C\left(\Delta^{\frac{1}{2}}h(\Delta) + \Delta^{\frac{1}{4}}(h(\Delta))^{\frac{1}{2}} - \frac{1}{\ln(\Delta)} + \Delta^{\frac{3}{8}}h(\Delta)\right. \\ &\left. + \Big(\frac{1}{\ln(\Delta^{-1})} + \Delta^{\frac{1}{4}}(h(\Delta))^{\frac{1}{2}}\Big)^{\frac{1}{2}}\right),\end{aligned}$$

which implies

$$\begin{aligned}\mathbb{E}\Big(\sup_{0\le u\le \xi}|e_\Delta(u\wedge\beta_{\Delta,R})|\Big) \le& H_1\int_0^\xi \mathbb{E}\Big(\sup_{0\le u\le \zeta}|e_\Delta(u\wedge\beta_{\Delta,R})|\Big)d\zeta \\ &+ C\Big(\frac{1}{\ln(\Delta^{-1})} + \Delta^{\frac{1}{4}}(h(\Delta))^{\frac{1}{2}}\Big)^{\frac{1}{2}}. \quad (88)\end{aligned}$$

Case 2: Suppose $\alpha \in (0, \frac{1}{2})$. In this case from (86) we have

$$\begin{aligned}\mathbb{E}\Big(\sup_{0\le u\le \xi}|S_{\delta,\epsilon,\Delta}(u)|\Big) \le& C\Big(\Delta^{\frac{1}{4}}\sqrt{h(\Delta)}\Big)^{1+2\alpha} \\ &+ 2^{\alpha+\frac{5}{2}}H_2\mathbb{E}\left(\Big(\sup_{0\le u\le \xi}|e_\Delta(\xi\wedge\beta_{\Delta,R})|\Big)^{\frac{1}{2}}\Big(\int_0^{\xi\wedge\beta_{\Delta,R}}|e_\Delta(\zeta)|^{2\alpha}d\zeta\Big)^{\frac{1}{2}}\right).\end{aligned}$$

By Young's inequality and Lemma 7 we have

$$\begin{aligned}\mathbb{E}\Big(\sup_{0\le u\le \xi}|S_{\delta,\epsilon,\Delta}(u)|\Big) \le& C\left(\Big(\Delta^{\frac{1}{4}}\sqrt{h(\Delta)}\Big)^{1+2\alpha} + \big(\sqrt{\Delta}h(\Delta)\big)^{4\alpha^2}\right) \\ &+ \frac{1}{2}\mathbb{E}\Big(\sup_{0\le u\le \xi}|e_\Delta(\xi\wedge\beta_{\Delta,R})|\Big)\Big). \quad (89)\end{aligned}$$

Inserting (89) in (85) and by choosing $\varepsilon = \Delta^{\frac{1}{2}}h(\Delta)$ and $\delta = 2$, we get

$$\mathbb{E}\Big(\sup_{0\le u\le \xi}|e_\Delta(u\wedge\beta_{\Delta,R})|\Big) \le 2H_1\int_0^\xi \mathbb{E}\Big(\sup_{0\le u\le \zeta}|e_\Delta(u\wedge\beta_{\Delta,R})|\Big)d\zeta + C\big(\sqrt{\Delta}h(\Delta)\big)^{4\alpha^2}. \quad (90)$$

Hence, by using the Gronwall inequality on (88) and (90) we have the required assertion. □

Theorem 4. Let $\nu(r) = \eta_h r^{1+\bar{\rho}}$ and $h(\Delta) = \eta_h \Delta^{-\omega\epsilon}$ in which $\bar{\rho} \geq \frac{\rho}{2}$, and

$$\epsilon \in \big(0, \min\{\frac{1}{4\omega}, \frac{3}{4-2\omega}\}\big),$$

for any $\omega \in (0,2)$. Let Assumptions 1 and 2 hold, and also let Δ^* be sufficiently small such that $\Delta^{\frac{1-2\omega\epsilon}{4}} \leq \frac{1}{\ln \Delta^{-1}}$ for all $\Delta \in (0, \Delta^*]$. Then

$$\mathbb{E}\Big(\sup_{0\leq\zeta\leq T}|x(\zeta)-y_\Delta(\zeta)|\Big) \leq \begin{cases} C\Big(\frac{1}{\ln(\Delta^{-1})}\Big)^{\frac{1}{2}}, & \text{if} \quad \alpha = 0, \\ C\Big(\Delta^{2\alpha^2(1-2\omega\epsilon)} + \Delta^{\frac{1-\epsilon}{2}}\Big), & \text{if} \quad \alpha \in (0, \frac{1}{2}). \end{cases} \tag{91}$$

Proof. For any $R > |x_0|$, we can decompose the left side of (91) into two parts as below:

$$\mathbb{E}\Big(\sup_{0\leq\zeta\leq T}|x(\zeta)-y_\Delta(\zeta)|\Big) \leq \mathbb{E}\Big(\sup_{0\leq u\leq \xi}|e_\Delta(u\wedge\beta_{\Delta,R})|\Big)+\mathbb{E}\Big(\sup_{0\leq\zeta\leq T}|e_\Delta(\zeta)|I_{\{\beta_{\Delta,R}\leq T\}}\Big). \tag{92}$$

Let $p > 1 + \frac{1+\bar{\rho}}{2\omega\epsilon}$ be fixed. By the Young inequality, that is

$$z_1 z_2 \leq \frac{\delta}{p} z_1^p + \frac{p-1}{p\delta^{1/(p-1)}} z_2^{1/(p-1)}, \qquad \text{for all } z_1, z_2 \in [0, +\infty), \qquad \text{for all } \delta > 0,$$

and Lemmas 1 and 5, we can deduce

$$\begin{aligned} \mathbb{E}\Big(\sup_{0\leq\zeta\leq T}|e_\Delta(\zeta)|I_{\{\beta_{\Delta,R}\leq T\}}\Big) &\leq \frac{\delta}{p}\mathbb{E}\Big(\sup_{0\leq\zeta\leq T}|e_\Delta(\zeta)|^p\Big) + \frac{p-1}{p\delta^{1/p-1}}\mathbb{P}(\beta_{\Delta,R} \leq T) \\ &\leq \frac{C\delta}{p} + \frac{p-1}{p\delta^{1/p-1}R^p}. \end{aligned} \tag{93}$$

By choosing

$$\delta = \Delta^{(1-\epsilon)/2}, \qquad R = \Delta^{-(1-\epsilon)/2(p-1)},$$

we see that $\nu^{(-1)}(h(\Delta)) \geq R$. Therefore, by substituting (93) into (92) and applying Lemma 9, we have

$$\mathbb{E}\Big(\sup_{0\leq\zeta\leq T}|x(\zeta)-y_\Delta(\zeta)|\Big) \leq \begin{cases} C\Big(\big(\frac{1}{\ln(\Delta^{-1})} + \Delta^{\frac{1-2\omega\epsilon}{4}}\big)^{\frac{1}{2}} + \Delta^{\frac{1-\epsilon}{2}}\Big), & \text{if} \quad \alpha = 0, \\ C\Big(\Delta^{2\alpha^2(1-2\omega\epsilon)} + \Delta^{\frac{1-\epsilon}{2}}\Big), & \text{if} \quad \alpha \in (0, \frac{1}{2}), \end{cases}$$

which completes the proof of the theorem. □

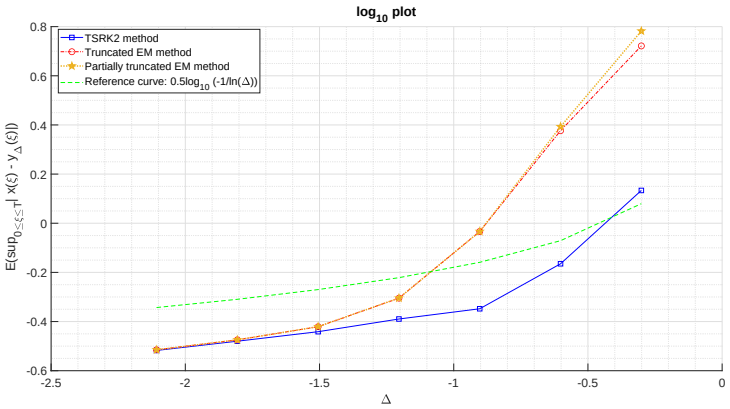


Figure 1: The supremum convergence norm (95) in $[0, 2]$ as a function of step size Δ for Example 1 with $\epsilon = 0.0001$, $\omega = 0.5$, and $\alpha = 0$.

5 Numerical experiments

In Section 3, we derived TSRK2 methods with free parameters α_1, α_2, and θ. In this section, we follow [17] and set them to below:

$$\alpha_1 = \alpha_2 = \frac{1}{2} \qquad \text{and} \qquad \theta = \frac{\omega_1}{\omega_0}, \tag{94}$$

in which $\omega_0 = 1.9$ and $\omega_1 = 0.8184$. We confirm the performance of the new method in terms of accuracy and stability in contrast to the truncated EM method [21, 32] and the modified partially truncated EM method [33]. In this context, we investigate

$$\mathbb{E}\Big(\sup_{0\le\zeta\le T}|x(\zeta) - y_\Delta(\zeta)|\Big) \approx \frac{1}{2000}\sum_{i=1}^{2000}\max_{n=1,\dots,M}\big|x^{(i)}(t_n) - y_\Delta^{(i)}(t_n)\big|, \tag{95}$$

for a given $\Delta = T/N$ to measure the accuracy of the methods.

Example 1. Consider the following SDE with Hölder continuous diffusion coefficient

$$\begin{aligned} dx(\xi) &= \Big(\lambda_1 x(\xi) - \lambda_2 x^3(\xi)\Big)d\xi + \mu|x(\xi)|^{\frac{1}{2}+\alpha}dW(\xi), \quad \xi \in [0, T], \\ x(0) &= x_0 \in \mathbb{R}, \end{aligned} \tag{96}$$

where $\lambda_1 \in \mathbb{R}$ and $\lambda_2, \mu \in \mathbb{R}^+$.

The drift and diffusion coefficients are $a(z) = \lambda_1 z - \lambda_2 z^3$ and $b(z) = \mu|z|^{\frac{1}{2}+\alpha}$, respectively. We can easily show that, for $z_1, z_2 \in \mathbb{R}$,

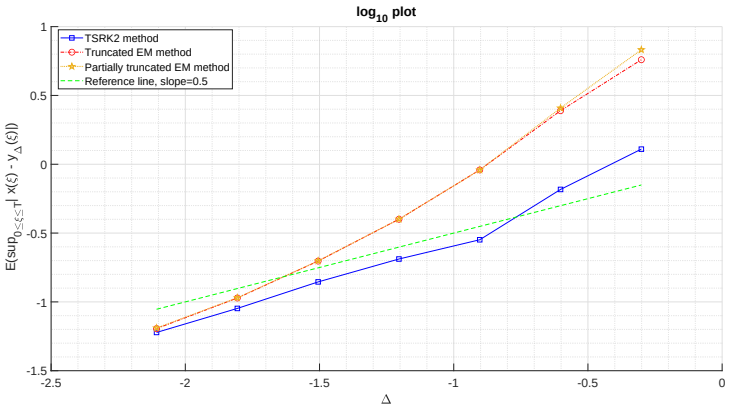


Figure 2: The supremum convergence norm (95) in $[0,2]$ as a function of step size Δ for Example 1 with $\epsilon = 0.0001$, $\omega = 0.5$, and $\alpha = 0.25$.

$$|a(z_1) - a(z_2)|^2 \leq \tilde{K}_1(1 + |z_1|^4 + |z_2|^4)|z_1 - z_2|^2,$$

where $\tilde{K}_1 := 3\max\{\lambda_1^2, \frac{9}{4}\lambda_2^2\}$. In other words, Assumption 1 is satisfied with $\rho = 4$. Moreover, we have

$$(z_1 - z_2)(a(z_1) - a(z_2)) \leq \lambda_1(z_1 - z_2)^2, \quad |b(z_1) - b(z_2)| \leq \mu|z_1 - z_2|^{\frac{1}{2}+\alpha}.$$

That is to say, Assumption 2 is fulfilled. Concerning (10), we set $\eta_h = 4\sqrt{\tilde{K}_1}$. Then, we have

$$\sup_{0<|z_1|\vee|z_2|\leq r} \frac{|a(z_1) - a(z_2)|}{|z_1 - z_2|} \leq \eta_h\big(\frac{1+r^2}{2}\big) \leq \eta_h r^3,$$

and

$$\sup_{|z|\leq r} |a(z)| \leq (|\lambda_1| + \lambda_2)r^3 \leq \eta_h r^3,$$

for all $r \geq 1$. To apply Theorem 4, we set $\nu(r) = \eta_h r^3$ and $h(\Delta) = \eta_h \Delta^{-\omega\epsilon}$. We can therefore conclude that the truncated Runge–Kutta solution (27) satisfies

$$\mathbb{E}\Big(\sup_{0\leq\zeta\leq T} |x(\zeta) - y_\Delta(\zeta)|\Big) \leq \begin{cases} C\Big(\frac{1}{\ln(\Delta^{-1})}\Big)^{\frac{1}{2}}, & \text{if} \quad \alpha = 0, \\ C\Big(\Delta^{2\alpha^2(1-2\omega\epsilon)} + \Delta^{\frac{1-\epsilon}{2}}\Big), & \text{if} \quad \alpha \in (0, \frac{1}{2}). \end{cases}$$

Let us now test the efficiency of the TSRK2 method for problem (96) in comparison with the truncated EM method [21] and the modified partially truncated EM method [33] for different values of the Hölder parameter α. Since we do not know the exact solution of (96), we search for a numerical

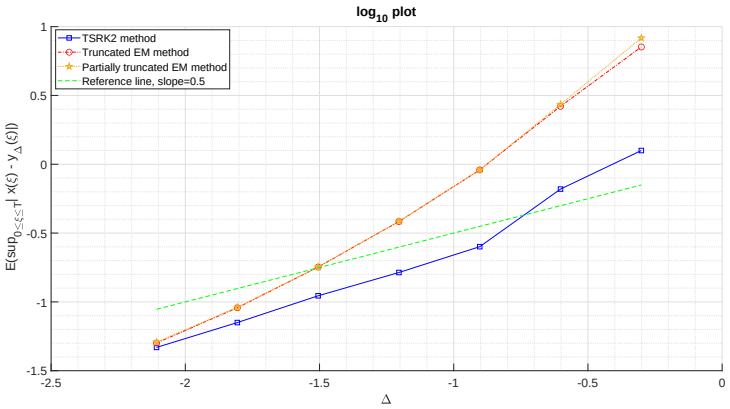


Figure 3: The supremum convergence norm (95) in $[0, 2]$ as a function of step size Δ for Example 1 with $\epsilon = 0.0001$, $\omega = 0.5$, and $\alpha = 0.45$.

solution with $\Delta = 2^{-18}$ using the truncated EM method and use it instead of the exact solution. Figures 1–3 show simulations of (95) for the TSRK2 method with parameters (94), the truncated EM method, and the modified partially truncated EM method with $\omega = 0.5$. Here we take $\lambda_1 = -1.5$, $\lambda_2 = 10$, and $\mu = 1.5$ as parameters of (96) and apply the methods over $0 \leq \xi \leq 2$ with step sizes, $\Delta = 2^{i-2}$ for $i = 1, 2, \ldots, 9$. In these figures, the error of the methods is plotted as a function of seven-step sizes Δ in log_{10} for different values of α $(0, 0.25$ and 0.45, respectively).

Figure 1 shows that, compared to the reference curve $0.5 \log_{10}(-\frac{1}{\ln(\Delta)})$, the TSRK2 method gives better convergence results than the truncated EM method and the modified partially truncated EM method for step size $2^{-5} \leq \Delta \leq 2^{-1}$. However, for sufficiently small Δ (e.g., $\Delta \in (0, 2^{-5}]$), all methods have the same slope values as in Theorem 4. Increasing the value of the Hölder parameter α, it can be seen in Figures 2 and 3 that for the step size $2^{-5} \leq \Delta \leq 2^{-1}$, the TSRK2 method has a better convergence rate than the other methods compared to the reference line with slope $1/2$, but for sufficiently small $\Delta \leq 2^{-6}$, all methods have the same convergence rate close to 0.5 as in Theorem 4. The simulation results clearly show that the TSRK2 method confirms the convergence results stated in Theorem 4 and that the new method is the most efficient method in terms of accuracy compared to the truncated EM method and the modified partially truncated EM method.

Example 2. Consider the scalar nonlinear Itô SDE with a one-dimensional Wiener process

$$\begin{aligned} dx(t) &= \Big(\lambda_1 x(\xi) - \lambda_2 x^5(\xi)\Big) d\xi + \mu |x(\xi)|^{\frac{1}{2}+\alpha} dW(\xi), \quad t \in [0, T], \\ x(0) &= 1, \end{aligned} \tag{97}$$

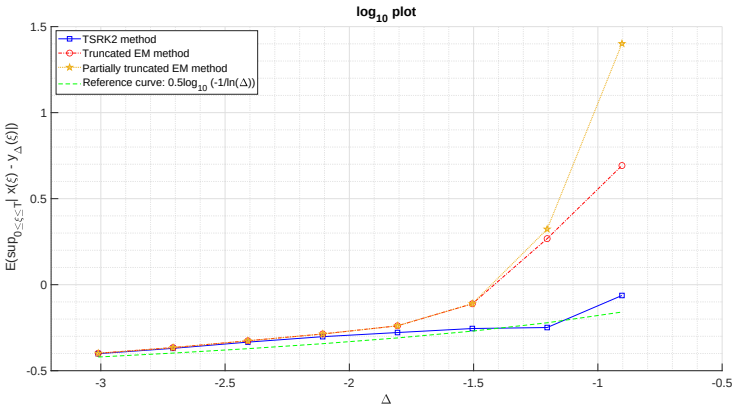


Figure 4: The supremum convergence norm (95) in $[0, 2]$ as a function of step size Δ for Example 2 with $\epsilon = 0.1$, $\omega = 0.5$, and $\alpha = 0$.

where $\lambda_1 \in \mathbb{R}$ and $\lambda_2, \mu \in \mathbb{R}^+$.

The drift and diffusion coefficients are $a(z) = \lambda_1 z - \lambda_2 z^5$ and $b(z) = \mu|z|^{\frac{1}{2}+\alpha}$, respectively. We can easily show that, for $z_1, z_2 \in \mathbb{R}$,

$$|a(z_1) - a(z_2)|^2 \leq \tilde{K}_2(1 + |z_1|^8 + |z_2|^8)|z_1 - z_2|^2,$$

where $\tilde{K}_2 := 3\max\{\lambda_1^2, \frac{25}{4}\lambda_2^2\}$. In other words, Assumption 1 is satisfied with $\rho = 8$. Moreover, we have

$$(z_1 - z_2)(a(z_1) - a(z_2)) \leq \lambda_1(z_1 - z_2)^2, \quad |b(z_1) - b(z_2)| \leq \mu|z_1 - z_2|^{\frac{1}{2}+\alpha}.$$

Concerning (10), we set $\eta_\nu = 4\sqrt{\tilde{K}_2}$. Then, we conclude

$$\sup_{0<|z_1|\vee|z_2|\leq r} \frac{|a(z_1) - a(z_2)|}{|z_1 - z_2|} \leq \eta_h\Big(\frac{1 + r^4}{2}\Big) \leq \eta_h r^5,$$

and

$$\sup_{|z|\leq r} |a(z)| \leq (|\lambda_1| + \lambda_2)r^5 \leq \eta_h r^5,$$

for all $r \geq 1$. To apply Theorem 4, we set $\nu(r) = \eta_\nu r^5$ and $h(\Delta) = \eta_\nu \Delta^{-\omega\epsilon}$. Figures 4–6 show the simulation of (95) of the TSRK2 method for equation (97) as a function of step size Δ for Hölder $\alpha = 0$, $\alpha = 0.25$, and $\alpha = 0.45$, respectively. Since there is no explicit solution, we use the truncated EM solution with $\Delta = 2^{-18}$ as a suitable approximation to the exact solution. Here we take $\lambda_1 = -15$, $\lambda_2 = 20$, and $\mu = 3$ as stiff parameters of (97) and apply the methods over $0 \leq \xi \leq 2$ with different step sizes $\Delta \in \{2^{-3}, 2^{-4}, \cdots, 2^{-10}\}$.

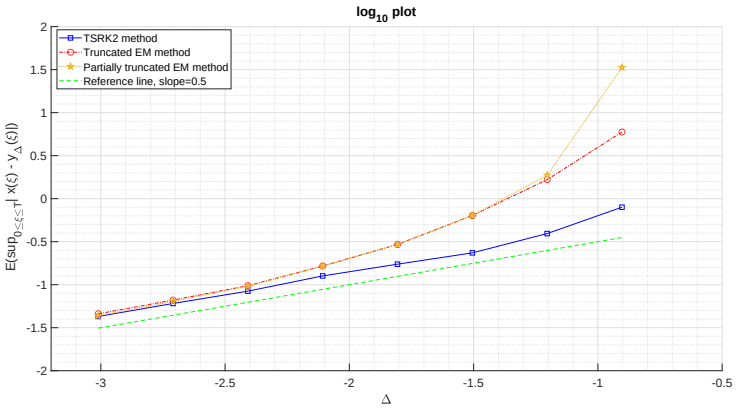


Figure 5: The supremum convergence norm (95) in $[0, 2]$ as a function of step size Δ for Example 2 with $\epsilon = 0.1$, $\omega = 0.5$ and $\alpha = 0.25$.

Figure 4, corresponding to $\alpha = 0$, shows that the TSRK2 method gives better convergence results than the truncated EM method and the modified partially truncated EM method, especially for the step size $2^{-6} \leq \Delta \leq 2^{-3}$, compared to the reference curve $0.5 \log_{10}(-\frac{1}{\ln(\Delta)})$. For the step size $\Delta = 2^{-3}$, for example, the new method has an error of 0.8655, while the truncated EM method and the modified partially truncated EM method reach an error of 4.9260 and 25.1876, respectively. It should be emphasized that the modified partially truncated method EM is not applicable for $\Delta \geq 2^{-2}$. However, for sufficiently small Δ (e.g., $\Delta \leq 2^{-6}$), all methods have the same slope values as in Theorem 4.

In Figures 4 and 5, we present the supremum norm of the methods for larger values of $\alpha = 0.25$ and $\alpha = 0.45$, respectively. These figures also show that the new method has a better convergence rate compared to the other methods mentioned, especially when the step sizes are not very small.

6 Conclusion

In this work, we have developed an explicit TSRK2 scheme for nonlinear one-dimensional Itô-SDEs with one-sided Lipschitz and local Lipschitz drift conditions and continuous Hölder diffusion condition. We proved the moment boundedness and convergence properties of the approximate solutions at a time T and in a finite time interval $[0, T]$, respectively. We have demonstrated the efficiency of the new method using some numerical examples with different Hölder parameters $\alpha \in [0, \frac{1}{2})$ and compared our method with the truncated EM method and the modified partially truncated EM method in terms of their strong convergence performance. The numerical simulations

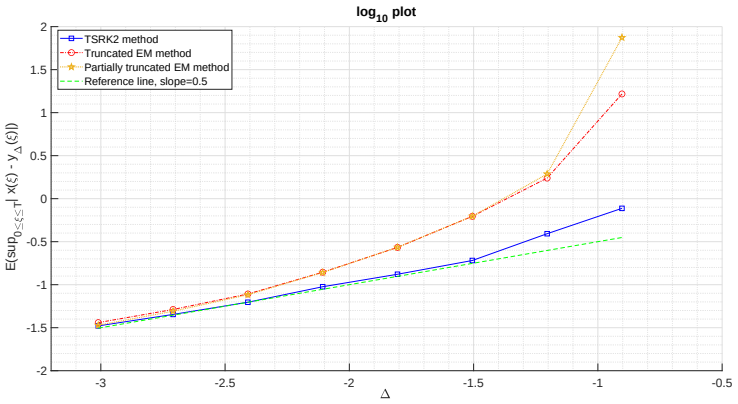


Figure 6: The supremum convergence norm (95) in $[0, 2]$ as a function of step size Δ for Example 2 with $\epsilon = 0.1$, $\omega = 0.5$ and $\alpha = 0.45$.

we performed confirmed the theoretical convergence results and also showed that the TSRK2 method is robust to changes in the Hölder parameters of the SDEs.

Acknowledgements

The authors are grateful to the reviewers and the editor for their constructive comments.

References

[1] Aït-Sahalia, Y. *Transition densities for interest rate and other nonlinear diffusions*, J. Fainance 54(4) (1999) 1361–1395.

[2] Buckwar, E. and Kelly, C. *Towards a systematic linear stability analysis of numerical methods for systems of stochastic differential equations*, SIAM J. Numer. Anal. 48(1) (2010), 298–321.

[3] Debrabant, K. and Rößler, A. *Diagonally drift-implicit Runge–Kutta methods of weak order one and two for Itô SDEs and stability analysis*, Appl. Numer. Math. 59(3) (2009) 595–607.

[4] Emmanuel, C. and Mao, X. *Truncated EM numerical method for generalised Aït-Sahalia-type interest rate model with delay*, J. Comput. Appl. Math. 383 (2021), 113137.

[5] Falsone, G. *Stochastic differential calculus for Gaussian and non-Gaussian noises: A critical review*, Commum. Nonlinear. Sci. 56 (2018), 198–216.

[6] Fatemion Aghda, A.S., Hosseini, S.M. and Tahmasebi, M. *Analysis of non-negativity and convergence of solution of the balanced implicit method for the delay Cox–Ingersoll–Ross model*, Appl. Numer. Math. 118 (2017), 249–265.

[7] Greenhalgh, D., Liang Y., and Mao, X. *Demographic stochasticity in the SDE SIS epidemic model*, Discrete Contin. Dyn. Syst. Ser. B. 20(9) (2015) 2859–2884.

[8] Haghighi, A. *New S-ROCK methods for stochastic differential equations with commutative noise*, Iran. J. Numer. Anal. Optim. 9(1) (2019), 105–126.

[9] Haghighi, A. and Rößler, A. *Split-step double balanced approximation methods for stiff stochastic differential equations*, Int. J. Appl. Math. 96(5) (2019), 1030–1047.

[10] Higham, D.J., Mao, X. and Stuart, A. M. *Strong convergence of Euler-type methods for nonlinear stochastic differential equations*, SIAM J. Numer. Anal. 40(3) (2002) 1041–1063.

[11] Hu, L., Li, X. and Mao, X. *Convergence rate and stability of the truncated Euler–Maruyama method for stochastic differential equations*, J. Comput. Appl. Math. 337 (2018) 274–289.

[12] Hutzenthaler, M. and Jentzen, A. *Numerical approximations of stochastic differential equations with non-globally Lipschitz continuous coefficients*, Providence, RI: American Mathematical Society (AMS), 2015.

[13] Hutzenthaler, M., Jentzen, A. and Kloeden, P.E. *Strong and weak divergence in finite time of Euler's method for stochastic differential equations with non-globally Lipschitz continuous coefficients*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci 467 (2011) 1563–1576.

[14] Hutzenthaler, M., Jentzen, A. and Kloeden, P.E. *Strong convergence of an explicit numerical method for SDEs with non-globally Lipschitz continuous coefficients*, Ann. Appl. Probab. 22(4) (2012) 1611–1641.

[15] Ikeda, N. and Watanabe, S. *Stochastic Differential Equations and Diffusion Process*, North-Holand, Amesterdam, 1981.

[16] Kloeden, P. and Platen, E. *Numerical solution of stochastic differential equations*, Vol. 23, Springer-Verlag, Berlin, 1999.

[17] Komori, Y. and Burrage, K. *Strong first order S-ROCK methods for stochastic differential equations*, J. Comput. Appl. Math. 242 (2013), 261–274.

[18] Li, X., Mao X. and Yin, G. *Explicit numerical approximations for stochastic differential equations in finite and infinite horizons: truncation methods, convergence in pth moment and stability*, IMA J. Numer. Anal. 39(2) (2018), 847–892.

[19] Liu, W. and Mao, X. *Strong convergence of the stopped Euler-Maruyama method for nonlinear stochastic differential equations*, Appl. Math. Comput. 223 (2013), 389–400.

[20] Mao, X. *Stochastic Differential Equations and Applications*, 2nd edition, Horwood, Chichester, UK, 2007.

[21] Mao, X. *The truncated Euler–Maruyama method for stochastic differential equations*, J. Comput. Appl. Math. 290 (2015) 370–384.

[22] Mao, X. *Convergence rates of the truncated Euler–Maruyama method for stochastic differential equations*, J. Comput. Appl. Math. 296 (2016) 362–375.

[23] Mao, X. and Szpruch, L. *Strong convergence and stability of implicit numerical methods for stochastic differential equations with non-globally Lipschitz continuous coefficients*, J. Comput. Appl. Math. 238 (2013) 14–28.

[24] Milstein, G. N. *Numerical Integration of Stochastic Differential Equations*, Kluwer Academic, Dordrecht, 1995.

[25] Milstein, G.N., Platen, E. and Schurz, H. *Balanced implicit methods for stiff stochastic system*, SIAM J. Numer. Anal. 35(3) (1998) 1010–1019.

[26] Øksendal, B. *Stochastic Differential Equations: An Introduction with Applications*, 6th edition, Springer, Berlin, 2003.

[27] Rößler, A. *Runge–Kutta methods for the strong approximation of solutions of stochastic differential equations*, SIAM J. Numer. Anal. 48(3) (2010) 922–952.

[28] Sabanis, S. *A note on tamed Euler approximations*, Electron. Comm. Probab. 18(47) (2013) 1–10.

[29] Soheili, R.A., Amini, M. and Soleymani, F. *A family of Chaplygin-type solvers for Itô stochastic differential equations*, Appl. Math. Comput. 340 (2019) 296–304.

[30] Tretyakov, M. and Zhang, Z. *A fundamental mean-square convergence theorem for SDEs with locally Lipschitz coefficients and its applications*, SIAM J. Numer. Anal. 51(6) (2013) 3135–3162.

[31] Yamada, T. and Watanabe, S. *On the uniqueness of stochastic differential equations*, J. Math. Kyoto Univ. 11 (1971) 155–167.

[32] Yang, H., Fuke, W., Kloeden P.E. and Mao, X. *The truncated Euler–Maruyama method for stochastic differential equations with Hölder diffusion coefficients*, J. Comput. Appl. Math. 366 (2020) 112379.

[33] Yang, H. and Huang, J. *Convergence and stability of modified partially truncated Euler–Maruyama method for nonlinear stochastic differential equations with Hölder continuous diffusion coefficient*, J. Comput. Appl. Math. 404 (2022) 113895.

How to cite this article
Haghighi, A., Modified Runge–Kutta method with convergence analysis for nonlinear stochastic differential equations with Hölder continuous diffusion coefficient. *Iran. j. numer. anal. optim.*, 2023; 13(2): 285-316. https://doi.org/10.22067/ijnao.2022.78723.1181

Iranian Journal of Numerical Analysis and Optimization
Vol. 13, No. 2, 2023, pp 317–335
https://doi.org/10.22067/ijnao.2022.79205.1190
https://ijnao.um.ac.ir/

How to cite this article
Research Article

Two-layer shallow water formula with slope and uneven bottom solved by finite volume method

U. Habibah*, I.R Lina, and W.M Kusumawinahyu

Abstract

This paper proposes a numerical approach to solve a two-layer shallow water formula with a slope and uneven bottom. The finite volume method (FVM) is applied to solve the shallow water model because the method is suitable for computational fluid dynamics problems. Rather than pointwise approximations at grid points, the FVM breaks the domain into grid cells and approximates the total integral over grid cells. The shallow water model is examined in two cases, the shallow water model in the steady state and the unsteady state. The quadratic upstream interpolation for convective kinetics (QUICK) is chosen to get the discretization of the space domain since it is a third-order scheme, which provides good accuracy, and this scheme proves its numerical stability. The advantage of the QUICK method is that the main coefficients are positive and satisfy the requirements for conservativeness, boundedness, and transportation. An explicit scheme is used to get the discretization of the time domain. Finally, the numerical solution of the steady state model shows that the flow remains unchanged. An unsteady-state numerical solution produces instability (wavy at the bottom layer). Moreover, the larger slope results in higher velocity and higher depth at the second layer.

*Corresponding author

Received 16 October 2022; revised 3 December 2022; accepted 13 December 2022

Ummu Habibah
Department of Mathematics, Faculty of Mathematics and Natural Sciences, University of Brawijaya, Malang, Indonesia. e-mail: ummu_habibah@ub.ac.id

Ismi Rizqa Lina
Department of Data Science, University of Insan Cita Indonesia, Jakarta, Indonesia. e-mail: ismi@uici.ac.id

Wuryansari Muharini Kusumawinahyu
Department of Mathematics, Faculty of Mathematics and Natural Sciences, University of Brawijaya, Malang, Indonesia. e-mail: wmuharini@ub.ac.id

AMS subject classifications (2020): Primary 76-10, 65M08.

Keywords: Two-layer shallow water; Finite volume method; QUICK scheme; Explicit scheme.

1 Introduction

A flow with constant density where the horizontal scale is much bigger than the vertical scale is called shallow water. Shallow water is a thin layer of fluid of constant density in hydrostatic equilibrium, surrounded by a solid surface below and a free surface above, and is likely in contact with another low-inertia fluid. The simplest and most useful models in geophysical fluid dynamics are single-layer models because they can explain the rotational effect with a simple structure without considering the effects of complex stratification. We can investigate the effect of the layers by adding a layer to them. The two-layer in the geometry of shallow water is not only a layered liquid model, but also it is an excellent model of a wide range of ocean and atmospheric phenomena; see [23].

One way to depict fluid flow is as a multilayer flow, where one layer flows on top of another. This difference in a layer is caused by temperature differences, which lead to different densities. Ocean water can be viewed as a fluid with two layers. The temperature of the water in the first layer, which is sunlight-exposed water, is greater than the temperature of the second layer, which is below the first layer; see [23].

Shallow water flow in two layers, especially its exact and numerical solution, is an interesting subject to be studied. They particularly find the best-governing equation that can describe physical phenomena.

The two-layer fluid flow model over the inclined and flat bottom was investigated in [17]. The model utilized a non-Newtonian fluid. They used the flow of mud in the water to illustrate the model. Then, this model was solved analytically and numerically. The McCormack's scheme was used to obtain the numerical solution since this scheme is suitable for systems that meet the hyperbolic conservation law, and it has second-order accuracy. The numerical simulation of stratified fluid, supposed to be two-layer shallow water flows through a channel with irregular geometry, was studied in [5]. They proposed extensions of the Q-schemes of van Leer and Roe to get the numerical solution of the model. The central upwind was used to solve the two-layer shallow water equations [10].

Furthermore, the fluid flow model with inclined and uneven bottom was developed in [25]. In this model, it was assumed that the fluid had a single layer and flowed in horizontal and vertical directions. The constructed model was completed analytically and numerically using the finite difference method by applying periodic boundary conditions. A numerical scheme stability analysis was also performed using the Fourier analysis developed by Von

Neumann. The stability analysis revealed that the flow was stable, which means that the error of a numerical solution will not grow infinitely with time.

The analytical and numerical methods were developed quickly to find the best solution for the model. The analytical method is the best method to find the solution of the mathematical model since this method does not have any errors that appear from the truncation of the method itself. Unfortunately, analytical methods sometimes require some complex calculations and are very difficult to solve. To handle this problem, numerical methods can be used to approximate the solutions of the governing equations. The classical finite difference method, in which derivatives are approximated by finite difference, is a well-known method to approximate the solution. Discontinuities become a problem when they occur in the solution. Such a method can be used as well with certain treatments near discontinuities in the solution. The finite volume method (FVM) is a method based on an integral form instead of the differential equation. Rather than pointwise approximations at grid points, the FVM breaks the domain into grid cells and approximates the total integral over grid cells. This integral is divided by the volume of the cell. These values are changed at each step by the flux passing through the edges of the grid cells. The main problem is to determine the proper numerical flux function that properly approximates the correct flux [11].

The study of the two-layer shallow water system (flat bottom) analytically and numerically was investigated in [3]. For an analytical solution, they considered a relaxation approach, which offers greater decoupling and accessible eigenstructures. The numerical solution of a two-layer shallow water equation using the FVM with an explicit scheme was investigated in [7]. Then, Swartenbroekx, Zech, and Soares-Frazao [22] predicted the bed load transport induced by dam break waves by proposing two-dimensional (2D) two-layer shallow water equations, where the upper layer was made of clear water, and the lower layer was made of a dense mixture of water and moving grains. The governing equation of the model was solved by applying Harten–Lax–Van Lear Finite Volume Scheme and a well-balanced Roe method [9]. Recently, Muhammad [13] simulated the flowing fluid using the FVM via OpenFOAM.

On the other hand, the flat-bottomed shallow hydraulic analogy of the Gibraltar Strait was developed with wind stress in mind [6]. In this model, the flow was assumed to be quasi-horizontal. The y component was ignored ,and the density was uniform in each layer. The model was numerically solved using the FVM based on the Roe-Riemann scheme approach. Chiapolino and dan Saure [8] also derived a two-layer shallow water flow model with some limitations by ignoring the vertical velocity component and assuming that the velocity is uniform across the cross-sections of each layer. This model was solved numerically using the Godunov scheme, which is an approximation of the Riemann problem at each cell boundary. A method for the numerical solution of the nonlinear regularized long wave equation

was studied in [20]. They [19] proposed an implicit method for the time discretization based on the θ-weighted and finite difference methods, while the spatial discretization is described with the help of the finite difference scheme derived from the local radial basis function method (the local collocation method). The localized meshless collocation method was studied in [16, 15]. A new weighted essentially non-oscillatory (WENO), a finite difference Hermite RBF-WENO scheme, and a modified weighted essentially non-oscillatory (WENO) scheme are procedures for solving hyperbolic conservation laws were studied in [18, 1, 2].

The study conducted in [4] obtained a numerical solution of the advection-reaction-diffusion equation using the FVM. In the FVM, the integral form is used to discretize the equation. The discretization of space used the Quadratic Upwind Interpolation for Convective Kinematics (QUICK) scheme, while the discretization of time used an explicit, implicit, or Crank Nicolson scheme.

In this paper, we solve numerically a mathematical formula of a two-layer shallow water flow with slope and uneven bottom by using the FVM with a QUICK scheme to discretize the space domain and explicit scheme to discretize the time domain. The new one in this article is how to find a solution of the model when the bottom of the shallow water is not flat. The FVM is used to solve the model because this method is suitable for computational fluid dynamics problems. The QUICK method is chosen since it is a third-order scheme that provides good accuracy, and this scheme have been proven its numerical stability. The advantages of this method are that the main coefficients are positive and satisfy the requirements for conservativeness, boundedness, and transportiveness [24]. A numerical simulation of the solutions is presented to describe the behavior of this model.

The organization of this work is as follows: Section 2 derives the governing equations of the models in steady and unsteady states. In Section 3, the FVM is constructed as a numerical solution with the QUICK scheme for discretization. Section 4 provides numerical simulations of the shallow water flows with several given parameters. Finally, Section 5 ends up with a brief conclusion.

2 Governing equations

The governing equation of the two-layer shallow water with slope and the uneven bottom is described in Figure 1. It is a non-dimensional form after deriving from continuity and momentum equations of shallow water equations. For detailed calculating, revisit [12]. The equation reads as follows:

$$\frac{\partial}{\partial t}\left(\eta_1 - \eta_2\right) + \frac{\partial}{\partial x}\left(u_1(\eta_1 - \eta_2)\right) = 0, \tag{1}$$

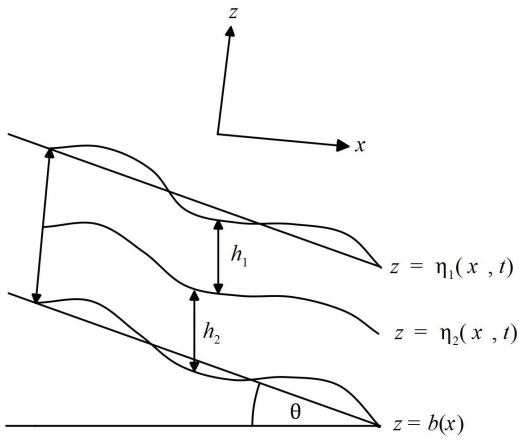


Figure 1: Two-layer shallow water flow with incline and uneven bottom [12].

$$\frac{\partial \eta_2}{\partial t} + \frac{\partial}{\partial x}\left(u_2(\eta_2 - b)\right) = 0, \tag{2}$$

$$\frac{\partial u_1}{\partial t} + u_1 \frac{\partial u_1}{\partial x} = -g\cos\theta \frac{\partial \eta_1}{\partial x} + g\sin\theta, \tag{3}$$

$$\frac{\partial u_2}{\partial t} + u_2 \frac{\partial u_2}{\partial x} = (g' - g)\cos\theta \frac{\partial \eta_1}{\partial x} - g'\cos\theta \frac{\partial \eta_2}{\partial x} + g\sin\theta. \tag{4}$$

We introduce the non-dimensional variables as follows:

$$x = L\widetilde{x}, \quad z = H\widetilde{z}, \quad t = \frac{L}{U}\widetilde{t}, \quad \eta_1 = H\widetilde{\eta}_1,$$
$$\eta_2 = H\widetilde{\eta}_2, \quad b = H\widetilde{b}, \quad q_1 = UH\widetilde{q}_1, \quad q_2 = UH\widetilde{q}_2,$$

and

$$U = (Hg')^{\frac{1}{2}}, \quad L = \frac{U^2}{g}, \quad \gamma = \frac{g'}{g},$$

where U is a typical velocity, L is typical length, and T is a typical time scale, and H is mean height. The two-layer shallow water with slope and uneven bottom in non-dimensional form is

$$\frac{\partial \eta_1}{\partial t} - \frac{\partial \eta_2}{\partial t} + \frac{\partial u_1\eta_1}{\partial x} - \frac{\partial u_1\eta_2}{\partial x} = 0, \tag{5}$$

$$\frac{\partial \eta_2}{\partial t} + \frac{\partial u_2\eta_2}{\partial x} - \frac{\partial u_2 b}{\partial x} = 0, \tag{6}$$

$$\frac{\partial u_1}{\partial t} + \frac{\partial}{\partial x}\left(\frac{u_1^2}{2}\right) = -\frac{\cos\theta}{\gamma}\frac{\partial \eta_1}{\partial x} + \sin\theta, \tag{7}$$

$$\frac{\partial u_2}{\partial t} + \frac{\partial}{\partial x}\left(\frac{u_2^2}{2}\right) = \left(1 - \frac{1}{\gamma}\right)\cos\theta\frac{\partial \eta_1}{\partial x} - \cos\theta\frac{\partial \eta_2}{\partial x} + \sin\theta. \tag{8}$$

We call the bottom layer as layer one, and the upper layer as layer two. We introduce the variables used in each layer, $p_1(x,t)$ and $p_2(x,t)$ are pressure forces in layers one and two. The upper layer has a pressure force $p_1(x,t)$, which is lower than the pressure force at the bottom layer $p_2(x,t)$. Then, $\eta_1(x,t)$ is the free-surface movement vertically while $\eta_2(x,t)$ is the distance between layers one and two's interface height, where $\eta_2(x,t) < z < \eta_1(x,t)$ and $b(x) < z < \eta_2(x,t)$. The thickness of the water columns is $h_1(x,t)$ and $h_2(x,t)$. The uneven bottom topography is $b(x)$, where, in general, $h_2(x,t) = \eta_2(x,t) - b(x)$, for the flat bottom $\eta_2(x,t) = h_2(x,t)$, and θ is the slope of the bottom surface. Moreover, $\gamma = g'/g$, where g is the gravitational force, and $g' = g(\frac{\rho_2 - \rho_1}{\rho_1})$ is the reduced gravity, while ρ_2 and ρ_1 are density in layers one and two. The proposed model, (1)–(4), is a powerful model of many geophysically interesting phenomena as well as being physically realizable in the laboratory [23].

2.1 Steady-state two-layer shallow water flow

In this part, shallow water is assumed to be in a steady state. This means at every point, the mass fluxes over the depth and energy are constant [14]. It is assumed that the height and water debit are constant with respect to time [21]. Equations (5)–(8) become

$$\frac{\partial u_1\eta_1}{\partial x} - \frac{\partial u_1\eta_2}{\partial x} = 0, \tag{9}$$

$$\frac{\partial u_2\eta_2}{\partial x} - \frac{\partial u_2 b}{\partial x} = 0, \tag{10}$$

$$0 = -\frac{\cos\theta}{\gamma}\frac{\partial \eta_1}{\partial x} + \sin\theta, \tag{11}$$

$$0 = \left(1 - \frac{1}{\gamma}\right)\cos\theta\frac{\partial \eta_1}{\partial x} - \cos\theta\frac{\partial \eta_2}{\partial x} + \sin\theta, \tag{12}$$

where $x \in [0, K]$ in which K is a constant at the boundary space domain. The equation system (9)–(12) is a function of spatial variable, such that the initial conditions for the equation system (9)–(12) are given as follows:

$$\eta_1(x,0) = \tan\theta x + A\sin(3x), \tag{13}$$

$$\eta_2(x,0) = \tan\theta x + A\sin(3x) - 3, \tag{14}$$

$$b(x) = \tan\theta x + K\tan\theta + A(\cos(5x) - \sin(3x)) - 4, \tag{15}$$

and the boundary conditions are given by

$$\eta_1(0) = \tan\theta(0) + A\sin(3(0)), \tag{16}$$
$$\eta_2(0) = \tan\theta(0) + A\sin(3(0)) - 3, \tag{17}$$
$$\eta_1(K) = \tan\theta(K) + A\sin(3(K)), \tag{18}$$
$$\eta_2(K) = \tan\theta(K) + A\sin(3(K)) - 3, \tag{19}$$
$$\frac{\partial\eta_1}{\partial x} = \frac{\partial\eta_2}{\partial x} = \tan\theta\gamma, \tag{20}$$

where A is wave amplitude.

The steady state of (9)–(12) can be applied to river at rest conditions. It seems that the river consists of two-layers. The temperature of the first layer will be lower than the temperature of the second layer when the river is exposed to sunshine. As a result, the first layer pressure is lower than the second layer pressure. The layers compete with one another when the interface between the two layers varies from position to position (wavy). Similarly, on an uneven base, generally, topography and bottom layer will exert forces on one another. Because it is in a steady state, the bottom shape is unevenly followed by the first and second levels. We will show this in numerical simulations.

2.2 Unsteady-state two-layer shallow water flow

For an unsteady state two-layer shallow water flow, it is assumed that the flow of shallow water at the upper free-surface, the gravitational force, is imposed on the fluid as the restoring force. Because we consider hydrostatic balance, the changes of surface velocity, and height are quite small around the depth of the h surface wave at constant conditions [23]. The depth of the fluid is as much as the displacement or movement of the free surface. We linearize (5)–(8) by defining new variables to describe the displacement of the fluid at rest; that is,

$$h_1(x,t) = H_1 + h_1'(x,t) = H_1 + (\eta_1'(x,t) - \eta_2'(x,t)), \tag{21}$$
$$h_2(x,t) = H_2 + h_2'(x,t) = H_2 + (\eta_2'(x,t) - b(x)), \tag{22}$$
$$u_1(x,t) = u_1'(x,t), \tag{23}$$
$$u_2(x,t) = u_2'(x,t). \tag{24}$$

Equations (21)–(24) are substituted into (5)–(8). Then, the variable following the derivation of steady state two-layer shallow water flow is non-dimensionalized. We obtain the following equations:

$$\frac{\partial\eta_1}{\partial t} - \frac{\partial\eta_2}{\partial t} = 0, \tag{25}$$
$$\frac{\partial\eta_2}{\partial t} + \frac{\partial u_2}{\partial x} - \frac{\partial u_2 b}{\partial x} = 0, \tag{26}$$

$$\frac{\partial u_1}{\partial t} = -\frac{\cos\theta}{\gamma}\frac{\partial \eta_1}{\partial x} + \sin\theta, \tag{27}$$

$$\frac{\partial u_2}{\partial t} = \left(1 - \frac{1}{\gamma}\right)\cos\theta\frac{\partial \eta_1}{\partial x} - \cos\theta\frac{\partial \eta_2}{\partial x} + \sin\theta, \tag{28}$$

where $x \in [0, K], t \in [0, L]$, with the initial conditions of the flow is stationary and the speed of the water remains along the channel

$$\eta_1(x, 0) = \tan\theta x + A\sin(3x), \tag{29}$$

$$\eta_2(x, 0) = \tan\theta x + A\sin(3x) - 3, \tag{30}$$

$$b(x) = \tan\theta x + K\tan\theta + A(\cos(5x) - \sin(3x)) - 4, \tag{31}$$

$$u_1(x, 0) = u_2(x, 0) = 0, \tag{32}$$

and the boundary condition is assumed that the flow is not moving, so that the velocity is equal to zero. The boundary condition of unsteady state two-layer shallow water flow is periodic functions as follows:

$$\eta_1(0, t) = \tan\theta(0) + A\sin(3(0)), \tag{33}$$

$$\eta_2(0, t) = \tan\theta(0) + A\sin(3(0)) - 3, \tag{34}$$

$$\eta_1(K, t) = \tan\theta(K) + A\sin(3(K)), \tag{35}$$

$$\eta_2(K, t) = \tan\theta(K) + A\sin(3(K)) - 3, \tag{36}$$

$$\frac{\partial \eta_1(0, t)}{\partial x} = \frac{\partial \eta_2(0, t)}{\partial x} = \tan\theta\gamma + 3A\cos(3(0)), \tag{37}$$

$$u_1(0, t) = u_2(0, t) = 0, \tag{38}$$

where $x > 0$ and $t > 0$.

3 Finite volume method (FVM)

The FVM is a technique using integral form rather than differential equations. Pointwise estimates at grid points are preferred. The domain is divided into grid cells by the FVM, which then approximates the entire integral over the grid cells. This integral is divided by the volume of the cells. The value changes at each step depending on the flow passing through the edges of the grid cells. The main problem is determining the appropriate numerical flow function that is directly close to the true flow [11].

By following the procedure in [24], we integrate the governing equations for steady and unsteady state, (9)–(12) and (25)–(28), respectively, along control volume and time. Furthermore, we use QUICK scheme to discretize space domain and explicit scheme to discretize time domain.

Here, we summarize the procedure of the proposed method as follows:

1. **Grid generation**
   The first step in the FVM is to divide the domain into discrete control volumes. Let us place a number of nodal points in the space between A and B. The boundaries (or faces) of control volumes are positioned mid-way between adjacent nodes. Thus each node is surrounded by a control volume or cell. It is common practice to set up control volumes near the edge of the domain in such a way that the physical boundaries coincide with the control volume boundaries.

2. **Discretisation**
   The key step of the FVM is the integration of the governing equation (or equations) over a control volume to yield a discretized equation at its nodal point P.

3. **Solution of equations**
   The resulting system of linear algebraic equations is then solved to obtain the distribution of the solution at nodal points.

3.1 Steady-state FVM

To solve (9)–(12) by using the FVM, first, we integrate (10) with respect to Control Volume (CV) as follows:

$$\int_{CV} \left(\frac{\partial u_2 \eta_2}{\partial x} - \frac{\partial u_2 b}{\partial x} \right) dV = 0. \tag{39}$$

This is the equation related to the first layer of shallow water flow. Since the model of two-layer shallow water flow has one dimension in space, the CV can be described as in [24]. Then (39) becomes

$$A\left(u_{2e} - u_{2w}\right)\left(\eta_{2e} - \eta_{2w}\right) = A\left(u_{2e} - u_{2w}\right)\left(b_e - b_w\right), \tag{40}$$

where index e is the east face and w is the west face. By applying QUICK discretization, we transform the control face to be nodal and after simplifying (40). Then we get

$$\frac{1}{8}\eta_{2WW} - \frac{7}{8}\eta_{2W} + \frac{3}{8}\eta_{2P} + \frac{3}{8}\eta_{2E} = \frac{1}{8}b_{WW} - \frac{7}{8}b_W + \frac{3}{8}b_P + \frac{3}{8}b_E. \tag{41}$$

Equation (41) can be written as a linear equation of the system by substituting $i = 2, 3, \ldots, N-1$ and applying initial and boundary condition such that we have

$$\begin{bmatrix} \frac{3}{8} & \frac{3}{8} & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ \frac{-7}{8} & \frac{3}{8} & \frac{3}{8} & 0 & \cdots & 0 & 0 & 0 & 0 \\ \frac{1}{8} & \frac{-7}{8} & \frac{3}{8} & \frac{3}{8} & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & \frac{1}{8} & \frac{-7}{8} & \frac{3}{8} & \frac{3}{8} \\ 0 & 0 & 0 & 0 & \cdots & 0 & \frac{1}{8} & \frac{-7}{8} & \frac{3}{8} \end{bmatrix} \begin{bmatrix} \eta_{2\,2} \\ \eta_{2\,3} \\ \eta_{2\,4} \\ \vdots \\ \eta_{2\,K-2} \\ \eta_{2\,K-1} \end{bmatrix} = \begin{bmatrix} F_2 \\ F_3 \\ F_4 \\ \vdots \\ F_{K-2} \\ F_{K-1} \end{bmatrix}, \tag{42}$$

with

$$\begin{bmatrix} F_2 \\ F_3 \\ F_4 \\ \vdots \\ F_{K-2} \\ F_{K-1} \end{bmatrix} = \begin{bmatrix} -\frac{6}{8}b_{i-1} + \frac{1}{8}\Delta x T1 + \frac{3}{8}b_i + \frac{3}{8}b_{i+1} + \frac{6}{8}\eta_{2\,i-1} - \frac{1}{8}\Delta x \tan(\theta)\gamma \\ \frac{1}{8}b_{i-2} - \frac{7}{8}b_{i-1} + \frac{3}{8}b_i + \frac{3}{8}b_{i+1} - \frac{1}{8}\eta_{2\,i-2} \\ \frac{1}{8}b_{i-2} - \frac{7}{8}b_{i-1} + \frac{3}{8}b_i + \frac{3}{8}b_{i+1} \\ \vdots \\ \frac{1}{8}b_{i-2} - \frac{7}{8}b_{i-1} + \frac{3}{8}b_i + \frac{3}{8}b_{i+1} \\ \frac{1}{8}b_{i-2} - \frac{7}{8}b_{i-1} + \frac{3}{8}b_i + \frac{3}{8}b_{i+1} - \frac{3}{8}\eta_{2\,i+1} \end{bmatrix},$$

where $T_1 = \tan\theta + A(5\sin(-5(0)) - 3\cos(-3x(0)))$. The methods for solving a system of linear equations can be used to determine the value of η_{2i}, for example, Thomas's algorithm.

Furthermore, similar to the first layer, the discretization equation of the second layer is obtained by integrating (9) with respect to CV

$$\int_{CV} \left(\frac{\partial u_1 \eta_1}{\partial x} - \frac{\partial u_1 \eta_2}{\partial x} \right) dV = 0. \tag{43}$$

Discretization QUICK is applied to (43) by following the procedure in [24] and applying initial and boundary conditions. Then, we arrive at a system of linear equations as follows:

$$\begin{bmatrix} \frac{3}{8} & \frac{3}{8} & 0 & 0 & \cdots & 0 & 0 & 0 & 0 \\ \frac{-7}{8} & \frac{3}{8} & \frac{3}{8} & 0 & \cdots & 0 & 0 & 0 & 0 \\ \frac{1}{8} & \frac{-7}{8} & \frac{3}{8} & \frac{3}{8} & \cdots & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \cdots & \frac{1}{8} & \frac{-7}{8} & \frac{3}{8} & \frac{3}{8} \\ 0 & 0 & 0 & 0 & \cdots & 0 & \frac{1}{8} & \frac{-7}{8} & \frac{3}{8} \end{bmatrix} \begin{bmatrix} \eta_{1\,2} \\ \eta_{1\,3} \\ \eta_{1\,4} \\ \vdots \\ \eta_{1\,K-2} \\ \eta_{1\,K-1} \end{bmatrix} = \begin{bmatrix} G_2 \\ G_3 \\ G_4 \\ \vdots \\ G_{K-2} \\ G_{K-1} \end{bmatrix}, \tag{44}$$

with

$$\begin{bmatrix} G_2 \\ G_3 \\ G_4 \\ \vdots \\ G_{K-2} \\ G_{K-1} \end{bmatrix} = \begin{bmatrix} -\frac{6}{8}\eta_{2i-1} + \frac{1}{8}\Delta x\tan(\theta)\gamma + \frac{3}{8}\eta_{2i} + \frac{3}{8}\eta_{2i+1} + \frac{6}{8}\eta_{2i-1} - \frac{1}{8}\Delta x\tan(\theta)\gamma \\ \frac{1}{8}\eta_{2i-2} - \frac{7}{8}\eta_{2i-1} + \frac{3}{8}\eta_{2i} + \frac{3}{8}\eta_{2i+1} - \frac{1}{8}\eta_{1i-2} \\ \frac{1}{8}\eta_{2i-2} - \frac{7}{8}\eta_{2i-1} + \frac{3}{8}\eta_{2i} + \frac{3}{8}\eta_{2i+1} \\ \vdots \\ \frac{1}{8}\eta_{2i-2} - \frac{7}{8}\eta_{2i-1} + \frac{3}{8}\eta_{2i} + \frac{3}{8}\eta_{2i+1} \\ \frac{1}{8}\eta_{2i-2} - \frac{7}{8}\eta_{2i-1} + \frac{3}{8}\eta_{2i} + \frac{3}{8}\eta_{2i+1} - \frac{3}{8}\eta_{1i+1} \end{bmatrix}.$$

The value of η_{1i} can be obtained by the methods for solving system of linear equations.

3.2 Unsteady-state FVM

For unsteady-state two-layer shallow water flow, we integrated (25)–(28) with respect to CV and time. The procedure to get the integration (25)–(28) with respect to CV in the unsteady state, shallow water flow is similar to the procedure to calculate the integral of the governing equation in the steady state. We apply QUICK discretization. The difference is we integrate the governing equation with respect to time, and apply an explicit scheme for discretization. First, we integrate (25) to yield

$$\int_t^{t+\Delta t}\int_{CV}\left(\frac{\partial\eta_1}{\partial t} - \frac{\partial\eta_2}{\partial t}\right)dV\,dt = 0. \tag{45}$$

The result of the integration of (45) is

$$(\eta_{1P} - {\eta_1}_P^0)A\Delta x - (\eta_{2P} - {\eta_2}_P^0)A\Delta x = 0, \tag{46}$$

where $\eta_{1,P}$ and $\eta_{2,P}$ are the height of surface at node P and time $t+\Delta t$, $\eta_{1,P}^0$ and $\eta_{2,p}^0$ are the height of surface at node P and time t. By simplifying (46) and writing the index P becomes i, we get the height of the first layer

$${\eta_1}_i^{n+1} = {\eta_1}_i^n + {\eta_2}_i^{n+1} - {\eta_2}_i^n, \tag{47}$$

where the subscribe n denotes time t, and $n+1$ denotes time $t+\Delta t$. We apply the procedure in [24] for the rest of (26)–(28), so that we obtain

$${\eta_2}_i^{n+1} = {\eta_2}_i^n - \frac{\Delta t}{\Delta x}\left(\frac{1}{8}{u_2}_{i-2}^n - \frac{7}{8}{u_2}_{i-1}^n + \frac{3}{8}{u_2}_i^n + \frac{3}{8}{u_2}_{i+1}^n\right)\left(1 - \left(\frac{1}{8}b_{i-2} - \frac{7}{8}b_{i-1} + \frac{3}{8}b_i + \frac{3}{8}b_{i+1}\right)\right). \tag{48}$$

$$u_1{}_i^{n+1} = u_1{}_i^{n} - \frac{\Delta t}{\Delta x}\frac{\cos\theta}{\gamma}\left(\frac{1}{8}\eta_1{}_{i-2}^{n} - \frac{7}{8}\eta_1{}_{i-1}^{n} + \frac{3}{8}\eta_1{}_{i}^{n} + \frac{3}{8}\eta_1{}_{i+1}^{n}\right) + \sin\theta\Delta t. \tag{49}$$

$$u_2{}_i^{n+1} = u_2{}_i^{n} + \frac{\Delta t}{\Delta x}\left(1 - \frac{1}{\gamma}\right)\cos\theta\left(\frac{1}{8}\eta_1{}_{i-2}^{n} - \frac{7}{8}\eta_1{}_{i-1}^{n}\frac{3}{8}\eta_1{}_{i}^{n} + \frac{3}{8}\eta_1{}_{i+1}^{n}\right) - \frac{\Delta t}{\Delta x}\cos\theta\left(\frac{1}{8}\eta_2{}_{i-2}^{n} - \frac{7}{8}\eta_2{}_{i-1}^{n} + \frac{3}{8}\eta_2{}_{i}^{n} + \frac{3}{8}\eta_2{}_{i+1}^{n}\right) + \sin\theta\Delta t. \tag{50}$$

The explicit solver is easily programmed, yet the time step t must be limited to stabilize the scheme. Experience has shown that we can simply enforce the Von Neumann criteria for the single step algorithm,

$$\left|\frac{\Delta t}{\Delta x}\right| < \frac{1}{2}. \tag{51}$$

4 Numerical simulation

Numerical simulation aims to illustrate movement of shallow water flows based on the numerical scheme obtained in the previous section. Numerical simulations are performed under the environment of MATLAB R2018a based on MacOS MOJAVE version 10.14.6, which the processor is 1.6 GHz Intel Core i5. Numerical solutions of steady-state of two-layer shallow water flow are calculated at $x \in [0, 5]$ by taking $\Delta x = 0.1$ and $\gamma = 1$ [17]. Figure 2 depicts two shallow water layers at rest, where the water velocity is zero and there is no time-dependent flow movement. Changes at the bottom surface are followed by depth changes at the first and second chamber levels. When we increase the slope, the depth of the second layer is higher than the first layer. The two-layer shallow water flow conditions remain stable for each slope of (θ). Negative sign of slope has physical interpretation, that is, water flows from the left side to the right side. Figure 3 shows the velocity profiles of steady-state of two-layer shallow water flow at $t = 0.025$.

Table 1: Comparison of the CPU run times of the proposed method depends on the slopes.

$\theta = -10$	$\theta = -30$	$\theta = -60$
1,376413 second	1,536349 second	1,544937 second

Furthermore, the numerical simulation for unsteady-state two-layer shallow water flow can be seen in Figures 4 and 5. The parameter (γ), slope (θ), and space domain for the simulation are chosen with the same values for the steady-state. The time domain is $t \in [0, 0.1]$ with $\Delta t = 0.001$. The simulated water wave height and velocity are calculated when $t = 0.025$ and $t = 0.1$. The height profile of unsteady-state two-layer shallow water flow at $t = 0.025$

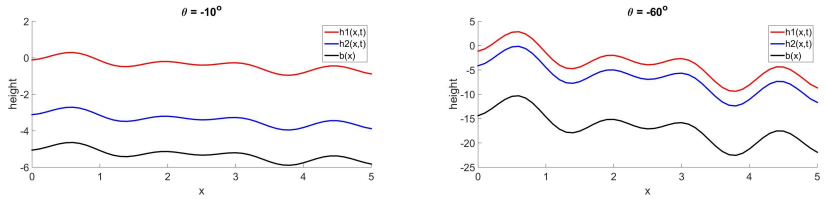


Figure 2: The numerical solution of steady-state of two-layer shallow water flow with $\theta = 10^0$, $\theta = 30^0$, and $\theta = 60^0$.

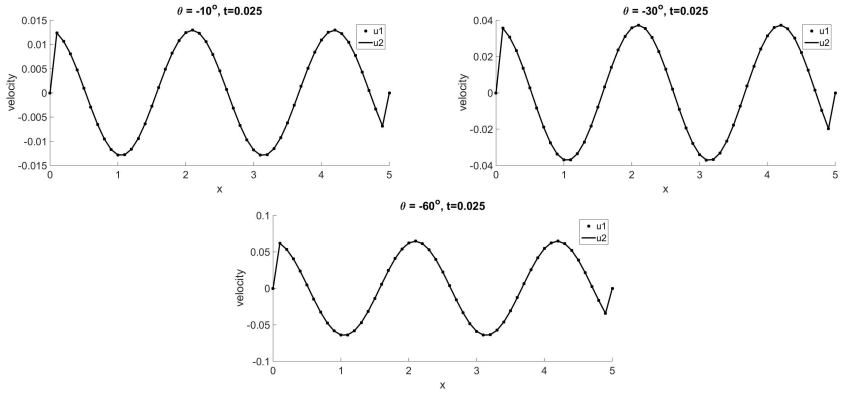


Figure 3: The velocity profiles of steady-state of two-layer shallow water flow at $t = 0.025$.

is shown in Figure 4, while the velocity profiles of unsteady-state of two-layer shallow water flow at $t = 0.025$ is shown in Figure 5.

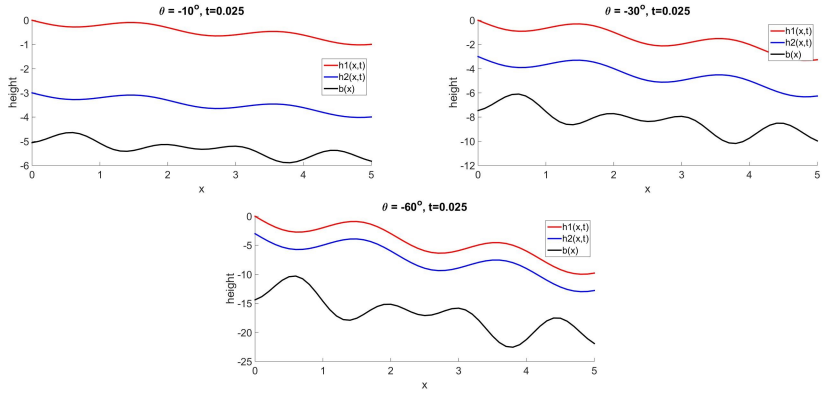


Figure 4: The height profiles of unsteady-state two-layer shallow water flow at $t = 0.025$.

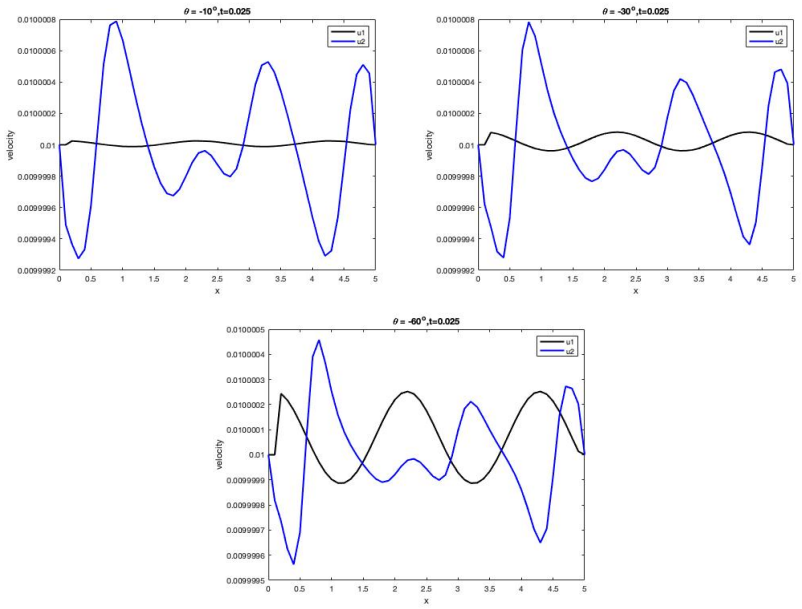


Figure 5: The velocity profiles of unsteady-state of two-layer shallow water flow at $t = 0.025$.

The numerical solution for unsteady shallow water produces instability (see Figure 5). It is because there is a compressive force between the layers. Similarly, if the liquid's bottom is not flat, the bottom layer and topography tend to push against one another. This type of force, known as form drag. Moreover, in Figure 5, when the slope is increased, the depth of the second layer is higher than the first layer, and the second layer has a more wavy shape than the first layer. In Figure 5, the flow has bigger velocity when the slope of bottom layer is increased. From Table 2, the velocity at layer one increases when the slope is increased. Similarly, the velocity at layer two increases when the slope is increased see Table 3. The last row in each Tables 2 and 3 shows the rate of velocity u_1 and u_2. The CPU run times of the proposed method is written in Table 1.

Table 2: Comparison of velocity u_1 depends on the slopes.

u_1 at $\theta = -10$	u_1 at $\theta = -30$	u_1 at $\theta = -60$
0,01	0,01	0,01
0,01	0,01	0,01
0,010000047	0,010000156	0,010000486
0,010000042	0,01000014	0,010000435
0,010000035	0,010000115	0,010000356
0,010000025	0,010000084	0,010000257
0,010000015	0,010000049	0,010000148
0,010000004	0,010000013	0,01000004
0,009999994	0,00999998	0,009999941
0,009999986	0,009999953	0,009999861
0,00999998	0,009999934	0,009999804
0,009999976	0,009999924	0,009999775
0,009999977	0,009999924	0,009999776
0,00999998	0,009999934	0,009999806
0,009999986	0,009999954	0,009999864
0,009999995	0,009999982	0,009999946
0,010000005	0,010000015	0,010000046
0,010000015	0,010000051	0,010000154
0,010000026	0,010000086	0,010000263
0,010000035	0,010000117	0,010000361
0,010000043	0,010000141	0,010000439
0,010000047	0,010000157	0,010000489
0,010000049	0,010000162	0,010000504
0,010000047	0,010000156	0,010000485
0,010000042	0,010000139	0,010000431
0,010000034	0,010000114	0,010000351
0,010000025	0,010000082	0,010000251
0,010000014	0,010000047	0,010000142
0,010000003	0,010000011	0,010000034
0,009999993	0,009999979	0,009999936
0,009999985	0,009999952	0,009999857
0,009999979	0,009999933	0,009999801
0,009999976	0,009999923	0,009999774
0,009999977	0,009999924	0,009999776
0,00999998	0,009999935	0,009999809
0,009999986	0,009999956	0,009999868
0,009999995	0,009999984	0,009999951
0,010000005	0,010000017	0,010000051
0,010000016	0,010000053	0,01000016
0,010000027	0,010000088	0,010000268
0,010000036	0,010000118	0,010000366
0,010000043	0,010000142	0,010000442
0,010000047	0,010000157	0,01000049
0,010000049	0,010000162	0,010000504
0,010000047	0,010000155	0,010000483
0,010000042	0,010000138	0,010000428
0,010000034	0,010000112	0,010000346
0,010000024	0,01000008	0,010000245
0,010000013	0,010000045	0,010000136
0,010000003	0,010000009	0,010000028
0,01	0,01	0,01
0,510000685	0,51000228	0,510007164

Table 3: Comparison of velocity u_2 depends on the slopes.

u_2 at $\theta = -10$	u_2 at $\theta = -30$	u_2 at $\theta = -60$
0,01	0,01	0,01
0,009999236	0,009999431	0,009999728
0,009999233	0,009999349	0,009999664
0,009999146	0,00999918	0,009999542
0,009999225	0,009999147	0,009999474
0,009999545	0,009999439	0,009999622
0,010000052	0,010000062	0,01000004
0,010000542	0,01000064	0,010000417
0,010000821	0,010000851	0,010000502
0,010000857	0,010000764	0,010000407
0,010000736	0,010000576	0,010000283
0,01000055	0,01000039	0,010000182
0,010000354	0,010000236	0,010000106
0,010000169	0,010000109	0,010000049
0,01	0,010000001	0,010000001
0,009999849	0,009999904	0,009999958
0,009999721	0,009999815	0,009999917
0,009999628	0,00999974	0,00999988
0,009999584	0,009999691	0,009999852
0,009999598	0,009999681	0,009999842
0,009999659	0,009999714	0,009999853
0,009999739	0,00999977	0,009999879
0,009999804	0,009999822	0,009999904
0,009999833	0,009999847	0,009999917
0,009999824	0,009999841	0,009999915
0,009999795	0,009999816	0,009999902
0,00999977	0,009999793	0,009999889
0,009999785	0,009999802	0,009999893
0,00999987	0,00999988	0,009999935
0,01000004	0,010000044	0,010000027
0,010000262	0,010000256	0,010000145
0,010000465	0,010000426	0,01000023
0,01000058	0,010000443	0,010000214
0,010000494	0,010000344	0,010000159
0,010000353	0,010000229	0,010000102
0,010000189	0,010000116	0,01000005
0,010000014	0,010000006	0,010000002
0,009999831	0,009999893	0,009999953
0,009999641	0,009999768	0,009999897
0,009999446	0,00999962	0,009999826
0,009999266	0,00999945	0,009999736
0,009999147	0,009999283	0,009999632
0,009999163	0,009999199	0,009999553
0,009999371	0,009999324	0,009999592
0,009999739	0,009999699	0,009999808
0,010000128	0,010000137	0,010000082
0,010000397	0,010000409	0,010000239
0,0100005	0,010000471	0,010000259
0,010000565	0,010000492	0,010000257
0,01	0,01	0,01
0,509997101	0,509997386	0,509998561

5 Conclusion

In this research, we offered a numerical solution for a two-layer shallow water flow with slope and uneven bottom for both steady-state and unsteady-state. The QUICK and explicit schemes were used to determine the numerical solution through the FVM. The steady-state numerical solution yielded stable results, while unsteady-state numerical solutions produced unstable results. The momentum that can be added to or taken from a flow, is what leads to this instability. This happened when there are differences in the position or surface wave shape between the first and second layers. The layers exert a compressive force on one another. Similar to this, if the fluid's bottom is not flat, the bottom layer and the topography will generally exert forces on one another. Moreover, the simulations yielded the larger slope resulted in higher velocity and higher depth at second layer.

Acknowledgments

The authors would like to express their gratitude to Prof. Sukir Maryanto for valuable comments of the research, and LPPM of Brawijaya University.

References

[1] Abedian, R. *A modified high-order symmetrical WENO scheme for hyperbolic conservation laws*, Int. J. Nonlinear Sci. Numer. Simul. 2022.

[2] Abedian, R. *A finite difference Hermite RBF-WENO scheme for hyperbolic conservation laws*, Int. J. Numer. Methods Fluids, 94(6) (2022), 583–607.

[3] Abgrall, R. and Karni, S. *Two-layer shallow water system: A relaxation approach*, SIAM J. Sci. Comput. 31(3) (2009), 1603–1627.

[4] Arachchige, J.P. and Pettet. G.J. *A finite volume method with linearisation in time for the solution of advection–reaction–diffusion systems*, Appl. Math. Comput. 231 (2014), 445–462.

[5] Castro, M.J., Garcia-Rodriguez, J.A., Gonzalez-Vida, J.M., Macias, J., Pares, C. and Vazquez-Cendon, M.E. *Numerical simulation of two-layer shallow water flows through channels with irregular geometry*, J. Comput. Phys. 195 (2004), 2002–235.

[6] Chakir, M., Ouazar, D., and Taik, A. *Roe scheme for two-layer shallow water equations: Application to the strait of Gibraltar*, Math. Model. Nat. Phenom. 4(5) (2009), 114–127.

[7] Chen, S C., and Peng, S.H. *Two-dimensional numerical model of two-layer shallow water equations for confluence simulation*, Adv. Water Resour. 29 (2006), 1608–1617.

[8] Chiapolino, A. and dan Saurel, R. *Models and methods for two-layer shallow water flows*, J. Comput. Phys. 371 (2018), 1043–1066.

[9] Cristo, C.D., Greco, M., Iervolino, M., Martino, R. and Vacca, A. *A remark of finite volume methods for 2D shallow water equations over irregular bottom topography*, J. Hydraul. Res. 59 (2021), 337–344.

[10] Kurganov, A. and Petrova, G. *Central-upwind schemes for two-layer shallow water equations*, SIAM J. Sci. Comput. 31(3) (2009), 1742–1773.

[11] Leveque, R.A. *Finite-volume methods for hyperbolic problems*. Cambridge University Press, 2002.

[12] Lina, I.R., Habibah, U., and Kusumawinahyu, W.M. *Mathematical modelling of two layer shallow water flow with incline and uneven bottom*, J. Phys. Conf. Ser. 1563 (2020), 0122019.

[13] Muhammad, N. *Finite volume method for simulation of flowing fluid via OpenFOAM*, Eur. Phys. J. Plus, 136 (10) (2021) 1–22.

[14] Mungkasi, S. *Finite volume methods for one-dimensional shallow water equations*. Ph.D. Thesis, Australian National University, 2008.

[15] Nikan, O. and Avazzadeh, Z.*Coupling of the Crank–Nicolson scheme and localized meshless technique for viscoelastic wave model in fluid flow*, J. Comput. Appl. Math. 398 (2021), 113695.

[16] 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 Dyn. 106 (2021), 783–813.

[17] Pascal, J.P. *A model for two-layer moderate Reynolds number flow down an incline*, Int. J. Non-Linear Mech. 36(6) (2001), 977–985.

[18] Rasoulizadeh, M.N., Ebadi, M.J., Avazzadeh, Z., and Nikan, O. *A high-order symmetrical weighted hybrid ENO-ux limiter scheme for hyperbolic conservation laws*, Comput. Phys. Commun. 185 (2014), 106–127.

[19] Rasoulizadeh, M.N., Ebadi, M.J., Avazzadeh, Z. and Nikan, O. *An efficient local meshless method for the equal width equation in fluid mechanics*, Eng. Anal. Bound Elem. 131 (2021), 258–268.

[20] Rasoulizadeh, M.N., Nikan, O. and Avazzadeh, Z. *The impact of LRBF-FD on the solutions of the nonlinear regularized long wave equation*, Math. Sci. 15 (2021), 365–376.

[21] Remi, A. and Smadar, K. *Two-layer shallow water system: A relaxation approach*, SIAM J. Sci. Comput. 31(3) (2009), 1603–1627.

[22] Swartenbroekx, C., Zech, Y.V. and Soares-Frazao, S. *Two-dimensional two-layer shallow water model for dam break flows with significant bed load transport*, Int. J. Numeric. Met, Fluids, 73 (2013), 477–508.

[23] Vallis, G.K. *Atmospheric and oceanic fluid dynamics*. Cambridge University Press, USA, 2006.

[24] Versteeg, H. K. and Malalasekera, W. *An introduction to computational fluid dynamics: The finite volume method*, Pearson Education Limited, 2007.

[25] Yadav, A., Chakraborty, S. and Usha, R. *Steady solution of an inverse problem in gravity-driven shear-thinning film flow: Reconstruction of an uneven bottom substrate*, J. Non-Newton. Fluid Mech. 219 (2015), 65–77.

How to cite this article
Habibah, U., Lina, I.R. and Kusumawinahyu, W.M., Two-layer shallow water formula with slope and uneven bottom solved by finite volume method. *Iran. j. numer. anal. optim.*, 2023; 13(2): 317-335. https://doi.org/10.22067/ijnao.2022.79205.1190

Iranian Journal of Numerical Analysis and Optimization
Vol. 13, No. 2, 2023, pp 336–353
https://doi.org/10.22067/ijnao.2022.73126.1069
https://ijnao.um.ac.ir/

CC BY NC

How to cite this article
Research Article

An operational matrix method for solving a class of nonlinear Volterra integral equations arising in steady activation of a skeletal muscle

S. Torkaman*, M. Heydari and G.B. Loghmani

Abstract

The problem of the steady activation of a skeletal muscle is one of the applicable phenomena in real life that can be modeled by a Volterra integral equation. The current research aims to investigate this problem by using an effective operational matrix-based method. For this purpose, the operational matrix of integration is derived for the barycentric rational cardinal basis functions. Then, by utilizing the obtained operational matrix and without using any collocation points, the governing integral equation is reduced to a system of nonlinear algebraic equations. Convergence analysis of the proposed numerical method is studied thoroughly. Moreover, the obtained numerical results based on the proposed method with acceptable computational times support the theoretical results and reveal the accuracy and efficiency of the method.

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

Keywords: Volterra integral equations; Linear barycentric rational interpolants; Operational matrix.

*Corresponding author

Received 17 October 2021; revised 17 August 2022; accepted 15 December 2022

Sorayya Torkaman
Department of Mathematical Sciences, Yazd University, Yazd, Iran.
e-mail: s.torkaman2009@gmail.com

Mohammad Heydari
Department of Mathematical Sciences, Yazd University, Yazd, Iran.
e-mail: m.heydari@yazd.ac.ir

Ghasem Barid Loghmani
Department of Mathematical Sciences, Yazd University, Yazd, Iran.
e-mail: loghmani@yazd.ac.ir

A Introduction

Mathematical modeling of most problems in mechanics, mathematical biology, and economics are expressed as integral equations. Skeletal or striated muscles are one of the three types of muscles in the body. They connect to bones and control body movements, maintain body posture, protect internal organs, support the entry and exit points of the body, and regulate body temperature. The steady activation of a skeletal muscle is a phenomenon with complex and subtle physiological processes that can be described by Volterra integral equations (VIEs) [5, 8, 10]. The motor units (MUs) develop a skeletal muscle; each MUs contains a motoneuron (MN) and the muscle fibers it innervates. Electrical signals, coming from higher motor centers and propagating to the motoneurons along with a network of nerve fibers, compose the input of MUs. In this problem, MUs are ordered according to their maximal (tetanic) contraction forces t. Consider the muscle force $f(t)$ as a function of the last recruited MU for an arbitrary but fixed network. Clearly, $f(t)$ is an increasing function of t. As mentioned in [5, 8], the muscle force $f(t)$ can be stated with the following VIE:

$$f(t) = \int_a^t k(s, f(s), f(t))ds. \tag{1}$$

Equation (1) is a subclass of VIEs and can be called non-standard VIEs because, in addition to $f(s)$, its kernel also includes $f(t)$. Let ρ be a density function that represents the MU population. Suppose that $t_{\min}$ and $t_{\max}$ be the tetanic forces of the weakest and strongest MUs and let $\int_{t_{\min}}^{t_{\max}} \rho(s)ds$ be the number of MUs in the pool. The VIE for (1) takes the form [5]

$$f(v) = \int_{t_{\min}}^{v} K(f(v), f(u))h(u)du, \quad v \in [t_{\min}, t_{\max}], \tag{2}$$

where

$$K(f(v), f(u)) = 1 - c\exp\Big(-\alpha\frac{f(v) - f(u)}{f(u) + \Delta}\Big),$$

and $h(v) = v\rho(v)$, $\alpha > 0$, $\Delta > 0$, and $0 < c < 1$. The integral equation (2) governs the activation of a muscle, and it is not clear whether this equation has a solution or not. Because ρ is a strictly positive function almost everywhere, the function $H(t) = \int_{t_{\min}}^{t} h(s)ds$ is strictly increasing on the interval $[t_{\min}, t_{\max}]$ and hence invertible. $H(t)$ represents the force of the muscle when all MUs up to level t produce their tetanic force and $H(t_{\max}) = F_{\max}$. Now, by considering the change of variable $u = H^{-1}(x)$, the VIE (2) can be reduced to

$$f(v) = \int_{H(t_{\min})}^{H(v)} K(f(v), f(H^{-1}(x)))dx, \quad v \in [t_{\min}, t_{\max}]. \tag{3}$$

Then by assuming $H(v) = t$ and equivalently $v = H^{-1}(t)$, the integral equation (3) can be written as

$$Y(t) = \int_0^t \Big(1 - c\exp\Big(-\alpha\frac{Y(t) - Y(x)}{Y(x) + \Delta}\Big)\Big)dx, \quad t \in [0, T], \tag{4}$$

where $Y(t) = f(H^{-1}(t))$ and $T = F_{\max}$. Finally, the muscle force function can be obtained as $f(t) = Y(H(t))$. The existence and uniqueness of a non-negative solution of the VIE (4) are studied in [5].

The main purpose of this paper is to develop a computational method based on the linear barycentric rational interpolants (LBRIs) and associated integral operational matrix to approximate the problem of steady activation of a skeletal muscle (1). The LBRIs were introduced by Berrut [2] and then generalized by Floater and Hormann [4]. The family of Floater–Hormann (FH) interpolants has no real poles and has a high order of convergence for enough smooth functions. Furthermore, the FH interpolants offer a better choice than the polynomial interpolants, which are ill-conditioned and yield Runge's phenomenon in the case of equispaced interpolation nodes. Also, the LBRIs can be applied to approximate derivatives and integrals of a given function [1, 6]. Due to the attractive features of the LBRIs, many researchers are interested in developing numerical methods based on such interpolants to solve functional equations such as differential equations [13, 11], Volterra and Fredholm integral equations [3, 7], and integro-differential integral equations [1].

The paper is organized as follows: In Sections B and C, the LBRIs and the corresponding integral operational matrix are described, respectively. In Section D, the numerical technique for solving the VIE (1) based on the integral operational matrix is introduced. In Section E, the convergence analysis of the proposed method is provided. In Section F, the obtained numerical results are reported. Finally, a brief conclusion is given in Section G.

B The LBRIs

Let $n \in \mathbb{N}$, let $f : [a, b] \to \mathbb{R}$, let $\Omega = \{a \le t_0 < t_1 < \cdots < t_n \le b\}$ be a set of nodes on the interval $[a, b]$, and let $f(t_j),\ j = 0, 1, \ldots, n$ be the corresponding given data. The LBRIs can be defined as follows:

$$I_n(t) = \sum_{j=0}^{n} f(t_j)\phi_j(t), \tag{5}$$

where

$$\phi_j(t) = \frac{\frac{w_j}{t - t_j}}{\sum_{i=0}^{n} \frac{w_i}{t - t_i}}, \quad j = 0, 1, \ldots, n, \tag{6}$$

are barycentric cardinal basis functions with the following features:

- $\sum_{j=0}^{n} \phi_j(t) = 1,$
- $\phi_j(t_i) = \delta_{ij}, \quad i, j = 0, 1, \ldots, n,$

where δ_{ij} is the Kronecker delta. Also, the nonzero real numbers $\{w_j\}_{j=0}^{n}$ are barycentric weights independent of $f(t_j)$. Floater and Hormann [4] proposed a family of LBRIs by introducing w_j as

$$w_j = (-1)^j \sum_{i=\max(j-d,0)}^{\min(j,n-d)} \Big(\prod_{k=i,k\neq j}^{i+d} \frac{1}{\mid t_j - t_k \mid} \Big), \quad j = 0, 1, \ldots, n, \tag{7}$$

where $0 \leq d \leq n$ is an integer parameter. Using the barycentric weights (7), the LBRI (5) can be rewritten as follows [4]:

$$I_n(t) = \frac{\sum_{j=0}^{n-d} \lambda_j(t) p_j(t)}{\sum_{j=0}^{n-d} \lambda_j(t)}, \quad \lambda_j(t) = \frac{(-1)^j}{(t-t_j)\ldots(t-t_{j+d})}, \quad j = 0, 1, \ldots, n-d, \tag{8}$$

where $p_j(t)$ is the polynomial of degree at most d interpolates $f(t)$ at local nodes $\{x_{j+k}\}_{k=0}^{d}$. In [4], it is proved that the interpolant (8) has no real poles.

Theorem 1. [4] *Let $f \in C^{d+2}[a, b]$. Then*

$$\|f(t) - I_n(t)\|_\infty \leq C h^{d+1},$$

where

$$C = (1 + \gamma\mu) \begin{cases} (b-a)\frac{\|f^{(d+2)}\|_\infty}{d+2}, & (n-d) \text{ odd}, \\ \left((b-a)\frac{\|f^{(d+2)}\|_\infty}{d+2} + \frac{\|f^{(d+1)}\|_\infty}{d+1} \right), & (n-d) \text{ even}, \end{cases}$$

$$\mu = \max_{1\leq i\leq n-2} \min \left(\frac{t_{i+1} - t_i}{t_i - t_{i-1}}, \frac{t_{i+1} - t_i}{t_{i+2} - t_{i+1}} \right),$$

$$\gamma = \begin{cases} 1, & d = 0, \\ 0, & d > 1, \end{cases}, \qquad h = \max_{0\leq i\leq n-1} (t_{i+1} - t_i).$$

Similar to the one-dimensional LBRIs (6), one can define two-dimensional LBRIs. For this purpose, two-dimensional barycentric rational basis func-

tions can be defined as

$$\psi_{ij}(t,s) = \phi_i(t)\phi_j(s), \quad i,j = 0,1,\ldots,n,$$

where $\{\phi_i(t)\}_{i=0}^n$ are the basis functions (6). The basis functions $\{\psi_{ij}(t,s)\}$ have the following properties:

- $\sum_{i=0}^{n}\sum_{j=0}^{n}\psi_{ij}(t,s) = 1,$
- $\psi_{ij}(t_k,s_l) = \delta_{ik}\delta_{jl}.$

According to the Lagrange property, any two-dimensional function $f(t,s)$ can be approximated as

$$f(t,s) \simeq \sum_{i=0}^{n}\sum_{j=0}^{n} f(t_i,s_j)\psi_{ij}(t,s). \tag{9}$$

Also, by using the barycentric weights defined in (7), the two-dimensional LBRI (9) can be rewritten in the form of the two-dimensional FH interpolants as

$$I_n(t,s) = \frac{\sum_{i=0}^{n-d}\sum_{j=0}^{n-d}\lambda_i(t)\lambda_j(s)p_{i,j}(t,s)}{\sum_{i=0}^{n-d}\sum_{j=0}^{n-d}\lambda_i(t)\lambda_j(s)}, \tag{10}$$

where

$$\lambda_i(t) = \frac{(-1)^i}{(t-t_i)\ldots(t-t_{i+d})},$$
$$\lambda_j(s) = \frac{(-1)^j}{(s-s_j)\ldots(s-s_{j+d})}, \quad i,j = 0,1,\ldots,n,$$

in which $p_{i,j}(t,s)$ is the polynomial that interpolates f at local nodes $\{(t_{i+k},s_{j+l})\}_{k,l=0}^{d}$. In the following, by applying Theorem 1, the error bound of the two-dimensional Floater–Hormann interpolant (10) is estimated.

Theorem 2. *Let $f(t,s) \in C^{d+2}([a,b]\times[a,b])$. Then*

$$\|f(t,s) - I_n(t,s)\|_\infty = O(h^{d+1}).$$

Proof. Suppose that

$$\Omega_t = \{a \le t_0 < t_1 < \cdots < t_n \le b\},$$
$$\Omega_s = \{a \le s_0 < s_1 < \cdots < s_n \le b\},$$

and let $\{(t_i, s_j)\}_{i,j=0}^{n}$ be the interpolation nodes on $[a,b]\times[a,b]$. Then

$$f(t,s) = I_n(t,s), \qquad \text{for all } (t,s)\in \Omega_t \times \Omega_s.$$

Now, the error interpolation can be written as

$$f(t,s) - I_n(t,s) = \frac{\sum_{i=0}^{n-d}\sum_{j=0}^{n-d} \lambda_i(t)\lambda_j(s)(f(t,s) - p_{ij}(t,s))}{\sum_{i=0}^{n-d}\sum_{j=0}^{n-d} \lambda_i(t)\lambda_j(s)}, \tag{11}$$

for any $(t,s)\in[a,b]\times[a,b]-\Omega_t\times\Omega_s$. By extending the Newton form of polynomial interpolant error and using the commuting property of the divided differences described in [9], the Newton error formula for a two-dimensional polynomial interpolation can be represented as

$$\begin{aligned} f(t,s) - p_{ij}(t,s) =& \pi_{d+1}(t)[t_i, t_{i+1}, \ldots, t_{i+d}, t]f \\ &+ \pi_{d+1}(s)[s_j, s_{j+1}, \ldots, s_{j+d}, s]f \\ &- \pi_{d+1}(t)\pi_{d+1}(s)[s_j, s_{j+1}, \ldots, s_{j+d}, s][t_i, t_{i+1}, \ldots, t_{i+d}, t]f, \end{aligned} \tag{12}$$

where $\pi_{d+1}(t) = \prod_{r=0}^{d}(t - t_{i+r})$ and $\pi_{d+1}(s) = \prod_{r=0}^{d}(s - s_{j+r})$. Replacing (12) in (11) yields

$$\begin{aligned} f(t,s) - I_n(t,s) =& \frac{\sum_{i=0}^{n-d}(-1)^i[t_i, t_{i+1}, \ldots, t_{i+d}, t]f}{\sum_{i=0}^{n-d}\lambda_i(t)} \\ &+ \frac{\sum_{j=0}^{n-d}(-1)^j[s_j, s_{j+1}, \ldots, s_{j+d}, s]f}{\sum_{j=0}^{n-d}\lambda_j(s)} \\ &- \frac{\sum_{i=0}^{n-d}\sum_{j=0}^{n-d}(-1)^{i+j}[t_i, t_{i+1}, \ldots, t_{i+d}, t][s_j, s_{j+1}, \ldots, s_{j+d}, s]f}{\sum_{i=0}^{n-d}\sum_{j=0}^{n-d}\lambda_i(t)\lambda_j(s)}. \end{aligned} \tag{13}$$

To complete the proof, it is enough to compute an upper bound for the numerators and a lower bound for the denominators of the used quotient in (13). In [4], the upper bounds for the numerator of the first and second quotients applied in (13) are derived. Moreover, the following inequalities hold [4]:

$$\Big|\sum_{i=0}^{n-d}\lambda_i(t)\Big| \geq \frac{1}{(1+\gamma\mu)d!h^{d+1}}, \quad \Big|\sum_{j=0}^{n-d}\lambda_j(s)\Big| \geq \frac{1}{(1+\gamma\mu)d!h^{d+1}}, \tag{14}$$

where γ and μ are defined in Theorem 1. Now, it is enough to get an upper bound on the numerator of the third quotient in (13). Let $n-d$ be odd. Following the proof of Theorem 2 in [4], one can get

$$\begin{aligned}
&\sum_{i=0}^{n-d}\sum_{j=0}^{n-d}(-1)^{i+j}[t_i,t_{i+1},\ldots,t_{i+d},t][s_j,s_{j+1},\ldots,s_{j+d},s]f\\
&=-\sum_{i=0}^{n-d}\sum_{\substack{j=0\\ j\ \text{even}}}^{n-d-1}(-1)^i(s_{j+d+1}-s_j)[t_i,t_{i+1},\ldots,t_{i+d},t][s_j,s_{j+1},\ldots,s_{j+d+1},s]f\\
&=\sum_{\substack{i=0\\ i\ \text{even}}}^{n-d-1}\sum_{\substack{j=0\\ j\ \text{even}}}^{n-d-1}(t_{i+d+1}-t_i)(s_{j+d+1}-s_j)\\
&\quad\times[t_i,t_{i+1},\ldots,t_{i+d+1},t][s_j,s_{j+1},\ldots,s_{j+d+1},s]f,
\end{aligned}$$

and therefore

$$\begin{aligned}
&\Big|\sum_{i=0}^{n-d}\sum_{j=0}^{n-d}(-1)^{i+j}[t_i,t_{i+1},\ldots,t_{i+d},t][s_j,s_{j+1},\ldots,s_{j+d},s]f\Big|\\
&\quad\leq\sum_{\substack{i=0\\ i\ \text{even}}}^{n-d-1}\sum_{\substack{j=0\\ j\ \text{even}}}^{n-d-1}\frac{|t_{i+d+1}-t_i||s_{j+d+1}-s_j|}{((d+2)!)^2}\Big|\frac{\partial^{2d+4}}{\partial t^{d+2}\partial s^{d+2}}f\Big|\\
&\quad\leq\frac{(d+1)^2(b-a)^2}{((d+2)!)^2}\Big\|\frac{\partial^{2d+4}}{\partial t^{d+2}\partial s^{d+2}}f\Big\|_\infty.
\end{aligned}\tag{15}$$

Let $n-d$ be even. In a similar manner, by following the proof of Theorem 2 in [4], the following result is obtained:

$$\begin{aligned}
&\sum_{i=0}^{n-d}\sum_{j=0}^{n-d}(-1)^{i+j}[t_i,t_{i+1},\ldots,t_{i+d},t][s_j,s_{j+1},\ldots,s_{j+d},s]f\\
&=\sum_{\substack{i=0\\ i\ \text{even}}}^{n-d-2}\sum_{\substack{j=0\\ j\ \text{even}}}^{n-d-2}(t_{i+d+1}-t_i)(s_{j+d+1}-s_j)\\
&\quad\times[t_i,t_{i+1},\ldots,t_{i+d+1},t][s_j,s_{j+1},\ldots,s_{j+d+1},s]f\\
&\quad-\sum_{\substack{i=0\\ i\ \text{even}}}^{n-d-2}(t_{i+d+1}-t_i)[t_i,t_{i+1},\ldots,t_{i+d+1},t][s_{n-d},s_{n-d+1},\ldots,s_n,s]f
\end{aligned}$$

$$- \sum_{\substack{j=0 \\ j \text{ even}}}^{n-d-2} (s_{j+d+1} - s_j)[t_{n-d}, t_{n-d+1}, \ldots, t_n, t][s_j, s_{j+1}, \ldots, s_{j+d+1}, s]f$$
$$+ [t_{n-d}, t_{n-d+1}, \ldots, t_n, t][s_{n-d}, s_{n-d+1}, \ldots, s_n, s]f,$$

and therefore

$$\Big| \sum_{i=0}^{n-d} \sum_{j=0}^{n-d} (-1)^{i+j} [t_i, t_{i+1}, \ldots, t_{i+d}, t][s_j, s_{j+1}, \ldots, s_{j+d}, s]f \Big|$$
$$\leq \frac{(d+1)^2(b-a)^2}{((d+2)!)^2} \Big\| \frac{\partial^{2d+4}}{\partial t^{d+2} \partial s^{d+2}} f \Big\|_\infty + \frac{b-a}{d!(d+2)!} \Big\| \frac{\partial^{2d+3}}{\partial t^{d+2} \partial s^{d+1}} f \Big\|_\infty$$
$$+ \frac{b-a}{d!(d+2)!} \Big\| \frac{\partial^{2d+3}}{\partial t^{d+1} \partial s^{d+2}} f \Big\|_\infty + \frac{1}{((d+1)!)^2} \Big\| \frac{\partial^{2d+2}}{\partial t^{d+1} \partial s^{d+1}} f \Big\|_\infty. \qquad (16)$$

Finally, by using the relations (14), (15), and (16), the desired result can be obtained as follows:

• If $n - d$ is odd, then

$$\Big\| f(s,t) - I_n(s,t) \Big\|_\infty \leq \frac{(1+\gamma\mu)(b-a)}{d+2} \left(\Big\| \frac{\partial^{d+2}}{\partial t^{d+2}} f \Big\|_\infty + \Big\| \frac{\partial^{d+2}}{\partial s^{d+2}} f \Big\|_\infty \right) h^{d+1}$$
$$+ \frac{(1+\gamma\mu)^2(b-a)^2}{(d+2)^2} \Big\| \frac{\partial^{2d+4}}{\partial t^{d+2} \partial s^{d+2}} f \Big\|_\infty h^{2d+2}.$$

• If $n - d$ is even, then

$$\Big\| f(t,s) - I_n(t,s) \Big\|_\infty \leq \frac{(1+\gamma\mu)(b-a)}{d+2} \left(\Big\| \frac{\partial^{d+2}}{\partial t^{d+2}} f \Big\|_\infty + \Big\| \frac{\partial^{d+2}}{\partial s^{d+2}} f \Big\|_\infty \right) h^{d+1}$$
$$+ \frac{1+\gamma\mu}{d+1} \left(\Big\| \frac{\partial^{d+1}}{\partial t^{d+1}} f \Big\|_\infty + \Big\| \frac{\partial^{d+1}}{\partial s^{d+1}} f \Big\|_\infty \right) h^{d+1}$$
$$+ \frac{(1+\gamma\mu)^2(b-a)^2}{(d+2)^2} \Big\| \frac{\partial^{2d+4}}{\partial t^{d+2} \partial s^{d+2}} f \Big\|_\infty h^{2d+2}$$
$$+ \frac{(1+\gamma\mu)^2}{(d+1)^2} \Big\| \frac{\partial^{2d+2}}{\partial t^{d+1} \partial s^{d+1}} f \Big\|_\infty h^{2d+2}$$
$$+ \frac{(1+\gamma\mu)^2(b-a)}{(d+1)(d+2)} h^{2d+2}$$
$$\times \left(\Big\| \frac{\partial^{2d+3}}{\partial t^{d+2} \partial s^{d+1}} f \Big\|_\infty + \Big\| \frac{\partial^{2d+3}}{\partial t^{d+1} \partial s^{d+2}} f \Big\|_\infty \right).$$

□

C Integral operational matrix

In this section, a brief review of the integral operational matrix based on the barycentric rational basis functions described in [14, 12] is provided. Suppose that $\Phi(t)$ is an $(n+1)\times 1$ vector as follows:

$$\Phi(t) = [\phi_0(t), \phi_1(t), \ldots, \phi_n(t)]^T, \tag{17}$$

where $\phi_i(t),\ i = 0, 1, \ldots, n$, are defined in (6). The integration of vector $\Phi(t)$ in (17) can be expressed as

$$\int_a^t \Phi(s)ds = \Big[\int_a^t \phi_0(s)ds, \int_a^t \phi_1(s)ds, \ldots, \int_a^t \phi_n(s)ds\Big]^T. \tag{18}$$

Using (5), any elements of the vector (18) can be estimated as

$$\int_a^t \phi_i(s)ds \simeq \sum_{j=0}^{n} p_{ij}\phi_j(t) = \mathbf{P_i}\Phi(t), \quad i = 0, 1, \ldots, n. \tag{19}$$

where

$$\mathbf{P_i} = [p_{i0}, p_{i1}, \ldots, p_{in}], \quad i = 0, 1, \ldots, n.$$

Replacing (19) in (1) yields

$$\int_a^t \Phi(s)ds \simeq \Big[\mathbf{P_0}\Phi(t), \mathbf{P_1}\Phi(t), \ldots, \mathbf{P_n}\Phi(t)\Big]^T = \mathbf{P}\Phi(t),$$

in which

$$\mathbf{P} = \Big[\mathbf{P_0}, \mathbf{P_1}, \ldots, \mathbf{P_n}\Big]^T.$$

Hence, the operational matrix of integration can be defined as follows.

Definition 1. Let $\Phi(t)$ be the vector stated in (17). The operational matrix of integration $\mathbf{P}$, based on the barycentric basis functions, can be represented as

$$\int_a^t \Phi(s)ds \simeq \mathbf{P}\Phi(t),$$

where $\mathbf{P} = (p_{ij})$ and p_{ij} can be computed as

$$p_{ij} = \int_a^{t_j} \phi_i(s)ds, \quad i, j = 0, 1, \ldots, n.$$

Remark 1. Let A be an $(n+1)\times(n+1)$ matrix. Using (10), the function $\Phi^T(t)A\Phi(t)$ can be approximated as follows:

$$\Phi^T(t)A\Phi(t) \simeq \sum_{j=0}^{n} \Phi^T(t_j)A\Phi(t_j)\phi(t).$$

Due to the Lagrange property of the basis functions (6), one can get

$$\Phi^T(t)A\Phi(t) \simeq \hat{\mathbf{A}}\Phi(t), \quad \hat{\mathbf{A}} = [A_{00}, A_{11}, \ldots, A_{nn}],$$

where $\hat{\mathbf{A}}$ is a row vector composed of the diagonal elements of matrix A.

D Description of the numerical method

In this section, an operational matrix-based method will be discussed to solve the nonlinear VIE (1). For this purpose, by using (5), the function $f(t)$ is approximated as

$$f(t) \simeq f_n(t) = \sum_{i=0}^{n} f(t_i)\phi_i(t) = F^T\Phi(t), \tag{20}$$

where

$$F = [f(t_0), f(t_1), \ldots, f(t_n)]^T,$$

is the unknown vector and should be determined to compute the approximate solution of the problem. At first, using (20), the function $k(s, f_n(s), f_n(t))$ is estimated respect to the variable s as

$$k\Big(s, f_n(s), f_n(t)\Big) \simeq K(t)^T\Phi(s), \tag{21}$$

where

$$K(t) = \Big[k\Big(s_0, f_n(s_0), f_n(t)\Big), k\Big(s_1, f_n(s_1), f_n(t)\Big), \ldots, k\Big(s_n, f_n(s_n), f_n(t)\Big)\Big]^T,$$

and $s_i = t_i,\ i = 0, 1, \ldots, n$. From (20) and according to the Lagrange property of the basis functions $\phi_i(t),\ i = 0, 1, \ldots, n$, the vector $K(t)$ can be rewritten as

$$K(t) = \Big[k\Big(s_0, f(s_0), f_n(t)\Big), k\Big(s_1, f(s_1), f_n(t)\Big), \ldots, k\Big(s_n, f(s_n), f_n(t)\Big)\Big]^T.$$

Again using (5), each component of the vector $K(t)$ is approximated as

$$k\Big(s_i, f(s_i), f_n(t)\Big) \simeq K_i^T \Phi(t), \quad i = 0, 1, \ldots, n, \tag{22}$$

where

$$K_i = \Big[k\Big(s_i, f(s_i), f(t_0)\Big), k\Big(s_i, f(s_i), f(t_1)\Big), \ldots, k\Big(s_i, f(s_i), f(t_n)\Big)\Big]^T.$$

Substituting (22) into (21) yields

$$k\Big(s, f_n(s), f_n(t)\Big) \simeq \Phi^T(t) K^T \Phi(s), \tag{23}$$

where K is a $(n+1) \times (n+1)$ matrix and $K_{ij} = k(s_i, f(s_i), f(t_j))$. Now, by substituting (20) and (23) into (1), one can obtain

$$F^T \Phi(t) - \Phi(t)^T K^T \int_a^t \Phi(s) ds = 0.$$

Using the integral operational matrix defined in Definition 1, we arrive at

$$F^T \Phi(t) - \Phi(t)^T K^T \mathbf{P} \Phi(t) = 0. \tag{24}$$

Finally, applying Remark 1 leads to the following nonlinear algebraic system:

$$F^T - \hat{\mathbf{A}} = 0, \tag{25}$$

where $\hat{\mathbf{A}}$ is a row vector constructed of the diagonal elements of the matrix $K^T \mathbf{P}$. Thus, the unknown vector F can be computed by solving the nonlinear system of algebraic equations (25).

E Convergence analysis

In this section, the convergence analysis of the proposed numerical scheme is presented. To this end, we utilize the following norms to measure the approximation error:

$\|v(t)\|_\infty = \max\limits_{t \in [a,b]} v(t)$, for real function $v : [a, b] \to \mathbb{R}$,

$\|V(t)\|_\infty = \max\limits_{1 \le i \le k} \|v_i(t)\|_\infty$, for vector function $V(t) = [v_1(t), v_2(t), \ldots, v_k(t)]^T$, $v_i : [a, b] \to \mathbb{R}$.

Lemma 1. *Let $\Phi(t)$ be the vector defined in (17) with the weights (7). Then,*

$$\Big\| \int_a^t \Phi(s) ds - \mathbf{P} \Phi(t) \Big\|_\infty = O(h^{d+1}).$$

Proof. By integrating from the vector $\Phi(t)$, one obtains

$$\Big\| \int_a^t \Phi(s)ds - \mathbf{P}\Phi(t) \Big\|_\infty = \max_{0\le i\le n} \Big| \int_a^t \phi_i(s)ds - \mathbf{P_i}\Phi(t) \Big|, \tag{26}$$

where $\mathbf{P_i}$ is the ith row of the integral operational matrix $\mathbf{P}$. According to (19) and Theorem 1, one can get

$$\Big\| \int_a^t \phi_i(s)ds - \mathbf{P_i}\Phi(t) \Big\|_\infty \le h^{d+1}(1+\gamma\mu) \times \begin{cases} (b-a)\frac{\|\phi_i^{(d+2)}\|_\infty}{d+2}, & (n-d) \text{ odd} \\ \left((b-a)\frac{\|\phi_i^{(d+2)}\|_\infty}{d+2} + \frac{\|\phi_i^{(d+1)}\|_\infty}{d+1}\right), & (n-d) \text{ even} \end{cases} \tag{27}$$

Substituting (27) into (26), the desired result is obtained as

$$\Big\| \int_a^t \Phi(s)ds - \mathbf{P}\Phi(t) \Big\|_\infty \le Ch^{d+1},$$

where

$$C = (1+\gamma\mu) \begin{cases} (b-a)\frac{\|\Phi^{(d+2)}\|_\infty}{d+2}, & (n-d) \text{ odd} \\ \left((b-a)\frac{\|\Phi^{(d+2)}\|_\infty}{d+2} + \frac{\|\Phi^{(d+1)}\|_\infty}{d+1}\right), & (n-d) \text{ even} \end{cases}$$

and

$$\Phi^{(d)}(t) = [\phi_0^{(d)}(t), \phi_1^{(d)}(t), \dots, \phi_n^{(d)}(t)]^T, \quad \| \Phi^{(d)}(t) \|_\infty = \max_{0\le i\le n} |\phi_i^{(d)}(t)|.$$

□

Lemma 2. *Consider A an $(n+1)\times(n+1)$ matrix, and let $\Phi(t)$ be the vector defined in (17) with the weights (7). Then,*

$$\|\Phi^T(t)A\Phi(t) - \hat{\mathbf{A}}\Phi(t)\|_\infty = O(h^{d+1}),$$

where $\hat{\mathbf{A}}$ is a row vector composed of the diagonal elements of the matrix A.

Proof. As mentioned in Remark 1, we have

$$\Phi^T(t)A\Phi(t) \simeq \hat{\mathbf{A}}\Phi(t).$$

Finally, by applying the weights (7) and using Theorem 1, the desired result is acquired. □

Theorem 3. *Let $f \in C^{d+2}[a,b]$, let $k \in C^{d+2}([a,b],\mathbb{R},\mathbb{R})$, and let the kernel k satisfy the following Lipschitz condition:*

$$|k(s,y_1,z_1)-k(s,y_2,z_2)| \le L(|y_1-y_2|+|z_1-z_2|),$$
$$\text{for all } (s,y_1,z_1),(s,y_2,z_2) \in [a,b]\times\mathbb{R}^2.$$

Also, consider $f_n(t)=F^T\Phi(t)$ the approximate solution of (1) such that F is obtained from solving the nonlinear system of algebraic equations (25) and $2L(b-a)<1$. Then

$$\|f(t)-f_n(t)\|_\infty \to 0, \quad \text{as} \quad n\to+\infty.$$

Proof. The residual function of the VIE (1) can be computed as follows:

$$\mathbf{RES}_n(t) = f_n(t) - \int_a^t k\Big(s, f_n(s), f_n(t)\Big)ds. \tag{28}$$

Now, by subtracting relation (24) from (28) and utilizing equation (25), the residual function $\mathbf{RES}_n(t)$ can be rewritten as

$$\mathbf{RES}_n(t) = \underbrace{(F^T-\hat{\mathbf{A}})}_{0}\Phi(t) - \int_a^t \Big(k\Big(s,f_n(s),f_n(t)\Big) - \Phi^T(t)K^T\Phi(s)\Big)ds$$
$$-\Phi^T(t)K^T\Big(\int_a^t \Phi(s)ds - \mathbf{P}\Phi(t)\Big) - (\Phi^T(t)K^T\mathbf{P}\Phi(t) - \hat{\mathbf{A}}\Phi(t)).$$

Therefore

$$\|\mathbf{RES}_n(t)\|_\infty \le (b-a)\Big\|k\Big(s,f_n(s),f_n(t)\Big) - \Phi^T(t)K^T\Phi(s)\Big\|_\infty$$
$$+\|\Phi^T(t)K^T\|_\infty\Big\|\int_a^t \Phi(s)ds - \mathbf{P}\Phi(t)\Big\|_\infty$$
$$+\|\Phi^T(t)K^T\mathbf{P}\Phi(t) - \hat{\mathbf{A}}\Phi(t)\|_\infty,$$

Employing Theorem 2, and Lemmas 1 and 2 leads to the following relation:

$$\|\mathbf{RES}_n(t)\|_\infty = O(h^{d+1}). \tag{29}$$

Let $E_n(t)=f(t)-f_n(t)$. Substituting $f(t)=f_n(t)+E_n(t)$ into the VIE (1), one can obtain

$$f_n(t)+E_n(t) = \int_a^t k\Big(s, f_n(s)+E_n(s), f_n(t)+E_n(t)\Big)ds. \tag{30}$$

Employing the residual function (28) and relation (30) yields the following equation:

$$E_n(t) = \int_a^t \Big(k(s, f_n(s) + E_n(s), f_n(t) + E_n(t)) - k(s, f_n(s), f_n(t))\Big)ds \\ -\mathbf{RES}_n(t).$$

Therefore

$$|E_n(t)| \leq \int_a^t \Big|k(s, f_n(s) + E_n(s), f_n(t) + E_n(t)) - k(s, f_n(s), f_n(t))\Big|ds \\ +|\mathbf{RES}_n(t)|.$$

Finally, applying the Lipschitz condition yields

$$|E_n(t)| \leq L \int_a^t (|E_n(s)| + |E_n(t)|)ds + |\mathbf{RES}_n(t)|,$$

and consequently

$$\|E_n(t)\|_\infty \leq \frac{\|\mathbf{RES}_n(t)\|_\infty}{1 - 2L(b-a)}.$$

From (29) and $2L(b-a) < 1$, the desired result is obtained as

$$\|E_n(t)\|_\infty \to 0 \quad \text{as} \quad h \to 0.$$

□

F Numerical results

In this section, the proposed method is applied based on the integral operational matrix to approximate the solution of the VIE (4). Consider $t_{\min} = 10$, $t_{\max} = 20$, $c = 0.9$, $\alpha = 1.14$, and $\rho(t) = \frac{1}{t}$, $t \in [t_{\min}, t_{\max}]$. To get the numerical results, the equidistant nodes are employed on the interval $[0, 10]$, and the algorithm associated with the proposed method is implemented using Maple 2015 software on a Core (TM) i7 PC with 2.70 GHz of CPU and 16 GB of RAM. Here, the fsolve command is used to solve the nonlinear system (25). Also, the rate of convergence of the numerical method is determined based on the following formula:

$$\text{Ratio} = \frac{\log(\|E_n\|_\infty) - \log(\|E_{n'}\|_\infty)}{\log(h) - \log(h')}, \tag{31}$$

where h and h' are defined in Theorem 1. Because the exact solution of (4) is unknown, we use the error of residual function to verify the efficiency and high accuracy of the proposed method. So, to commutating the rate of convergence of the method, the absolute error $E_n(t)$ in (31) is replaced by the error of residual function $\mathbf{RES}_n(t)$.

Figure 1 depicts the approximation solution $Y_n(t)$ and the residual function $|\mathbf{RES}_n(t)|$ with $n = 20$ and $d = 18$ for different values of Δ.

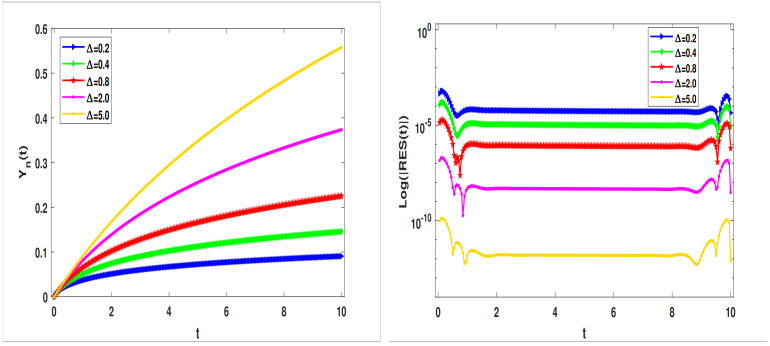


Figure 1: Graphs of $Y(t)$ and $|\mathbf{RES_n(t)}|$ with $n = 20$ and $d = 18$

Table 1 reports the numerical results for the VIE (4) for $\Delta = 2$, 5, and 10.

Table 1: Numerical results with $n = 20$ and $d = 19$

t	$\lvert\mathbf{RES}_n(t)\rvert$		
	$\Delta = 2$	$\Delta = 5$	$\Delta = 10$
0.0	0.000000	0.0000000	0.0000000
1.0	$3.98e-9$	$9.53e-13$	$1.67e-16$
2.0	$3.90e-9$	$9.46e-13$	$1.66e-16$
3.0	$3.82e-9$	$9.34e-13$	$1.65e-16$
4.0	$3.75e-9$	$9.23e-13$	$1.64e-16$
5.0	$3.69e-9$	$9.14e-13$	$1.63e-16$
6.0	$3.64e-9$	$9.05e-13$	$1.62e-16$
7.0	$3.60e-9$	$8.97e-13$	$1.63e-16$
8.0	$3.56e-9$	$8.80e-13$	$1.72e-16$
9.0	$3.51e-9$	$8.80e-13$	$2.21e-16$
10	$2.44e-9$	$4.74e-13$	$3.33e-16$
CPU time (s)	2.121	2.043	2.184

Table 2 represents the residual function error and the convergence rate for different values of d. According to this table, one can infer that the order of convergence is ≈ 4 for $d = 3$ and ≈ 3 for $d = 2$. So, the numerical results in Table 2 confirm the theoretical results stated in Theorem 3.

Table 2: Numerical results for $\|\mathbf{RES}_n(t)\|_\infty$ with $\Delta = 10$

n	h	$d = 3$	Ratio	CPU time(s)	$d = 2$	Ratio	CPU time(s)
4	2.50000	$1.91e-4$	–	0.343	$1.01e-3$	–	0.341
8	1.25000	$2.09e-5$	3.19	0.562	$1.38e-4$	2.87	0.515
16	0.62500	$1.67e-6$	3.65	1.373	$1.86e-5$	2.89	1.357
32	0.31250	$1.20e-7$	3.80	5.538	$2.46e-6$	2.92	5.491
64	0.15625	$7.80e-9$	3.94	27.690	$2.90e-7$	3.08	27.082

As it is clear from the graphical and tabulated results, the proposed method can be successfully employed to solve the VIEs (1).

G Conclusion

In this paper, the problem of the steady activation of a skeletal muscle has been investigated. Applying the numerical technique based on the integral operational matrix of the LBRIs, the approximate solution of the nonlinear VIE governing the problem was obtained. One of the advantages of the proposed method is that the FH interpolants are infinitely smooth rational interpolants and involve polynomial interpolants when $d = n, n-1$. Moreover, without using collocation points, the VIE (1) is reduced to a system of nonlinear algebraic equations. Numerical results confirmed the obtained theoretical results of Theorem 3. Thus, the method can be successfully employed to approximate the nonlinear VIEs with acceptable accuracy and computational times.

Acknowledgements

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

References

[1] Abdi, A. and Hosseini, A. *The barycentric rational difference-quadrature scheme for systems of Volterra integro-differential equations*, SIAM J. Sci. Comput. 40 (2018), A1936–A1960.

[2] Berrut, J.P. *Rational function for guaranteed and experimentally well-conditioned global interpolation*, Comput. Math. Appl. 15 (1988), 1–16.

[3] Berrut, J. P., Hosseini, S. A. and Klein, G. *The linear barycentric rational quadrature method for volterra integral equations*, SIAM J. Sci. Comput. 36 (2014), A105–A123.

[4] Floater, M.S. and Hormann, K. *Barycentric rational interpolation with no poles and high rates of approximation*, Numer. Math. 107(2) (2007), 315–331.

[5] Gabriel, J.P., Studer, L.M., Rüegg, D.G. and Schnetzer, M.A. *A mathematical model for the steady activation of a skeletal muscle*, SIAM J. Appl. Math. 68 (2008), 869–889.

[6] Klein, G. and Berrut, J.P. *Linear barycentric rational quadrature*, BIT, 52 (2012), 407–424.

[7] Liu, H., Huang, J., Pan, Y. and Zhang, J. *Barycentric interpolation collocation methods for solving linear and nonlinear high-dimensional Fredholm integral equations*, J. Comput. Appl. Math. 327 (2018), 141–154.

[8] Maleknejad, K., Torabi, P. and Mollapourasl, R. *Fixed point method for solving nonlinear quadratic Volterra integral equations*, Comput. Math. Appl. 62 (2011), 2555–2566.

[9] Phillips, G.M. *Interpolation and approximation by polynomials*, CMS Books in Mathematics, Springer, New York, NY, 2003.

[10] Studer, L.M., Ruegg, D.G. and Gabrie, J.P. *A model for steady isometric muscle activation*, Biol. Cybern. 80 (1999), 339–355.

[11] Torkaman, S., Heydari, M., Loghmani, G. B. and Ganji, D.D. *Barycentric rational interpolation method for numerical investigation of magnetohydrodynamics nanofluid flow and heat transfer in nonparallel plates with thermal radiation*, Heat Transfer-Asian Research, 49 (2020), 565–590.

[12] Torkaman, S., Heydari, M., Loghmani, G.B. and Wazwaz, A.M. *Numerical investigation of three-dimensional nanofluid flow with heat and mass transfer on a nonlinearly stretching sheet using the barycentric functions*, Int. J. Numer. Methods Heat Fluid Flow. 31 (2020), 783–808.

[13] Torkaman, S., Loghmani, G.B., Heydari, M. and Rashidi, M.M. *Novel numerical solutions of nonlinear heat transfer problems using the linear barycentric rational interpolation*, Heat Transfer-Asian Research, 48 (2019), 1318–1344.

[14] Torkaman, S., Loghmani, G.B., Heydari, M. and Rashidi, M.M. *An effective operational matrix method based on barycentric cardinal functions to study nonlinear MHD nanofluid flow and heat transfer*, Int. J. Mech. Eng. 5 (2020,) 51–63.

How to cite this article
Torkaman, S., Heydari, M. and Loghmani, G.B., An operational matrix method for solving a class of nonlinear Volterra integral equations arising in steady activation of a skeletal muscle. *Iran. j. numer. anal. optim.*, 2023; 13(2): 336-353. https://doi.org/10.22067/ijnao.2022.73126.1069

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 optimization. Articles must be written in English.
All submitted papers will be refereed and the authors may be asked to revise their manuscripts according to the referee's reports. The Editorial Board of the Journal keeps the right to accept or reject the papers for publication.
The papers with more than one authors, should determine the corresponding author. The e-mail address of the corresponding author must appear at the end of the manuscript or as a footnote of the first page.
It is strongly recommended to set up the manuscript by Latex or Tex, using the template provided in the web site of the Journal. Manuscripts should be typed double-spaced with wide margins to provide enough room for editorial remarks.

References should be arranged in alphabetical order by the surname of the first author as examples below:

[1] Brunner, H. *A survey of recent advances in the numerical treatment of Volterra integral and integro-differential equations*, J. Comput. Appl. Math. 8 (1982), 213-229.

[2] Stoer, J. and Bulirsch, R. *Introduction to Numerical Analysis*, Springer-verlag, New York, 2002.

Iranian Journal of Numerical Analysis and Optimization

CONTENTS Vol. 13, No. 2, pp 157-353

A three-free-parameter class of power series based iterative method for approximation of nonlinear equations solution . . 157
O. Ogbereyivwe and O. Izevbizua

Stability and Hopf Bifurcation in Leech Heart Interneuron Model 170
F. Parvizi, M. Razvan and Y. Alipour Fakhri

Approximate solutions to the Allen–Cahn equation using rational radial basis functions method 187
M. Shiralizadeh, A. Alipanah and M. Mohammadi

Regularization of the generalized auto-convolution Volterra integral equation of the first kind 205
S. Pishbin and A. Ebadi

The interpolated variational iteration method for solving a class of nonlinear optimal control problems 224
M. Shirazian

A robust method for optimal control problems governed by system of Fredholm integral equations in mechanics 243
A. Ebrahimzadeh

Fitted scheme for singularly perturbed time delay reaction-diffusion problems 262
M. Amsalu Ayele, A. Andargie Tiruneh and G. Adamu Derese

Modified Runge–Kutta method with convergence analysis for nonlinear stochastic differential equations with Hölder continuous diffusion coefficient 285
A. Haghighi

Two-layer shallow water formula with slope and uneven bottom solved by finite volume method 317
U. Habibah, I.R. Lina and W.M. Kusumawinahyu

An operational matrix method for solving a class of nonlinear Volterra integral equations arising in steady activation of a skeletal muscle 336
S. Torkaman, M. Heydari and G.B. Loghmani

web site: https://ijnao.um.ac.ir
Email: ijnao@um.ac.ir
ISSN-Print: 2423-6977
ISSN-Online: 2423-6969