Iranian Journal of Numerical Analysis and Optimization

Iranian Journal of Numerical Analysis and Optimization

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.

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

On optimality and duality for multiobjective interval-valued programming problems with vanishing constraints

B. Japamala Rani*, I. Ahmad and K. Kummari

Abstract

In this study, we explore the theoretical features of a multiobjective interval-valued programming problem with vanishing constraints. In view of this, we have defined a multiobjective interval-valued programming problem with vanishing constraints in which the objective functions are considered to be interval-valued functions, and we define an LU-efficient solution by employing partial ordering relations. Under the assumption of generalized convexity, we investigate the optimality conditions for a (weakly) LU-efficient solution to a multiobjective interval-valued programming problem with vanishing constraints. Furthermore, we establish Wolfe and Mond–Weir duality results under appropriate convexity hypotheses. The study concludes with examples designed to validate our findings.

---

*Corresponding author

B. Japamala Rani

[a]Department of Mathematics, School of Science, GITAM-Hyderabad Campus, Hyderabad-502329, India. e-mail: bjapamalarani@gmail.com
[b]Department of Mathematics, St. Ann's College for Women, Mehdipatnam, Hyderabad-500028, India.

Izhar Ahmad

[a]Department of Mathematics, King Fahd University of Petroleum and Minerals, Dhahran-31261, Saudi Arabia. e-mail: drizhar@kfupm.edu.sa
[b]Center for Intelligent Secure Systems, King Fahd University of Petroleum and Minerals, Dhahran-31261, Saudi Arabia.

Krishna Kummari

Department of Mathematics, School of Science, GITAM-Hyderabad Campus, Hyderabad-502329, India. e-mail: krishna.maths@gmail.com

Keywords: Multiobjective interval-valued optimization problem; vanishing constraints; (weakly) LU-efficient solution; duality.

1 Introduction

In modern mathematical research, the concept of mathematical programming with vanishing constraints has emerged as a novel type of constrained optimization problems. Formal analysis was conducted by Achtziger and Kanzow [1]. Dorsch, Shikhman, and Stein [9] presented a topological analysis of mathematical programs with vanishing constraints and introduced the new concept of a T-stationary point. By applying the concept of local regularization to mathematical programs with vanishing constraints, Hohesiel, Kanzow, and Schwartz [13] derived a new solution method for solving such a class of optimization problems and proved several convergence results. Later, to compute the mathematical problems involving vanishing constraints numerically, Hoheisel et al. [14] investigated and compared four regularization methods, each impacted by a single regularization parameter. The study of mathematical programming with vanishing constraints has a wide range of real-world applications, including the development of robot motion plans [8, 19], the design of optimal truss topologies for mechanical structures [11], and the design of nonlinear optimal control problems for mixed integers [20]. A multiobjective programming problem involves minimizing multiple objectives over a set of feasible solutions. Multiobjective programming is challenging due to the fact that the objectives for vector optimization problems compete with each other, and an improvement on one objective can reduce goals for other objectives. There is an enormous amount of literature on optimal conditions and numerous kinds of dualities in multiobjective programming problems (see, for example, [7, 22, 23]). A constraint qualification is an element critical to the existence of Lagrange multipliers in multiobjective optimization problems, as it allows Karush–Kuhn–Tucker optimality conditions to hold, thereby assisting with and enhancing optimization algorithms design. There have been several recent articles published on optimality, stationarity, criticality, and constraint qualification; for instance, we refer to [10, 12, 17]. Jayswal and Singh [18] studied about modified objective function approach for an equivalent η-approximated multiobjective optimization problem with vanishing constraints and also discussed saddle point criteria. The class of differentiable semi-infinite multiobjective programming problems with vanishing constraints was discussed by Antczak [4].

Using separate considerations of minimization and maximization, Ishibuchi and Tanaka [16] investigated multiobjective optimization problems in which the objective functions are interval-valued and developed an ordering relationship between two closed intervals. A general methodology proposed by Urli and Nadeau [26] provides a way of formulating the non-deterministic multiobjective linear programming problem with interval coefficients in a de-

terministic way and then solving it with an interactive approach. Under certain convexity assumptions, The Karush–Kuhn–Tucker necessary optimality conditions for nonlinear differentiable multiobjective programming problems with an interval-valued objective and constraint functions were derived by Hosseinzade and Hassanpour [15]. Studies on optimality conditions and different types of duality for multiobjective programming problems with interval objective function are quite widespread (refer to [6, 27, 28, 15, 21]). In this paper, we aim to investigate the optimality conditions and the duality results for multiobjective interval-valued programming problems with vanishing constraints under the Abadie constraint qualification.

Following is an outline of the rest of this paper: Section 2 consists of some basic definitions, background material, and the necessary optimality conditions. Section 3 represents the sufficient optimality conditions for multiobjective interval-valued optimization problems with vanishing constraints. In Sections 4 and 5, Wolfe type dual and Mond–Weir type dual are presented, and appropriate duality results are also discussed. Section 6 explores special cases. Finally, the paper is concluded in Section 7.

2 Preliminaries

This section contains a list of notations and basic definitions which will be used throughout the article. Let R^n be the Euclidean space with n-dimensions and R^n_+ be its nonnegative orthant. For a given a, $\Theta(a)$ is the system of the neighborhoods of a. For $A \subseteq R^n$, $span A$ and $pos A$ stands for its linear hull and convex cone (containing the origin) of A, respectively. Let $A \neq \phi$ and let the contingent cone of set A at the point a, be denoted by $\mathbb{T}(A, a)$. Let $I(R)$ be the set of all closed and bounded intervals in R. For the case where $\Lambda_1 \in I(R)$ is a closed interval, we use the notation $\Lambda_1 = [\alpha_0^L, \alpha_0^U]$, where α_0^L and α_0^U represent the minimum and maximum values of Λ_1, respectively. Let

$$\Lambda_1 = [\alpha_0^L, \alpha_0^U], \quad \Lambda_2 = [\beta_0^L, \beta_0^U] \in I(R).$$

Then we have

(i) $\Lambda_1 + \Lambda_2 = \{\alpha_0 + \beta_0 \mid \alpha_0 \in \Lambda_1 \text{ and } \beta_0 \in \Lambda_2\} = [\alpha_0^L + \beta_0^L, \alpha_0^U + \beta_0^U]$,

(ii) $-\Lambda_1 = \{\alpha_0 \mid \alpha_0 \in \Lambda_1\} = [-\alpha_0^U, -\alpha_0^L]$,

(iii) $\Lambda_1 - \Lambda_2 = \Lambda_1 + (-\Lambda_2) = [\alpha_0^L - \beta_0^U, \alpha_0^U - \beta_0^L]$,

(iv) $k\Lambda_1 = \{k\alpha_0 \mid \alpha_0 \in \Lambda_1\} = \begin{cases} [k\alpha_0^L, k\alpha_0^U], & \text{if } k \geq 0, \\ [k\alpha_0^U, k\alpha_0^L], & \text{if } k < 0, \end{cases}$ where k is a real number.

The real number $k \in R$ is equivalent to the closed interval $\Lambda_{1_k} = [k, k]$. Let $\Lambda_1 = [\alpha_0^L, \alpha_0^U] \in I(R)$ be a closed interval. We write the sum of an interval $\Lambda_1 \in I(R)$ and a real number k as $\Lambda_1 + \Lambda_{1_k}$. Thus, $\Lambda_1 + k = \Lambda_1 + \Lambda_{1_k} = [\alpha_0^L + k, \alpha_0^U + k]$.

For $\Lambda_1 = [\alpha_0^L, \alpha_0^U]$ and $\Lambda_2 = [\beta_0^L, \beta_0^U]$, the order relation $\preceq_{LU}$ is defined as follows:

(i) $\Lambda_1 \preceq_{LU} \Lambda_2$ if and only if $\alpha_0^L \leq \beta_0^L$ and $\alpha_0^U \leq \beta_0^U$.

(ii) $\Lambda_1 \prec_{LU} \Lambda_2$ if and only if $\Lambda_1 \preceq_{LU} \Lambda_2$ and $\Lambda_1 \neq \Lambda_2$.
It is obvious that, $\Lambda_1 \prec_{LU} \Lambda_2$ if and only if

$$\alpha_0^L < \beta_0^L \text{ and } \alpha_0^U < \beta_0^U,$$

$$\text{or,} \quad \alpha_0^L \leq \beta_0^L \text{ and } \alpha_0^U < \beta_0^U,$$

$$\text{or,} \quad \alpha_0^L < \beta_0^L \text{ and } \alpha_0^U \leq \beta_0^U.$$

Furthermore, for $\dot{u}, \dot{v} \in R^m$, we use the following notations:

(i). $\dot{u} \prec \dot{v} \Leftrightarrow \dot{u}_i < \dot{v}_i$, for all $i \in \{1, 2, \ldots, m\}$, $\dot{u} \not\prec \dot{v}$ is the negation of $\dot{u} \prec \dot{v}$

(ii). $\dot{u} \preceq \dot{v} \Leftrightarrow \begin{cases} \dot{u}_i \leq \dot{v}_i, \text{ for all } i \in \{1, 2, \ldots, m\} \\ \dot{u}_{i_0} < \dot{v}_{i_0}, \text{ for at least one } i_0 \in \{1, 2, \ldots, m\}, \end{cases}$
$\dot{u} \not\preceq \dot{v}$ is the negation of $\dot{u} \preceq \dot{v}$.

In the present analysis, we consider the following differentiable vector optimization problem with multiple interval-valued objective function with vanishing constraints (MIVVC):

$$\text{MIVVC} \quad \min \quad \vartheta(\xi) = (\vartheta_1(\xi), \vartheta_2(\xi), \ldots, \vartheta_m(\xi))$$
$$\text{subject to}$$
$$\tau_i(\xi) \leq 0, \quad \text{for all } i = 1, 2, \ldots, p,$$
$$\sigma_i(\xi) = 0, \quad \text{for all } i = 1, 2, \ldots, q,$$
$$\rho_i(\xi) \geq 0, \quad \text{for all } i = 1, 2, \ldots, r,$$
$$\omega_i(\xi)\rho_i(\xi) \leq 0, \quad \text{for all } i = 1, 2, \ldots, r,$$

where each $\vartheta_i : R^n \to I(R), i \in T = \{1, 2, \ldots, m\}$ is an interval-valued function; that is, $\vartheta_i(\xi) = [\vartheta_i^L(\xi), \vartheta_i^U(\xi)], i \in T$ and $\tau_i(i = 1, 2, \ldots, p)$, $\sigma_i(i = 1, 2, \ldots, q)$, $\rho_i, \omega_i(i = 1, 2, \ldots, r)$ are assumed to be continuously differentiable functions from $R^n \to R$. Let us denote $T_\tau := \{1, 2, \ldots, p\}$, $T_\sigma := \{1, 2, \ldots, q\}$, and $T_r := \{1, 2, \ldots, r\}$. The feasible solution set of MIVVC is given by

$$\mathbb{F}_{\mathbb{VC}} = \left\{ \xi \in R^n \mid \tau_i(\xi) \leq 0, \text{ for all } i = 1, 2, \ldots, p, \right.$$

$$\left. \begin{aligned} &\sigma_i(\xi) = 0, \quad \text{for all } i = 1, 2, \ldots, q, \\ &\rho_i(\xi) \geq 0, \quad \text{for all } i = 1, 2, \ldots, r, \\ &\omega_i(\xi)\rho_i(\xi) \leq 0, \quad \text{for all } i = 1, 2, \ldots, r \end{aligned} \right\}.$$

Definition 1. A point $a \in \mathbb{F}_{\mathbb{VC}}$ is said to be a locally LU-efficient solution of MIVVC, if there exists a neighborhood $U \in \Theta(a)$ such that there is no $\xi \in \mathbb{F}_{\mathbb{VC}} \cap \mathbb{U}$ satisfying

$$\vartheta(\xi) \preceq_{LU} \vartheta(a).$$

Definition 2. A point $a \in \mathbb{F}_{\mathbb{VC}}$ is said to be a locally weakly LU-efficient solution of MIVVC, if there exists a neighborhood $U \in \Theta(a)$ such that there is no $\xi \in \mathbb{F}_{\mathbb{VC}} \cap \mathbb{U}$ satisfying

$$\vartheta(\xi) \prec_{LU} \vartheta(a).$$

Let $a \in \mathbb{F}_{\mathbb{VC}}$ be any feasible solution of the MIVVC. The following index sets will be used:

$$T_+(a) := \{i \in T_r \mid \rho_i(a) > 0\},$$
$$T_0(a) := \{i \in T_r \mid \rho_i(a) = 0\}.$$

Furthermore, the index set T_+ can be divided into the following subsets

$$T_{+0}(a) := \{i \in T_r \mid \rho_i(a) > 0, \ \omega_i(a) = 0\},$$

$$T_{+-}(a) := \{i \in T_r \mid \rho_i(a) > 0, \ \omega_i(a) < 0\}.$$

Similarly, the index set T_0 can be partitioned in the following way

$$T_{0+}(a) := \{i \in T_r \mid \rho_i(a) = 0, \ \omega_i(a) > 0\},$$

$$T_{00}(a) := \{i \in T_r \mid \rho_i(a) = 0, \ \omega_i(a) = 0\},$$

$$T_{0-}(a) := \{i \in T_r \mid \rho_i(a) = 0, \ \omega_i(a) < 0\}.$$

Definition 3. A point $a \in \mathbb{F}_{\mathbb{VC}}$ is said to be a strong stationary point of MIVVC if and only if there exists $(\alpha^L, \alpha^U, \lambda^\tau, \lambda^\sigma, \lambda^\omega, \lambda^\rho) \in R_+^m \times R_+^m \times R^p \times R^q \times R^r \times R^r$ with $\sum\limits_{i \in T} \left(\alpha_i^L + \alpha_i^U\right) = 1, \lambda_{T_+(a)}^\rho = 0, \lambda_{T_{00}(a) \cup T_{0-}(a)}^\rho \geq 0,$
$\lambda_{T_{+-}(a) \cup T_{0+}(a) \cup T_{00}(a) \cup T_{0-}(a)}^\omega = 0$ and $\lambda_{T_{+0}(a)}^\omega \geq 0$ such that

$$\sum_{i \in T} \alpha_i^L \nabla \vartheta_i^L(a) + \sum_{i \in T} \alpha_i^U \nabla \vartheta_i^U(a) + \sum_{i \in T_\tau} \lambda_i^\tau \nabla \tau_i(a) + \sum_{i \in T_\sigma} \lambda_i^\sigma \nabla \sigma_i(a)$$
$$+ \sum_{i \in T_{+0}} \lambda_i^\omega \nabla \omega_i(a) - \sum_{i \in T_{0+} \cup T_{00} \cup T_{0-}} \lambda_i^\rho \nabla \rho_i(a) = 0.$$

Definition 4. A point $a \in \mathbb{F}_{\mathbb{VC}}$ is said to be a VC-stationary point of MIVVC if and only if there exists $(\alpha^L, \alpha^U, \lambda^\tau, \lambda^\sigma, \lambda^\omega, \lambda^\rho) \in R_+^m \times R_+^m \times R^p \times R^q \times$

$R^r \times R^r$ with $\sum\limits_{i\in T} \left(\alpha_i^L + \alpha_i^U\right) = 1, \lambda_{T_+(a)}^\rho = 0, \lambda_{T_{00}(a)\cup T_{0-}(a)}^\rho \geq 0,$

$\lambda_{T_{+-}(a)\cup T_{0+}(a)\cup T_{00}(a)\cup T_{0-}(a)}^\omega = 0$ and $\lambda_{T_{+0}(a)\cup T_{00}(a)}^\omega \geq 0$ such that

$$\sum_{i\in T} \alpha_i^L \nabla \vartheta_i^L(a) + \sum_{i\in T} \alpha_i^U \nabla \vartheta_i^U(a) + \sum_{i\in T_\tau} \lambda_i^\tau \nabla \tau_i(a) + \sum_{i\in T_\sigma} \lambda_i^\sigma \nabla \sigma_i(a)$$

$$- \sum_{i\in T_{0+}\cup T_{00}\cup T_{0-}} \lambda_i^\rho \nabla \rho_i(a) + \sum_{i\in T_{+0}} \lambda_i^\omega \nabla \omega_i(a) = 0.$$

For $a \in \mathbb{F}_{\mathbb{VC}}$ and $(\lambda^\tau, \lambda^\sigma, \lambda^\omega, \lambda^\rho) \in R^p \times R^q \times R^r \times R^r$, let us define

$$T_\tau^+(a) := \{i \in T_\tau(a) \mid \lambda_i^\tau > 0\},$$

$$T_\sigma^+(a) := \{i \in T_\sigma(a) \mid \lambda_i^\sigma > 0\},\ T_\sigma^-(a) := \{i \in T_\sigma(a) \mid \lambda_i^\sigma < 0\},$$

$$\hat{T}_+^+(a) := \{i \in T_+(a) \mid \lambda_i^\rho > 0\},$$

$$\hat{T}_0^+(a) := \{i \in T_0(a) \mid \lambda_i^\rho > 0\},\ \hat{T}_0^-(a) := \{i \in T_0(a) \mid \lambda_i^\rho < 0\},$$

$$\hat{T}_{0+}^+(a) := \{i \in T_{0+}(a) \mid \lambda_i^\rho > 0\},\ \hat{T}_{0+}^-(a) := \{i \in T_{0+}(a) \mid \lambda_i^\rho < 0\},$$

$$\hat{T}_{00}^+(a) := \{i \in T_{00}(a) \mid \lambda_i^\rho > 0\},\ \hat{T}_{00}^-(a) := \{i \in T_{00}(a) \mid \lambda_i^\rho < 0\},$$

$$\hat{T}_{0-}^+(a) := \{i \in T_{0-}(a) \mid \lambda_i^\rho > 0\},$$

$$T_{+0}^+(a) := \{i \in T_{+0}(a) \mid \lambda_i^\omega > 0\},\ T_{+0}^-(a) := \{i \in T_{+0}(a) \mid \lambda_i^\omega < 0\},$$

$$T_{+-}^+(a) := \{i \in T_{+-}(a) \mid \lambda_i^\omega > 0\},$$

$$T_{0+}^+(a) := \{i \in T_{0+}(a) \mid \lambda_i^\omega > 0\},\ T_{0+}^-(a) := \{i \in T_{0+}(a) \mid \lambda_i^\omega < 0\},$$

$$T_{00}^+(a) := \{i \in T_{00}(a) \mid \lambda_i^\omega > 0\},\ T_{00}^-(a) := \{i \in T_{00}(a) \mid \lambda_i^\omega < 0\},$$

$$T_{0-}^+(a) := \{i \in T_{0-}(a) \mid \lambda_i^\omega > 0\}.$$

Definition 5. Let $a \in \mathbb{F}_{\mathbb{VC}}$.

(i) The linearized cone of MIVVC at a is
$L(a) := \{d \in \mathbb{R}^n \mid \langle \nabla \tau_i(a),\ d \rangle \leq 0\ (i \in T_\tau), \langle \nabla \sigma_i(a),\ d \rangle = 0\ (i \in T_\sigma),$
$\langle \nabla \rho_i(a),\ d \rangle = 0\ (i \in T_{0+}), \langle \nabla \rho_i(a),\ d \rangle \geq 0\ (i \in T_{00} \cup T_{0-}),$
$\langle \nabla \omega_i(a),\ d \rangle \leq 0\ (i \in T_{+0})\}.$

(ii) The VC-linearized cone of MIVVC at a is
$L_{VC}(a) := \{d \in \mathbb{R}^n \mid \langle \nabla \tau_i(a),\ d \rangle \leq 0\ (i \in T_\tau), \langle \nabla \sigma_i(a),\ d \rangle = 0\ (i \in T_\sigma),$
$\langle \nabla \rho_i(a),\ d \rangle = 0\ (i \in T_{0+}), \langle \nabla \rho_i(a),\ d \rangle \geq 0\ (i \in T_{00} \cup T_{0-}),$
$\langle \nabla \omega_i(a),\ d \rangle \leq 0\ (i \in T_{+0} \cup T_{00})\}.$

Definition 6. The Abadie constraint qualification (MIVVC-ACQ) is said to hold at $a \in \mathbb{F}_{\mathbb{VC}}$ if

$$L(a) \subseteq \mathbb{T}(\mathbb{F}_{\mathbb{VC}}, a).$$

Definition 7. The vanishing Abadie constraint qualification (MIVVC-VACQ) is said to hold at $a \in \mathbb{F}_{\mathbb{VC}}$ if

$$L_{VC}(a) \subseteq \mathbb{T}(\mathbb{F}_{\mathbb{VC}}, a).$$

The following theorem can be written in a similar way to Proposition 1 of Tung [25].

Theorem 1 (Necessary optimality conditions). Let ξ_0 be a locally weakly LU-efficient solution of primal problem MIVVC and also further assume that if MIVVC-VACQ holds at ξ_0 and the set

$$\Delta_1 := pos \left(\bigcup_{i \in T_\tau} \nabla \tau_i(\xi_0) \ \cup \bigcup_{i \in T_{00} \cup T_{0-}} (-\nabla \rho_i(\xi_0)) \ \cup \bigcup_{i \in T_{+0} \cup T_{00}} \nabla \omega_i(\xi_0) \right)$$

$$+ \, span \left(\bigcup_{i \in T_\sigma} \nabla \sigma_i(\xi_0) \ \cup \bigcup_{i \in T_{0+}} \nabla \rho_i(\xi_0) \right)$$

is closed, then there exists $(\alpha^L, \alpha^U, \lambda^\tau, \lambda^\sigma, \lambda^\omega, \lambda^\rho) \in R_+^m \times R_+^m \times R^p \times R^q \times R^r \times R^r$ with $\sum\limits_{i \in T} (\alpha_i^L + \alpha_i^U) = 1, \lambda^\rho_{T_+(\xi_0)} = 0, \lambda^\rho_{T_{00}(\xi_0) \cup T_{0-}(\xi_0)} \geq 0, \lambda^\omega_{T_{+-}(\xi_0) \cup T_{0+}(\xi_0) \cup T_{0-}(\xi_0)} = 0$ and $\lambda^\omega_{T_{+0}(\xi_0) \cup T_{00}(\xi_0)} \geq 0$ such that

$$\sum_{i \in T} \alpha_i^L \nabla \vartheta_i^L(a) + \sum_{i \in T} \alpha_i^U \nabla \vartheta_i^U(a) + \sum_{i \in T_\tau} \lambda_i^\tau \nabla \tau_i(a) + \sum_{i \in T_\sigma} \lambda_i^\sigma \nabla \sigma_i(a)$$

$$- \sum_{i \in T_r} \lambda_i^\rho \nabla \rho_i(a) + \sum_{i \in T_r} \lambda_i^\omega \nabla \omega_i(a) = 0.$$

3 Sufficient optimality conditions

In this section, we establish sufficient optimality conditions for the problem MIVVC using the concept of generalized convexity.

Theorem 2. Let ξ_0 be a strong stationary point of MIVVC. Suppose that $\hat{T}_{0+}^- \cup T_{+0}^+ = \phi$ and τ_i $(i \in T_\tau)$, σ_i $(i \in T_\sigma^+)$, $-\sigma_i$ $(i \in T_\sigma^-)$, ω_i $(i \in T_{+0}^+)$, $-\rho_i$ $(i \in \hat{T}_{0+}^+ \cup \hat{T}_{00}^+ \cup \hat{T}_{0-}^+)$ are quasiconvex functions at ξ_0. If $\sum\limits_{i \in T} \alpha_i^L \vartheta_i^L(\cdot) + \sum\limits_{i \in T} \alpha_i^U \vartheta_i^U(\cdot)$ is pseudoconvex function at ξ_0, then ξ_0 is an LU-efficient solution of MIVVC.

Proof. Since ξ_0 is a strong stationary point of MIVVC, there exists $(\alpha^L, \alpha^U, \lambda^\tau, \lambda^\sigma, \lambda^\omega, \lambda^\rho) \in R_+^m \times R_+^m \times R^p \times R^q \times R^r \times R^r$ with $\sum\limits_{i \in T} (\alpha_i^L + \alpha_i^U) = 1, \lambda^\rho_{T_+} = 0, \lambda^\rho_{T_{00} \cup T_{0-}} \geq 0, \lambda^\omega_{T_{+-} \cup T_{0+} \cup T_{0-}} = 0$ and $\lambda^\omega_{T_{+0}} \geq 0$ such that

$$\sum_{i \in T} \alpha_i^L \nabla \vartheta_i^L(\xi_0) + \sum_{i \in T} \alpha_i^U \nabla \vartheta_i^U(\xi_0) + \sum_{i \in T_\tau} \lambda_i^\tau \nabla \tau_i(\xi_0) + \sum_{i \in T_\sigma} \lambda_i^\sigma \nabla \sigma_i(\xi_0)$$

$$- \sum_{i \in T_r} \lambda_i^\rho \nabla \rho_i(\xi_0) + \sum_{i \in T_r} \lambda_i^\omega \nabla \omega_i(\xi_0) = 0. \tag{1}$$

For an arbitrary $\xi \in \mathbb{F}_{\mathbb{VC}}$, we get $\tau_i(\xi) \leq 0 = \tau_i(\xi_0)$ for each $i \in T_\tau$. Thus the quasiconvexity at ξ_0 of τ_i $(i \in T_\tau)$ gives that

$$\langle \nabla \tau_i(\xi_0), \xi - \xi_0 \rangle \leq 0, \quad \text{for all } i \in T_\tau,$$

consequently, together with $\lambda_i^\tau \in \mathbb{R}^p$ leads that

$$\left\langle \sum_{i \in T_\tau} \lambda_i^\tau \nabla \tau_i(\xi_0), \xi - \xi_0 \right\rangle \leq 0. \tag{2}$$

We deduce from $\xi, \xi_0 \in \mathbb{F}_{\mathbb{VC}}$ that $\sigma_i(\xi) = \sigma_i(\xi_0) = 0, \quad \text{for all } i \in T_\sigma$, and hence,

$$\sigma_i(\xi) \leq \sigma_i(\xi_0) = 0, \quad \text{for all } i \in T_\sigma^+ \text{ and } -\sigma_i(\xi) \leq -\sigma_i(\xi_0) = 0, \quad \text{for all } i \in T_\sigma^-.$$

The above inequalities along with the quasiconvexity at ξ_0 of σ_i $(i \in T_\sigma^+)$ and $-\sigma_i$ $(i \in T_\sigma^-)$ ensure that

$$\langle \nabla \sigma_i(\xi_0), \xi - \xi_0 \rangle \leq 0, \quad \text{for all } i \in T_\sigma^+ \text{ and } \langle -\nabla \sigma_i(\xi_0), \xi - \xi_0 \rangle \leq 0, \quad \text{for all } i \in T_\sigma^-.$$

Thus, taking into account the definitions of T_σ^+, T_σ^- results in

$$\left\langle \sum_{i \in T_\sigma} \lambda_i^\sigma \nabla \sigma_i(\xi_0), \xi - \xi_0 \right\rangle \leq 0. \tag{3}$$

Again, we deduce from $\xi \in \mathbb{F}_{\mathbb{VC}}$ that $-\rho_i(\xi) \leq 0, \omega_i(\xi) \geq 0, \quad \text{for all } i \in T_r$. Thus,

$$\begin{cases} -\rho_i(\xi) \leq -\rho_i(\xi_0), & i \in \hat{T}_{0+}^+ \cup \hat{T}_{00}^+ \cup \hat{T}_{0-}^+, \\ \omega_i(\xi) \leq \omega_i(\xi_0), & i \in T_{+0}^+. \end{cases}$$

Therefore, the quasiconvexity of $-\rho_i, i \in \hat{T}_{0+}^+ \cup \hat{T}_{00}^+ \cup \hat{T}_{0-}^+$ and $\omega_i, i \in T_{+0}^+$ at ξ_0 yields that

$$\langle -\nabla \rho_i(\xi_0), \xi - \xi_0 \rangle \leq 0, \quad \text{for all } i \in \hat{T}_{0+}^+ \cup \hat{T}_{00}^+ \cup \hat{T}_{0-}^+, \tag{4}$$

$$\langle \nabla \omega_i(\xi_0), \xi - \xi_0 \rangle \leq 0, \quad \text{for all } i \in T_{+0}^+. \tag{5}$$

As $T_{+0}^+ \cup \hat{T}_{0+}^- = \phi$, we presume from (1)–(5) that

$$\left\langle \sum_{i \in T} \alpha_i^L \nabla \vartheta_i^L(\xi_0) + \sum_{i \in T} \alpha_i^U \nabla \vartheta_i^U(\xi_0), \xi - \xi_0 \right\rangle$$

$$= \left\langle \sum_{i \in T_\tau} \lambda_i^\tau \nabla \tau_i(\xi_0) + \sum_{i \in T_\sigma} \lambda_i^\sigma \nabla \sigma_i(\xi_0) - \sum_{i \in T_r} \lambda_i^\rho \nabla \rho_i(\xi_0) + \right.$$

$$\left. \sum_{i \in T_r} \lambda_i^\omega \nabla \omega_i(\xi_0), \xi - \xi_0 \right\rangle \geq 0, \tag{6}$$

for all $\xi \in \mathbb{F}_{\mathbb{VC}}$.

On the contrary, suppose ξ_0 is not an LU-efficient solution of MIVVC. This leads to the existence of a feasible point $\tilde{\xi} \in \mathbb{F}_{\mathbb{VC}}$ such that

$$\vartheta(\tilde{\xi}) \preceq_{LU} \vartheta(\xi_0);$$

that is, for $i \in T$,

$$\left\{ \begin{array}{l} \vartheta_i^L(\tilde{\xi}) < \vartheta_i^L(\xi_0) \\ \vartheta_i^U(\tilde{\xi}) \leq \vartheta_i^U(\xi_0) \end{array} \right. , \text{ or } \left\{ \begin{array}{l} \vartheta_i^L(\tilde{\xi}) \leq \vartheta_i^L(\xi_0) \\ \vartheta_i^U(\tilde{\xi}) < \vartheta_i^U(\xi_0) \end{array} \right. , \text{ or } \left\{ \begin{array}{l} \vartheta_i^L(\tilde{\xi}) < \vartheta_i^L(\xi_0) \\ \vartheta_i^U(\tilde{\xi}) < \vartheta_i^U(\xi_0) \end{array} \right. .$$

From the fact $\alpha^L \in R_+^m, \alpha^U \in R_+^m$ with $\sum_{i \in T} \left(\alpha_i^L + \alpha_i^U \right) = 1$, then above inequalities together yield

$$\sum_{i \in T} \alpha_i^L \vartheta_i^L(\tilde{\xi}) + \sum_{i \in T} \alpha_i^U \vartheta_i^U(\tilde{\xi}) < \sum_{i \in T} \alpha_i^L \vartheta_i^L(\xi_0) + \sum_{i \in T} \alpha_i^U \vartheta_i^U(\xi_0),$$

which by the pseudoconvexity of $\sum_{i \in T} \alpha_i^L \vartheta_i^L(\cdot) + \sum_{i \in T} \alpha_i^U \vartheta_i^U(\cdot)$, we obtain

$$\left\langle \sum_{i \in T} \alpha_i^L \nabla \vartheta_i^L(\xi_0) + \sum_{i \in T} \alpha_i^U \nabla \vartheta_i^U(\xi_0), \tilde{\xi} - \xi_0 \right\rangle < 0,$$

contradicting to (6). □

Theorem 3. Let ξ_0 be a strong stationary point of MIVVC. Suppose that $\hat{T}_{0+}^- \cup T_{+0}^+ = \phi$ and τ_i $(i \in T_\tau), \sigma_i$ $(i \in T_\sigma^+), -\sigma_i$ $(i \in T_\sigma^-), \omega_i$ $(i \in T_{+0}^+), -\rho_i$ $(i \in \hat{T}_{0+}^+ \cup \hat{T}_{00}^+ \cup \hat{T}_{0-}^+)$ are quasiconvex functions at ξ_0. If $\sum_{i \in T} \alpha_i^L \vartheta_i^L(\cdot) + \sum_{i \in T} \alpha_i^U \vartheta_i^U(\cdot)$ is strictly pseudoconvex function at ξ_0, then ξ_0 is a weakly LU-efficient solution of MIVVC.

Proof. Similar to the proof of Theorem 2, we get

$$\left\langle \sum_{i \in T} \alpha_i^L \nabla \vartheta_i^L(\xi_0) + \sum_{i \in T} \alpha_i^U \nabla \vartheta_i^U(\xi_0), \xi - \xi_0 \right\rangle$$

$$= \left\langle \sum_{i \in T_\tau} \lambda_i^\tau \nabla \tau_i(\xi_0) + \sum_{i \in T_\sigma} \lambda_i^\sigma \nabla \sigma_i(\xi_0) - \sum_{i \in T_r} \lambda_i^\rho \nabla \rho_i(\xi_0) + \right.$$

$$\left. \sum_{i \in T_r} \lambda_i^\omega \nabla \omega_i(\xi_0), \xi - \xi_0 \right\rangle \geq 0. \tag{7}$$

Reasoning by contraposition, assume that ξ_0 is not a weakly LU-efficient solution. Then there exists a feasible point $\tilde{\xi}$ satisfying

$$\vartheta(\tilde{\xi}) \prec_{LU} \vartheta(\xi_0);$$

that is, for $i \in T$,

$$\begin{cases} \vartheta_i^L(\tilde{\xi}) < \vartheta_i^L(\xi_0), \\ \vartheta_i^U(\tilde{\xi}) < \vartheta_i^U(\xi_0). \end{cases}$$

From the fact that $\alpha^L \in R_+^m, \alpha^U \in R_+^m$ with $\sum_{i \in T} \left(\alpha_i^L + \alpha_i^U \right) = 1$, and by the above inequalities, we get

$$\sum_{i \in T} \alpha_i^L \vartheta_i^L(\tilde{\xi}) + \sum_{i \in T} \alpha_i^U \vartheta_i^U(\tilde{\xi}) < \sum_{i \in T} \alpha_i^L \vartheta_i^L(\xi_0) + \sum_{i \in T} \alpha_i^U \vartheta_i^U(\xi_0).$$

By using the strictly pseudoconvexity of $\sum_{i \in T} \alpha_i^L \vartheta_i^L(\cdot) + \sum_{i \in T} \alpha_i^U \vartheta_i^U(\cdot)$ at $\tilde{\xi}$ on $\mathbb{F}_{\mathbb{VC}}$, we get

$$\left\langle \sum_{i \in T} \alpha_i^L \nabla \vartheta_i^L(\xi_0) + \sum_{i \in T} \alpha_i^U \nabla \vartheta_i^U(\xi_0), \tilde{\xi} - \xi_0 \right\rangle < 0,$$

contradicting to (7).   □

Now, we verify the sufficient optimality conditions by an example.

Example 1. Consider the following multiobjective interval-valued programming problem with vanishing constraints (MIVVC-1):

$$MIVVC - 1 \; \mathbb{R}_+ - \min \; \vartheta(\xi) = (\vartheta_1(\xi), \vartheta_2(\xi))$$
$$= \left([4\xi^2 - \xi, 4\xi^2 + \xi + 1], \; [\xi^2 - 2\xi, \xi^4 + 2\xi] \right)$$
$$\text{subject to}$$
$$\rho_1(\xi) = \xi \geq 0,$$
$$\omega_1(\xi)\rho_1(\xi) = (-1 - \xi)\xi \leq 0,$$

where $\vartheta_1^L(\xi) = 4\xi^2 - \xi, \vartheta_2^L(\xi) = \xi^2 - 2\xi, \vartheta_1^U(\xi) = 4\xi^2 + \xi + 1, \vartheta_2^U(\xi) = \xi^4 + 2\xi$, which is in the form of MIVVC with $m = 2, n = 1, p = q = 0$, and $r = 1$.

The feasible region of MIVVC-1 is $\mathbb{F}_{\mathbb{VC}1} = \{\xi \in R \mid \rho_1(\xi) \geq 0, \; \omega_1(\xi)\rho_1(\xi) \leq 0\}$.
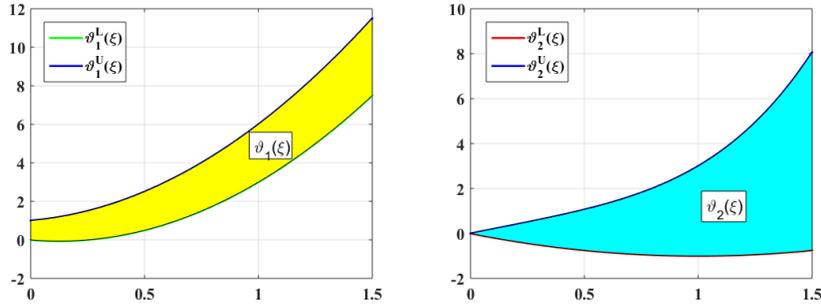
(a) Graphical view of $\vartheta_1(\xi) = [\vartheta_1^L(\xi), \vartheta_1^U(\xi)]$ (b) Graphical view of $\vartheta_2(\xi) = [\vartheta_2^L(\xi), \vartheta_2^U(\xi)]$
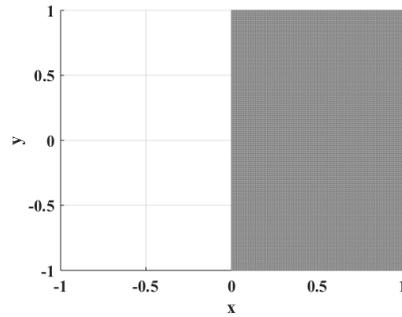
Graphical view of the feasible region of MIVVC-1

Note that $\xi_0 = 0$ is a feasible solution of MIVVC-1. By simple calculations, we get $\mathbb{T}(\mathbb{F}_{\mathbb{VC}1}, \xi_0) = \mathbb{F}_{\mathbb{VC}1}$, $\nabla\vartheta_1^L(\xi_0) = \{-1\}$, $\nabla\vartheta_2^L(\xi_0) = \{-2\}$, $\nabla\vartheta_1^U(\xi_0) = \{1\}$, $\nabla\vartheta_2^U(\xi_0) = \{2\}$, $\nabla\rho_1(\xi_0) = \{1\}$, $\nabla\omega_1(\xi_0) = \{-1\}$, $T_+ = T_{0+} = T_{0-} = \phi$, $T_{00} = \{1\}$,

$$\left(\bigcup_{i \in T_{00}} (-\nabla\rho_i(\xi_0))\right)^- = \{\xi \in R \mid \xi \geq 1\},$$

$$\left(\bigcup_{i \in T_{00}} (\nabla\omega_i(\xi_0))\right)^- = \{\xi \in R \mid \xi \geq 1\},$$

$$\left(\bigcup_{i \in T_{00}} (-\nabla\rho_i(\xi_0))\right)^- \cap \left(\bigcup_{i \in T_{00}} \nabla\omega_i(\xi_0)\right)^- = \{\xi \in R \mid \xi \geq 1\}.$$

Hence,

$$\left(\bigcup_{i\in T_{00}}(-\nabla\rho_i(\xi_0))\right)^- \cap \left(\bigcup_{i\in T_{00}}(\nabla\omega_i(\xi_0))\right)^- \subset T(\mathbb{F}_{\mathbb{VC}1},\xi_0),$$

yields that MIVVC-VACQ satisfies at ξ_0. Moreover,

$$\Delta_1 := pos\left(\bigcup_{i\in T_{00}}(-\nabla\rho_i(\xi_0))\cup\bigcup_{i\in T_{00}}\nabla\omega_i(\xi_0)\right) = \{\xi\in R \mid \xi \geq -1\}$$

is closed. Thus, all assumptions in Theorem 1 are satisfied. Then there exist $\alpha_1^L = \alpha_2^L = \frac{1}{2}, \alpha_1^U = \alpha_2^U = \frac{1}{2}, \lambda_1^\rho = 0, \lambda_1^\omega = 0$ such that (1) is satisfied at $\xi_0 = 0$ for the problem MIVVC-1. Furthermore, it can be easily observed that the hypothesis of Theorem 3 hold at $\xi_0 = 0$, and owing to the fact that for $\xi \neq \xi_0$, $\vartheta(\xi) \not\prec_{LU} \vartheta(\xi_0)$. Then, we assert that ξ_0 is a locally weakly LU-efficient solution of MIVVC-1.

4 The Wolfe type duality

In this section, we present the Wolfe type dual problem to MIVVC assuming that all the functions to be convex. For a given $\bar{u}$, $\Theta(\bar{u})$ is the system of the neighborhoods of $\bar{u}$. For $\xi_0 \in \mathbb{F}_{\mathbb{VC}}, (u, \alpha^L, \alpha^U, \lambda^\tau, \lambda^\sigma, \lambda^\omega, \lambda^\rho) \in R^n \times R_+^m \times R_+^m \times R^p \times R^q \times R^r \times R^r$ with $\sum_{i\in T}(\alpha_i^L + \alpha_i^U) = 1, \lambda_{T_+(\xi_0)}^\rho \geq 0, \lambda_{T_{0+}(\xi_0)}^\omega \leq 0,$ and $\lambda_{T_{+-}(\xi_0)\cup T_{0-}(\xi_0)}^\omega \geq 0$, we define

$$\mathcal{L}(u, \alpha^L, \alpha^U, \lambda^\tau, \lambda^\sigma, \lambda^\omega, \lambda^\rho) = \left(\vartheta_1(u) + \left(\sum_{i\in T_\tau}\lambda_i^\tau \tau_i(u) + \sum_{i\in T_\sigma}\lambda_i^\sigma \sigma_i(u)\right.\right.$$

$$\left.- \sum_{i\in T_r}\lambda_i^\rho \rho_i(u) + \sum_{i\in T_r}\lambda_i^\omega \omega_i(u)\right) e + \cdots$$

$$+\vartheta_m(u) + \left(\sum_{i\in T_\tau}\lambda_i^\tau \tau_i(u) + \sum_{i\in T_\sigma}\lambda_i^\sigma \sigma_i(u)-\right.$$

$$\left.\left.\sum_{i\in T_r}\lambda_i^\rho \rho_i(u) + \sum_{i\in T_r}\lambda_i^\omega \omega_i(u)\right) e\right),$$

where $e := (1,\ldots,1) \in R^m$. We consider the Wolfe type dual problem as follows:

$(WD_w(\xi_0)) \quad R_+^m - \max \mathcal{L}(u, \alpha^L, \alpha^U, \lambda^\tau, \lambda^\sigma, \lambda^\omega, \lambda^\rho)$

$\qquad$ subject to

$$\sum_{i\in T}\alpha_i^L \nabla \vartheta_i^L(u) + \sum_{i\in T}\alpha_i^U \nabla \vartheta_i^U(u) + \sum_{i\in T_\tau}\lambda_i^\tau \nabla \tau_i(u)$$

$$+ \sum_{i \in T_\sigma} \lambda_i^\sigma \nabla \sigma_i(u) - \sum_{i \in T_r} \lambda_i^\rho \nabla \rho_i(u) + \sum_{i \in T_r} \lambda_i^\omega \nabla \omega_i(u) = 0,$$

$$\sum_{i \in T} \left(\alpha_i^L + \alpha_i^U \right) = 1, \lambda_{T_+(\xi_0)}^\rho \geq 0, \lambda_{T_{0+}(\xi_0)}^\omega \leq 0 \text{ and}$$

$$\lambda_{T_{+-}(\xi_0) \cup T_{0-}(\xi_0)}^\omega \geq 0, (u, \alpha^L, \alpha^U, \lambda^\tau, \lambda^\sigma, \lambda^\omega, \lambda^\rho) \in$$
$$R^n \times R_+^m \times R_+^m \times R^p \times R^q \times R^r \times R^r.$$

The feasible set of $(WD_w(\xi_0))$ is defined by

$$\mathbb{F}_{\mathbb{VC}w}(\xi_0) := \Bigg\{ (u, \alpha^L, \alpha^U, \lambda^\tau, \lambda^\sigma, \lambda^\omega, \lambda^\rho) \in R^n \times R_+^m \times R_+^m \times R^p \times R^q \times R^r$$

$$\times R^r | \sum_{i \in T} \left(\alpha_i^L + \alpha_i^U \right) = 1, \lambda_{T_+(\xi_0)}^\rho \geq 0, \lambda_{T_{0+}(\xi_0)}^\omega \leq 0, \text{ and}$$

$$\lambda_{T_{+-}(\xi_0) \cup T_{0-}(\xi_0)}^\omega \geq 0, \sum_{i \in T} \alpha_i^L \nabla \vartheta_i^L(u) + \sum_{i \in T} \alpha_i^U \nabla \vartheta_i^U(u) +$$

$$\sum_{i \in T_\tau} \lambda_i^\tau \nabla \tau_i(u) + \sum_{i \in T_\sigma} \lambda_i^\sigma \nabla \sigma_i(u) - \sum_{i \in T_r} \lambda_i^\rho \nabla \rho_i(u) +$$

$$\sum_{i \in T_r} \lambda_i^\omega \nabla \omega_i(u) = 0 \Bigg\}.$$

The Wolfe type duality problem of MIVVC, which is not dependent on ξ_0, is

$$(WD_w): \qquad R_+^m - \max \mathcal{L}(\psi, \alpha^L, \alpha^U, \lambda^\tau, \lambda^\sigma, \lambda^\omega, \lambda^\rho)$$
$$\text{subject to}$$
$$(\psi, \alpha^L, \alpha^U, \lambda^\tau, \lambda^\sigma, \lambda^\omega, \lambda^\rho) \in \mathbb{F}_{\mathbb{VC}w} := \bigcap_{\xi_0 \in \mathbb{F}_{\mathbb{VC}}} \mathbb{F}_{\mathbb{VC}w}(\xi_0).$$

Definition 8. Let $\xi_0 \in \mathbb{F}_{\mathbb{VC}}$. Then $(\bar{u}, \bar{\alpha}^L, \bar{\alpha}^U, \bar{\lambda}^\tau, \bar{\lambda}^\sigma, \bar{\lambda}^\omega, \bar{\lambda}^\rho) \in \mathbb{F}_{\mathbb{VC}w}(\xi_0)$ is a locally LU-efficient solution of $(WD_w(\xi_0))$ (locally weakly LU-efficient solution of $(WD_w(\xi_0))$) if there exists $U \in \Theta(\bar{u})$ such that there is no $(u, \alpha^L, \alpha^U, \lambda^\tau, \lambda^\sigma, \lambda^\omega, \lambda^\rho) \in \mathbb{F}_{\mathbb{VC}w}(\xi_0) \cap U$ satisfying

$$\mathcal{L}(\bar{u}, \bar{\alpha}^L, \bar{\alpha}^U, \bar{\lambda}^\tau, \bar{\lambda}^\sigma, \bar{\lambda}^\omega, \bar{\lambda}^\rho) \preceq_{LU} \mathcal{L}(u, \alpha^L, \alpha^U, \lambda^\tau, \lambda^\sigma, \lambda^\omega, \lambda^\rho),$$
$$\left(\mathcal{L}(\bar{u}, \bar{\alpha}^L, \bar{\alpha}^U, \bar{\lambda}^\tau, \bar{\lambda}^\sigma, \bar{\lambda}^\omega, \bar{\lambda}^\rho) \prec_{LU} \mathcal{L}(u, \alpha^L, \alpha^U, \lambda^\tau, \lambda^\sigma, \lambda^\omega, \lambda^\rho) \right).$$

Theorem 4 (Weak Duality). Let $\xi \in \mathbb{F}_{\mathbb{VC}}$ and let $(\psi, \alpha^L, \alpha^U, \lambda^\tau, \lambda^\sigma, \lambda^\omega, \lambda^\rho) \in \mathbb{F}_{\mathbb{VC}w}$. Suppose that $\tau_i(i \in T_\tau^+(\xi)), \sigma_i(i \in T_\sigma^+(\xi)), -\sigma_i(i \in T_\sigma^-(\xi)), \rho_i(i \in \hat{T}_0^-(\xi)), -\rho_i(i \in \hat{T}_+^+(\xi) \cup \hat{T}_0^+(\xi)), \omega_i(i \in T_{+0}^+(\xi) \cup T_{+-}^+(\xi) \cup T_{00}^+(\xi) \cup T_{0-}^+(\xi)), -\omega_i(i \in T_{+0}^-(\xi) \cup T_{0+}^-(\xi) \cup T_{00}^-(\xi))$ are convex functions at ψ,

(i) If $\vartheta_i^L, \vartheta_i^U (i \in T)$ are convex functions at ψ, then

$$\vartheta(\xi) \nprec_{LU} \mathcal{L}(\psi, \alpha^L, \alpha^U, \lambda^\tau, \lambda^\sigma, \lambda^\omega, \lambda^\rho).$$

(ii) If $\vartheta_i^L, \vartheta_i^U (i \in T)$ are strictly convex functions at ψ, then

$$\vartheta(\xi) \npreceq_{LU} \mathcal{L}(\psi, \alpha^L, \alpha^U, \lambda^\tau, \lambda^\sigma, \lambda^\omega, \lambda^\rho).$$

Proof. For $\xi \in \mathbb{F}_{\mathbb{VC}}$ and $(\psi, \alpha^L, \alpha^U, \lambda^\tau, \lambda^\sigma, \lambda^\omega, \lambda^\rho) \in \mathbb{F}_{\mathbb{VC}w} = \bigcap\limits_{\xi_0 \in \mathbb{F}_{\mathbb{VC}}} \mathbb{F}_{\mathbb{VC}w}(\xi_0)$, one gets

$$\tau_i(\xi) \leq 0 \ (i \in T_\tau), \sigma_i(\xi) = 0 \ (i \in T_\sigma), \rho_i(\xi) \geq 0 \ (i \in T_r), \omega_i(\xi)\rho_i(\xi) \leq 0 \ (i \in T_r) \tag{8}$$

and

$$\sum_{i \in T} \alpha_i^L \nabla \vartheta_i^L(\psi) + \sum_{i \in T} \alpha_i^U \nabla \vartheta_i^U(\psi) + \sum_{i \in T_\tau} \lambda_i^\tau \nabla \tau_i(\psi)$$

$$+ \sum_{i \in T_\sigma} \lambda_i^\sigma \nabla \sigma_i(\psi) - \sum_{i \in T_r} \lambda_i^\rho \nabla \rho_i(\psi) + \sum_{i \in T_r} \lambda_i^\omega \nabla \omega_i(\psi) = 0 \tag{9}$$

with

$$\sum_{i \in T} \left(\alpha_i^L + \alpha_i^U\right) = 1, \lambda_{T_+(\xi)}^\rho \geq 0, \lambda_{T_{0+}(\xi)}^\omega \leq 0, \lambda_{T_{+-}(\xi) \cup T_{0-}(\xi)}^\omega \geq 0. \tag{10}$$

Therefore we conclude from (8), based on the convexity of τ_i $(i \in T_\tau^+(\xi)), \sigma_i$ $(i \in T_\sigma^+(\xi)), -\sigma_i$ $(i \in T_\sigma^-(\xi)), \rho_i$ $(i \in \hat{T}_0^-(\xi)), -\rho_i$ $(i \in \hat{T}_+^+(\xi) \cup \hat{T}_0^+(\xi)), \omega_i$ $(i \in T_{+0}^+(\xi) \cup T_{+-}^+(\xi) \cup T_{00}^+(\xi) \cup T_{0-}^+(\xi)), -\omega_i$ $(i \in T_{+0}^-(\xi) \cup T_{0+}^-(\xi) \cup T_{00}^-(\xi))$ at ψ and by the definitions of index sets that

$$\tau_i(\psi) + \langle \nabla \tau_i(\psi), \xi - \psi \rangle \leq \tau_i(\xi) \leq 0, \lambda_i^\tau > 0, \text{ for all } i \in T_\tau^+(\xi),$$

$$\sigma_i(\psi) + \langle \nabla \sigma_i(\psi), \xi - \psi \rangle \leq \sigma_i(\xi) = 0, \lambda_i^\sigma > 0, \text{ for all } i \in T_\sigma^+(\xi),$$

$$-\sigma_i(\psi) + \langle -\nabla \sigma_i(\psi), \xi - \psi \rangle \leq -\sigma_i(\xi) = 0, \lambda_i^\sigma < 0, \text{ for all } i \in T_\sigma^-(\xi),$$

$$\rho_i(\psi) + \langle \nabla \rho_i(\psi), \xi - \psi \rangle \leq \rho_i(\xi) = 0, \lambda_i^\rho < 0, \text{ for all } i \in \hat{T}_0^-(\xi),$$

$$-\rho_i(\psi) + \langle -\nabla \rho_i(\psi), \xi - \psi \rangle \leq -\rho_i(\xi) < 0, \lambda_i^\rho > 0, \text{ for all } i \in \hat{T}_+^+(\xi),$$

$$-\rho_i(\psi) + \langle -\nabla \rho_i(\psi), \xi - \psi \rangle \leq -\rho_i(\xi) < 0, \lambda_i^\rho > 0, \text{ for all } i \in \hat{T}_0^+(\xi),$$

$$\omega_i(\psi) + \langle \nabla \omega_i(\psi), \xi - \psi \rangle \leq \omega_i(\xi) = 0, \lambda_i^\omega > 0, \text{ for all } i \in T_{+0}^+(\xi) \cup T_{00}^+(\xi),$$

$$\omega_i(\psi) + \langle \nabla \omega_i(\psi), \xi - \psi \rangle \leq \omega_i(\xi) < 0, \lambda_i^\omega > 0, \text{ for all } i \in T_{+-}^+(\xi) \cup T_{0-}^+(\xi),$$

$$-\omega_i(\psi) + \langle -\nabla \omega_i(\psi), \xi - \psi \rangle \leq -\omega_i(\xi) = 0, \lambda_i^\omega > 0, \text{ for all } i \in T_{+0}^-(\xi) \cup T_{00}^-(\xi),$$

$$-\omega_i(\psi) + \langle -\nabla \omega_i(\psi), \xi - \psi \rangle \leq -\omega_i(\xi) < 0, \lambda_i^\omega < 0, \text{ for all } i \in T_{0+}^-(\xi).$$

The above inequalities imply that

$$\sum_{i \in T_\tau} \lambda_i^\tau \tau_i(\psi) + \sum_{i \in T_\sigma} \lambda_i^\sigma \sigma_i(\psi) - \sum_{i \in T_r} \lambda_i^\rho \rho_i(\psi) + \sum_{i \in T_r} \lambda_i^\omega \omega_i(\psi)$$

$$+ \left\langle \sum_{i \in T_\tau} \lambda_i^\tau \nabla \tau_i(\psi) + \sum_{i \in T_\sigma} \lambda_i^\sigma \nabla \sigma_i(\psi) - \sum_{i \in T_r} \lambda_i^\rho \nabla \rho_i(\psi) \right.$$

$$\left. + \sum_{i \in T_r} \lambda_i^\omega \nabla \omega_i(\psi), \xi - \psi \right\rangle \le 0. \tag{11}$$

By using (9) and (11), we obtain

$$\left\langle \sum_{i \in T} \alpha_i^L \nabla \vartheta_i^L(\psi) + \sum_{i \in T} \alpha_i^U \nabla \vartheta_i^U(\psi), \xi - \psi \right\rangle$$

$$= - \left\langle \sum_{i \in T_\tau} \lambda_i^\tau \nabla \tau_i(\psi) + \sum_{i \in T_\sigma} \lambda_i^\sigma \nabla \sigma_i(\psi) - \sum_{i \in T_r} \lambda_i^\rho \nabla \rho_i(\psi) + \sum_{i \in T_r} \lambda_i^\omega \nabla \omega_i(\psi), \xi - \psi \right\rangle$$

$$\ge \sum_{i \in T_\tau} \lambda_i^\tau \tau_i(\psi) + \sum_{i \in T_\sigma} \lambda_i^\sigma \sigma_i(\psi) - \sum_{i \in T_r} \lambda_i^\rho \rho_i(\psi) + \sum_{i \in T_r} \lambda_i^\omega \omega_i(\psi). \tag{12}$$

(i) Suppose to the contrary that

$$\vartheta(\xi) \prec_{LU} \mathcal{L}(\psi, \alpha^L, \alpha^U, \lambda^\tau, \lambda^\sigma, \lambda^\omega, \lambda^\rho). \tag{13}$$

Then, we deduce from (13) and $\alpha^L \in R_+^m, \alpha^U \in R_+^m$ that

$$\left\langle \alpha^L, \vartheta^L(\xi) - \mathcal{L}(\psi, \alpha^L, \alpha^U, \lambda^\tau, \lambda^\sigma, \lambda^\omega, \lambda^\rho) \right\rangle < 0,$$
$$\left\langle \alpha^U, \vartheta^U(\xi) - \mathcal{L}(\psi, \alpha^L, \alpha^U, \lambda^\tau, \lambda^\sigma, \lambda^\omega, \lambda^\rho) \right\rangle < 0,$$

which is equivalent to

$$\sum_{i=1}^m \alpha_i^L \left(\vartheta_i^L(\xi) - \vartheta_i^L(\psi) \right) - \sum_{i=1}^m \alpha_i^L \left(\sum_{i \in T_\tau} \lambda_i^\tau \tau_i(\psi) + \sum_{i \in T_\sigma} \lambda_i^\sigma \sigma_i(\psi) \right.$$

$$\left. - \sum_{i \in T_r} \lambda_i^\rho \rho_i(\psi) + \sum_{i \in T_r} \lambda_i^\omega \omega_i(\psi) \right) < 0,$$

$$\sum_{i=1}^m \alpha_i^U \left(\vartheta_i^U(\xi) - \vartheta_i^U(\psi) \right) - \sum_{i=1}^m \alpha_i^U \left(\sum_{i \in T_\tau} \lambda_i^\tau \tau_i(\psi) + \sum_{i \in T_\sigma} \lambda_i^\sigma \sigma_i(\psi) \right.$$

$$\left. - \sum_{i \in T_r} \lambda_i^\rho \rho_i(\psi) + \sum_{i \in T_r} \lambda_i^\omega \omega_i(\psi) \right) < 0.$$

On adding, we have

$$\sum_{i=1}^m \alpha_i^L \left(\vartheta_i^L(\xi) - \vartheta_i^L(\psi) \right) + \sum_{i=1}^m \alpha_i^U \left(\vartheta_i^U(\xi) - \vartheta_i^U(\psi) \right) - \sum_{i=1}^m \left(\alpha_i^L + \alpha_i^U \right)$$

$$\left(\sum_{i\in T_\tau}\lambda_i^\tau\tau_i(\psi)+\sum_{i\in T_\sigma}\lambda_i^\sigma\sigma_i(\psi)-\sum_{i\in T_r}\lambda_i^\rho\rho_i(\psi)+\sum_{i\in T_r}\lambda_i^\omega\omega_i(\psi)\right)<0.$$

It follows from $\sum\limits_{i=1}^{m}\left(\alpha_i^L+\alpha_i^U\right)=1$ that

$$\sum_{i=1}^{m}\alpha_i^L\left(\vartheta_i^L(\xi)-\vartheta_i^L(\psi)\right)+\sum_{i=1}^{m}\alpha_i^U\left(\vartheta_i^U(\xi)-\vartheta_i^U(\psi)\right)$$

$$<\left(\sum_{i\in T_\tau}\lambda_i^\tau\tau_i(\psi)+\sum_{i\in T_\sigma}\lambda_i^\sigma\sigma_i(\psi)-\sum_{i\in T_r}\lambda_i^\rho\rho_i(\psi)+\sum_{i\in T_r}\lambda_i^\omega\omega_i(\psi)\right).\qquad(14)$$

From the convexity of $\vartheta_i^L,\vartheta_i^U(i\in T)$ at ψ, we get

$$\begin{aligned}\left\langle\nabla\vartheta_i^L(\psi),\xi-\psi\right\rangle&\le\vartheta_i^L(\xi)-\vartheta_i^L(\psi),&&\text{for all } i\in T,\\\left\langle\nabla\vartheta_i^U(\psi),\xi-\psi\right\rangle&\le\vartheta_i^U(\xi)-\vartheta_i^U(\psi),&&\text{for all } i\in T,\end{aligned}$$

which leads that

$$\left\langle\sum_{i=1}^{m}\alpha_i^L\nabla\vartheta_i^L(\psi),\xi-\psi\right\rangle\le\sum_{i=1}^{m}\alpha_i^L\left(\vartheta_i^L(\xi)-\vartheta_i^L(\psi)\right),$$

$$\left\langle\sum_{i=1}^{m}\alpha_i^U\nabla\vartheta_i^U(\psi),\xi-\psi\right\rangle\le\sum_{i=1}^{m}\alpha_i^U\left(\vartheta_i^U(\xi)-\vartheta_i^U(\psi)\right).\qquad(15)$$

We deduce from the above inequalities and (14) that

$$\left\langle\sum_{i=1}^{m}\alpha_i^L\nabla\vartheta_i^L(\psi)+\sum_{i=1}^{m}\alpha_i^U\nabla\vartheta_i^U(\psi),\xi-\psi\right\rangle$$

$$<\left(\sum_{i\in T_\tau}\lambda_i^\tau\tau_i(\psi)+\sum_{i\in T_\sigma}\lambda_i^\sigma\sigma_i(\psi)-\sum_{i\in T_r}\lambda_i^\rho\rho_i(\psi)+\sum_{i\in T_r}\lambda_i^\omega\omega_i(\psi)\right),$$

which contradicts with (12).

(ii) Reasoning by contraposition, suppose that

$$\vartheta(\xi)\preceq_{LU}\mathcal{L}(\psi,\alpha^L,\alpha^U,\lambda^\tau,\lambda^\sigma,\lambda^\omega,\lambda^\rho).\qquad(16)$$

We deduce from (16) and $\alpha^L\in R_+^m,\alpha^U\in R_+^m$ that

$$\begin{cases}\left\langle\alpha^L,\vartheta^L(\xi)-\mathcal{L}(\psi,\alpha^L,\alpha^U,\lambda^\tau,\lambda^\sigma,\lambda^\omega,\lambda^\rho)\right\rangle<0,\\\left\langle\alpha^U,\vartheta^U(\xi)-\mathcal{L}(\psi,\alpha^L,\alpha^U,\lambda^\tau,\lambda^\sigma,\lambda^\omega,\lambda^\rho)\right\rangle\le0,\end{cases}$$

or

$$
\begin{cases}
\left\langle \alpha^L, \vartheta^L(\xi) - \mathcal{L}(\psi, \alpha^L, \alpha^U, \lambda^\tau, \lambda^\sigma, \lambda^\omega, \lambda^\rho) \right\rangle \le 0, \\
\left\langle \alpha^U, \vartheta^U(\xi) - \mathcal{L}(\psi, \alpha^L, \alpha^U, \lambda^\tau, \lambda^\sigma, \lambda^\omega, \lambda^\rho) \right\rangle < 0,
\end{cases}
$$

or

$$
\begin{cases}
\left\langle \alpha^L, \vartheta^L(\xi) - \mathcal{L}(\psi, \alpha^L, \alpha^U, \lambda^\tau, \lambda^\sigma, \lambda^\omega, \lambda^\rho) \right\rangle < 0, \\
\left\langle \alpha^U, \vartheta^U(\xi) - \mathcal{L}(\psi, \alpha^L, \alpha^U, \lambda^\tau, \lambda^\sigma, \lambda^\omega, \lambda^\rho) \right\rangle < 0,
\end{cases}
$$

which is equivalent to

$$
\sum_{i=1}^m \alpha_i^L \left(\vartheta_i^L(\xi) - \vartheta_i^L(\psi) \right) + \sum_{i=1}^m \alpha_i^U \left(\vartheta_i^U(\xi) - \vartheta_i^U(\psi) \right)
$$

$$
\le \sum_{i=1}^m \left(\alpha_i^L + \alpha_i^U \right) \left(\sum_{i \in T_\tau} \lambda_i^\tau \tau_i(\psi) + \sum_{i \in T_\sigma} \lambda_i^\sigma \sigma_i(\psi) - \sum_{i \in T_r} \lambda_i^\rho \rho_i(\psi) + \sum_{i \in T_r} \lambda_i^\omega \omega_i(\psi) \right).
$$

It follows from $\sum_{i=1}^m \left(\alpha_i^L + \alpha_i^U \right) = 1$ that

$$
\sum_{i=1}^m \alpha_i^L \left(\vartheta_i^L(\xi) - \vartheta_i^L(\psi) \right) + \sum_{i=1}^m \alpha_i^U \left(\vartheta_i^U(\xi) - \vartheta_i^U(\psi) \right)
$$

$$
\le \left(\sum_{i \in T_\tau} \lambda_i^\tau \tau_i(\psi) + \sum_{i \in T_\sigma} \lambda_i^\sigma \sigma_i(\psi) - \sum_{i \in T_r} \lambda_i^\rho \rho_i(\psi) + \sum_{i \in T_r} \lambda_i^\omega \omega_i(\psi) \right). \tag{17}
$$

From the strict convexity of $\vartheta_i^L, \vartheta_i^U (i \in T)$ at ψ, we get

$$
\begin{aligned}
\left\langle \nabla \vartheta_i^L(\psi), \xi - \psi \right\rangle &< \vartheta_i^L(\xi) - \vartheta_i^L(\psi), \quad \text{for all } i \in T \\
\left\langle \nabla \vartheta_i^U(\psi), \xi - \psi \right\rangle &< \vartheta_i^U(\xi) - \vartheta_i^U(\psi), \quad \text{for all } i \in T,
\end{aligned}
$$

which leads that

$$
\left\langle \sum_{i=1}^m \alpha_i^L \nabla \vartheta_i^L(\psi), \xi - \psi \right\rangle < \sum_{i=1}^m \alpha_i^L \left(\vartheta_i^L(\xi) - \vartheta_i^L(\psi) \right),
$$

$$
\left\langle \sum_{i=1}^m \alpha_i^U \nabla \vartheta_i^U(\psi), \xi - \psi \right\rangle < \sum_{i=1}^m \alpha_i^U \left(\vartheta_i^U(\xi) - \vartheta_i^U(\psi) \right). \tag{18}
$$

It follows from (17) and (18) that

$$
\left\langle \sum_{i=1}^m \alpha_i^L \nabla \vartheta_i^L(\psi) + \sum_{i=1}^m \alpha_i^U \nabla \vartheta_i^U(\psi), \xi - \psi \right\rangle
$$

$$
< \left(\sum_{i \in T_\tau} \lambda_i^\tau \tau_i(\psi) + \sum_{i \in T_\sigma} \lambda_i^\sigma \sigma_i(\psi) - \sum_{i \in T_r} \lambda_i^\rho \rho_i(\psi) + \sum_{i \in T_r} \lambda_i^\omega \omega_i(\psi) \right),
$$

contradicting to (12). □

Example 2. Consider the following multiobjective interval-valued programming problem with vanishing constraints (MIVVC-2):

$$MIVVC - 2 \qquad \mathbb{R}_+ - \min \ \vartheta(\xi) = (\vartheta_1(\xi), \vartheta_2(\xi))$$
$$= \left([2\xi + 4, e^{2\xi}], \ [-2\xi + 1, -\xi - 1]\right)$$
$$\text{subject to}$$
$$\rho_1(\xi) = \xi \geq 0,$$
$$\omega_1(\xi)\rho_1(\xi) = -e^{1-\xi}\xi \leq 0,$$

where $\vartheta_1^L(\xi) = 2\xi + 4, \vartheta_2^L(\xi) = -2\xi + 1, \vartheta_1^U(\xi) = e^{2\xi}, \vartheta_2^U(\xi) = -\xi - 1$, which is in the form of MIVVC with $m = n = 1$, $p = q = 0$ and $r = 1$. The feasible set of MIVVC-2 is $\mathbb{F}_{\mathbb{VC}2} = \{\xi \in R \mid \rho_1(\xi) \geq 0, \ \omega_1(\xi)\rho_1(\xi) \leq 0\}$. For any $\xi_0 \in \mathbb{F}_{\mathbb{VC}2}$, the corresponding Wolfe type dual problem to MIVVC-2 is given by

$$(WD_w(\xi_0) - 1)R_+^m - \max \mathcal{L}(u, \alpha^L, \alpha^U, \lambda^\omega, \lambda^\rho)$$
$$= \left([2u + 4, e^{2u}] + (-\lambda_1^\rho(u) + \lambda_1^\omega(-e^{1-u}))(1),\right.$$
$$\left. [-2u + 1, -u - 1] + (-\lambda_1^\rho(u) + \lambda_1^\omega(-e^{1-u}))(1)\right)$$
$$\text{subject to}$$
$$\alpha_1^L(2) + \alpha_1^U(2e^{2u}) + \alpha_2^L(-2) + \alpha_2^U(-1) - \lambda_1^\rho(1)$$
$$+ \lambda_1^\omega(-1) = 0,$$
$$\alpha_1^L + \alpha_1^U = 1, \alpha_2^L + \alpha_2^U = 1, \lambda_1^\rho \begin{cases} \geq 0, \ if 1 \in T_+(\xi_0), \\ \in R, \ if 1 \in T_0(\xi_0), \end{cases}$$
$$\lambda_1^\omega \begin{cases} \leq 0, \ if 1 \in T_{0+}(\xi_0), \\ \geq 0, \ if 1 \in T_{+-}(\xi_0) \cup T_{0-}(\xi_0), \\ \in R, \ if 1 \in T_{+0}(\xi_0) \cup T_{00}(\xi_0), \end{cases}$$

$$where \ (u, \alpha_1^L, \alpha_1^U, \alpha_2^L, \alpha_2^U, \lambda^\omega, \lambda^\rho) \in R \times R_+ \times R_+ \times R_+ \times R_+ \times R \times R.$$

Therefore, we get the following feasible set of problem $(WD_w(\xi_0) - 1)$:

$$(\mathbb{F}_{\mathbb{VC}w}(\xi_0) - 1) := \left\{(u, \alpha_1^L, \alpha_1^U, \alpha_2^L, \alpha_2^U, \lambda^\omega, \lambda^\rho) \in R^n \times R_+^m \times R_+^m \right.$$
$$\times R_+^m \times R_+^m \times R^r \times R^r | \alpha_1^L + \alpha_1^U = 1, \alpha_2^L + \alpha_2^U = 1,$$
$$\lambda_1^\rho \in R, \lambda_1^\omega \in R, \alpha_1^L \nabla \vartheta_1^L(u) + \alpha_1^U \nabla \vartheta_1^U(u) + \alpha_2^L \nabla \vartheta_2^L(u)$$
$$\left. + \alpha_2^U \nabla \vartheta_2^U(u) - \lambda_1^\rho \nabla \rho_1(u) + \lambda_1^\omega \nabla \omega_1(u) = 0\right\}.$$

By elementary calculations, we get $\nabla\vartheta_1^L(\xi_0) = \{2\}$, $\nabla\vartheta_1^U(\xi_0) = \{2\}$, $\nabla\vartheta_2^L(\xi_0) = \{-2\}$, $\nabla\vartheta_2^U(\xi_0) = \{-1\}$, $\nabla\rho_1(\xi_0) = \{1\}$, $\nabla\omega_1(\xi_0) = \{e\}$, $T_+ = T_{0+} = T_{0-} = \phi$, $T_{00} = \{1\}$.

Clearly, $(u, \alpha_1^L, \alpha_1^U, \alpha_2^L, \alpha_2^U, \lambda^\omega, \lambda^\rho) = (0, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, \frac{1}{2}, 0, \frac{1}{2})$ is a feasible solution to $(WD_w(\xi_0) - 1)$. We also note that $\xi_0 = 0$ is a feasible solution to MIVVC-2. On the other hand, it is easily verified that the hypothesis (i) and (ii) of Theorem 4 are satisfied at $u = 0$.

Theorem 5 (Strong duality). Let $\xi_0 \in \mathbb{F}_{\mathbb{VC}}$ be a locally weakly efficient solution of MIVVC. If MIVVC-VACQ holds at ξ_0 and the set Δ_1 is closed, then there exists $(\bar{\alpha}^L, \bar{\alpha}^U, \bar{\lambda}^\tau, \bar{\lambda}^\sigma, \bar{\lambda}^\omega, \bar{\lambda}^\rho) \in R_+^m \times R_+^m \times R^p \times R^q \times R^r \times R^r$ with $\bar{\lambda}_{T_+(\xi_0)}^\rho = 0, \bar{\lambda}_{T_{00}(\xi_0)\cup T_{0-}(\xi_0)}^\rho \geq 0, \bar{\lambda}_{T_{+-}(\xi_0)\cup T_{0+}(\xi_0)\cup T_{0-}(\xi_0)}^\omega = 0$ and $\bar{\lambda}_{T_{+0}(\xi_0)\cup T_{00}(\xi_0)}^\omega \geq 0$ such that $(\xi_0, \bar{\alpha}^L, \bar{\alpha}^U, \bar{\lambda}^\tau, \bar{\lambda}^\sigma, \bar{\lambda}^\omega, \bar{\lambda}^\rho) \in \mathbb{F}_{\mathbb{VC}w}(\xi_0)$ and $\vartheta(\xi_0) = \mathcal{L}(\xi_0, \bar{\alpha}^L, \bar{\alpha}^U, \bar{\lambda}^\tau, \bar{\lambda}^\sigma, \bar{\lambda}^\omega, \bar{\lambda}^\rho)$. Furthermore, assume that $\tau_i (i \in T_\tau^+(\xi_0)), \sigma_i \ (i \in T_\sigma^+(\xi_0)), -\sigma_i(i \in T_\sigma^-(\xi_0)), \rho_i(i \in \hat{T}_0^-(\xi_0)), -\rho_i(i \in \hat{T}_+^+(\xi_0) \cup \hat{T}_0^+(\xi_0)), \omega_i(i \in T_{+0}^+(\xi_0) \cup T_{+-}^+(\xi_0) \cup T_{00}^+(\xi_0) \cup T_{0-}^+(\xi_0)), -\omega_i(i \in T_{+0}^-(\xi_0) \cup T_{0+}^-(\xi_0) \cup T_{00}^-(\xi_0))$ are convex functions at ξ_0.

(i) If $\vartheta_i^L, \vartheta_i^U (i \in T)$ are convex functions at ξ_0, then $(\xi_0, \bar{\alpha}^L, \bar{\alpha}^U, \bar{\lambda}^\tau, \bar{\lambda}^\sigma, \bar{\lambda}^\omega, \bar{\lambda}^\rho)$ is a weakly LU-efficient solution of $WD_w(\xi_0)$.

(ii) If $\vartheta_i^L, \vartheta_i^U (i \in T)$ are strictly convex functions at ξ_0, then $(\xi_0, \bar{\alpha}^L, \bar{\alpha}^U, \bar{\lambda}^\tau, \bar{\lambda}^\sigma, \bar{\lambda}^\omega, \bar{\lambda}^\rho)$ is an LU-efficient solution of $WD_w(\xi_0)$.

Proof. In view of Theorem 1, there exists $(\bar{\alpha}^L, \bar{\alpha}^U, \bar{\lambda}^\tau, \bar{\lambda}^\sigma, \bar{\lambda}^\omega, \bar{\lambda}^\rho) \in R_+^m \times R_+^m \times R^p \times R^q \times R^r \times R^r$ with $\bar{\lambda}_{T_+(\xi_0)}^\rho = 0, \bar{\lambda}_{T_{00}(\xi_0)\cup T_{0-}(\xi_0)}^\rho \geq 0, \bar{\lambda}_{T_{+-}(\xi_0)\cup T_{0+}(\xi_0)\cup T_{0-}(\xi_0)}^\omega = 0$ and $\bar{\lambda}_{T_{+0}(\xi_0)\cup T_{00}(\xi_0)}^\omega \geq 0$ such that

$$\sum_{i\in T} \bar{\alpha}_i^L \nabla\vartheta_i^L(\xi_0) + \sum_{i\in T} \bar{\alpha}_i^U \nabla\vartheta_i^U(\xi_0) + \sum_{i\in T_\tau} \bar{\lambda}_i^\tau \nabla\tau_i(\xi_0) + \sum_{i\in T_\sigma} \bar{\lambda}_i^\sigma \nabla\sigma_i(\xi_0)$$

$$- \sum_{i\in T_r} \bar{\lambda}_i^\rho \nabla\rho_i(\xi_0) + \sum_{i\in T_r} \bar{\lambda}_i^\omega \nabla\omega_i(\xi_0) = 0.$$

Since $\bar{\lambda}^\tau \in R^p$, one has $\bar{\lambda}_i^\tau \tau_i(\xi_0) = 0$ for all $i \in T_\tau$, and thus, $\sum_{i\in T_\tau} \bar{\lambda}_i^\tau \tau_i(\xi_0) = 0$. The fact $\xi_0 \in \mathbb{F}_{\mathbb{VC}}$ guarantees that $\sum_{i\in T_\sigma} \bar{\lambda}_i^\sigma \sigma_i(\xi_0) = 0$. Moreover, we observe by $\bar{\lambda}_{T_+(\xi_0)}^\rho = 0$ and $\rho_i(\xi_0) = 0$ for all $i \in T_0(\xi_0)$ that $\sum_{i\in T_r} \bar{\lambda}_i^\rho \rho_i(\xi_0) = 0$. Analogously, as $\bar{\lambda}_{T_{+-}(\xi_0)\cup T_{0+}(\xi_0)\cup T_{0-}(\xi_0)}^\omega = 0$ and $\omega_i(\xi_0) = 0$ for all $i \in T_{+0}(\xi_0) \cup T_{00}(\xi_0)$, we know that $\sum_{i\in T_r} \bar{\lambda}_i^\omega \omega_i(\xi_0) = 0$. Thus, $(\xi_0, \bar{\alpha}^L, \bar{\alpha}^U, \bar{\lambda}^\tau, \bar{\lambda}^\sigma, \bar{\lambda}^\omega, \bar{\lambda}^\rho) \in \mathbb{F}_{\mathbb{VC}w}(\xi_0)$ and $\sum_{i\in T_\tau} \bar{\lambda}_i^\tau \tau_i(\xi_0) + \sum_{i\in T_\sigma} \bar{\lambda}_i^\sigma \sigma_i(\xi_0) - \sum_{i\in T_r} \bar{\lambda}_i^\rho \rho_i(\xi_0) + \sum_{i\in T_r} \bar{\lambda}_i^\omega \omega_i(\xi_0) = 0$ which is nothing else but the following equality $\vartheta(\xi_0) = \mathcal{L}(\xi_0, \bar{\alpha}^L, \bar{\alpha}^U, \bar{\lambda}^\tau, \bar{\lambda}^\sigma, \bar{\lambda}^\omega, \bar{\lambda}^\rho)$.

(i). Now, arguing by contradiction, let us suppose that $(\xi_0, \bar{\alpha}^L, \bar{\alpha}^U, \bar{\lambda}^\tau, \bar{\lambda}^\sigma, \bar{\lambda}^\omega, \bar{\lambda}^\rho)$ is not a weakly LU-efficient solution of $WD_w(\xi_0)$. By the definition, there

exists $(u, \alpha^L, \alpha^U, \lambda^\tau, \lambda^\sigma, \lambda^\omega, \lambda^\rho) \in \mathbb{F}_{\mathbb{VC}w}(\xi_0)$ such that

$$\mathcal{L}(\xi_0, \bar{\alpha}^L, \bar{\alpha}^U, \bar{\lambda}^\tau, \bar{\lambda}^\sigma, \bar{\lambda}^\omega, \bar{\lambda}^\rho) \prec_{LU} \mathcal{L}(u, \alpha^L, \alpha^U, \lambda^\tau, \lambda^\sigma, \lambda^\omega, \lambda^\rho).$$

This shows that $\vartheta(\xi_0) \prec_{LU} \mathcal{L}(u, \alpha^L, \alpha^U, \lambda^\tau, \lambda^\sigma, \lambda^\omega, \lambda^\rho)$, which contradicts with Theorem 4(i).

(ii). Reasoning to the contrary, let us assume that $(\xi_0, \bar{\alpha}^L, \bar{\alpha}^U, \bar{\lambda}^\tau, \bar{\lambda}^\sigma, \bar{\lambda}^\omega, \bar{\lambda}^\rho)$ is not an LU-efficient solution of $WD_w(\xi_0)$. Then it guarantees the existence of $(u, \alpha^L, \alpha^U, \lambda^\tau, \lambda^\sigma, \lambda^\omega, \lambda^\rho) \in \mathbb{F}_{\mathbb{VC}w}(\xi_0)$ such that

$$\mathcal{L}(\xi_0, \bar{\alpha}^L, \bar{\alpha}^U, \bar{\lambda}^\tau, \bar{\lambda}^\sigma, \bar{\lambda}^\omega, \bar{\lambda}^\rho) \preceq_{LU} \mathcal{L}(u, \alpha^L, \alpha^U, \lambda^\tau, \lambda^\sigma, \lambda^\omega, \lambda^\rho).$$

Consequently, $\vartheta(\xi_0) \preceq_{LU} \mathcal{L}(u, \alpha^L, \alpha^U, \lambda^\tau, \lambda^\sigma, \lambda^\omega, \lambda^\rho)$ which contradicts with Theorem 4(ii). $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \square$

Theorem 6 (Strict converse duality). Let $\tilde{\xi} \in \mathbb{F}_{\mathbb{VC}}$ be a locally weakly efficient solution of MIVVC such that MIVVC-VACQ holds at $\tilde{\xi}$ and the strong duality between the MIVVC and the $(WD_W(\tilde{\xi}))$ as in Theorem 5 holds. Also, let $(\tilde{\psi}, \tilde{\alpha}^L, \tilde{\alpha}^U, \tilde{\lambda}^\tau, \tilde{\lambda}^\sigma, \tilde{\lambda}^\omega, \tilde{\lambda}^\rho) \in \mathbb{F}_{\mathbb{VC}w}$ be an LU-efficient solution of $(WD_W(\tilde{\xi}))$. Moreover, Assume that $\vartheta_i^L, \vartheta_i^U (i \in T)$ are strictly convex functions and that $\tau_i(i \in T_\tau^+(\tilde{\xi})), \sigma_i(i \in T_\sigma^+(\tilde{\xi})), -\sigma_i(i \in T_\sigma^-(\tilde{\xi})), \rho_i(i \in \hat{T}_0^-(\tilde{\xi})), -\rho_i(i \in \hat{T}_+^+(\tilde{\xi}) \cup \hat{T}_0^+(\tilde{\xi})), \omega_i(i \in T_{+0}^+(\tilde{\xi}) \cup T_{+-}^+(\tilde{\xi}) \cup T_{00}^+(\tilde{\xi}) \cup T_{0-}^+(\tilde{\xi})), -\omega_i(i \in T_{+0}^-(\tilde{\xi}) \cup T_{0+}^-(\tilde{\xi}) \cup T_{00}^-(\tilde{\xi}))$ are convex functions at $\tilde{\psi}$, respectively. Then, $\tilde{\xi} = \tilde{\psi}$.

Proof. Suppose on the contrary, $\tilde{\xi} \neq \tilde{\psi}$. Then, by Theorem 5, there exist $\tilde{\xi} \in \mathbb{F}_{\mathbb{VC}}$ and $(\tilde{\psi}, \tilde{\alpha}^L, \tilde{\alpha}^U, \tilde{\lambda}^\tau, \tilde{\lambda}^\sigma, \tilde{\lambda}^\omega, \tilde{\lambda}^\rho) \in \mathbb{F}_{\mathbb{VC}w}$, and hence

$$\vartheta(\tilde{\xi}) = \mathcal{L}(\tilde{\psi}, \tilde{\alpha}^L, \tilde{\alpha}^U, \tilde{\lambda}^\tau, \tilde{\lambda}^\sigma, \tilde{\lambda}^\omega, \tilde{\lambda}^\rho). \tag{19}$$

The strict convexity of $\vartheta_i^L, \vartheta_i^U (i \in T)$ at $\tilde{\psi}$ gives that

$$\left\langle \sum_{i=1}^m \alpha_i^L \nabla \vartheta_i^L(\tilde{\psi}) + \sum_{i=1}^m \alpha_i^U \nabla \vartheta_i^U(\tilde{\psi}), \tilde{\xi} - \tilde{\psi} \right\rangle$$

$$< \left(\sum_{i \in T_\tau} \lambda_i^\tau \tau_i(\tilde{\psi}) + \sum_{i \in T_\sigma} \lambda_i^\sigma \sigma_i(\tilde{\psi}) - \sum_{i \in T_r} \lambda_i^\rho \rho_i(\tilde{\psi}) + \sum_{i \in T_r} \lambda_i^\omega \omega_i(\tilde{\psi}) \right). \tag{20}$$

The convexity of $\tau_i(i \in T_\tau^+(\tilde{\xi})), \sigma_i(i \in T_\sigma^+(\tilde{\xi})), -\sigma_i(i \in T_\sigma^-(\tilde{\xi})),$ $\rho_i(i \in \hat{T}_0^-(\tilde{\xi})), -\rho_i(i \in \hat{T}_+^+(\tilde{\xi}) \cup \hat{T}_0^+(\tilde{\xi})), \omega_i(i \in T_{+0}^+(\tilde{\xi}) \cup T_{+-}^+(\tilde{\xi}) \cup T_{00}^+(\tilde{\xi}) \cup T_{0-}^+(\tilde{\xi})), -\omega_i(i \in T_{+0}^-(\tilde{\xi}) \cup T_{0+}^-(\tilde{\xi}) \cup T_{00}^-(\tilde{\xi}))$ at $\tilde{\psi}$ and by the definitions of index sets imply that

$$\tau_i(\tilde{\psi}) + \left\langle \nabla \tau_i(\tilde{\psi}), \tilde{\xi} - \tilde{\psi} \right\rangle \leq \tau_i(\tilde{\xi}) = 0, \lambda_i^\tau > 0, \text{ for all } i \in T_\tau^+(\tilde{\xi}),$$

$$\sigma_i(\tilde{\psi}) + \left\langle \nabla \sigma_i(\tilde{\psi}), \tilde{\xi} - \tilde{\psi} \right\rangle \leq \sigma_i(\tilde{\xi}) = 0, \lambda_i^\sigma > 0, \text{ for all } i \in T_\sigma^+(\tilde{\xi}),$$

$$-\sigma_i(\tilde{\psi}) + \left\langle -\nabla\sigma_i(\tilde{\psi}), \tilde{\xi} - \tilde{\psi} \right\rangle \le -\sigma_i(\tilde{\xi}) = 0, \lambda_i^\sigma < 0, \text{ for all } i \in T_\sigma^-(\tilde{\xi}),$$

$$\rho_i(\tilde{\psi}) + \left\langle \nabla\rho_i(\tilde{\psi}), \tilde{\xi} - \tilde{\psi} \right\rangle \le \rho_i(\tilde{\xi}) = 0, \lambda_i^\rho < 0, \text{ for all } i \in \hat{T}_0^-(\tilde{\xi}),$$

$$-\rho_i(\tilde{\psi}) + \left\langle -\nabla\rho_i(\tilde{\psi}), \tilde{\xi} - \tilde{\psi} \right\rangle \le -\rho_i(\tilde{\xi}) < 0, \lambda_i^\rho > 0, \text{ for all } i \in \hat{T}_+^+(\tilde{\xi}),$$

$$-\rho_i(\tilde{\psi}) + \left\langle -\nabla\rho_i(\tilde{\psi}), \tilde{\xi} - \tilde{\psi} \right\rangle \le -\rho_i(\tilde{\xi}) < 0, \lambda_i^\rho > 0, \text{ for all } i \in \hat{T}_0^+(\tilde{\xi}),$$

$$\omega_i(\tilde{\psi}) + \left\langle \nabla\omega_i(\tilde{\psi}), \tilde{\xi} - \tilde{\psi} \right\rangle \le \omega_i(\tilde{\xi}) = 0, \lambda_i^\omega > 0, \text{ for all } i \in T_{+0}^+(\tilde{\xi}) \cup T_{00}^+(\tilde{\xi}),$$

$$\omega_i(\tilde{\psi}) + \left\langle \nabla\omega_i(\tilde{\psi}), \tilde{\xi} - \tilde{\psi} \right\rangle \le \omega_i(\tilde{\xi}) < 0, \lambda_i^\omega > 0, \text{ for all } i \in T_{+-}^+(\tilde{\xi}) \cup T_{0-}^+(\tilde{\xi}),$$

which implies that

$$\sum_{i\in T_\tau}\lambda_i^\tau\tau_i(\tilde{\xi}) + \sum_{i\in T_\sigma}\lambda_i^\sigma\sigma_i(\tilde{\xi}) - \sum_{i\in T_r}\lambda_i^\rho\rho_i(\tilde{\xi}) + \sum_{i\in T_r}\lambda_i^\omega\omega_i(\tilde{\xi}) + \left\langle \sum_{i\in T_\tau}\lambda_i^\tau\nabla\tau_i(\tilde{\psi})\right.$$

$$\left. + \sum_{i\in T_\sigma}\lambda_i^\sigma\nabla\sigma_i(\tilde{\psi}) - \sum_{i\in T_r}\lambda_i^\rho\nabla\rho_i(\tilde{\psi}) + \sum_{i\in T_r}\lambda_i^\omega\nabla\omega_i(\tilde{\psi}), \tilde{\xi} - \tilde{\psi} \right\rangle \le 0. \qquad (21)$$

On adding the inequalities (20) and (21) and by using the duality constraint (9) of $(WD_w(\tilde{\xi}))$, we have

$$\mathcal{L}(\tilde{\psi}, \tilde{\alpha}^L, \tilde{\alpha}^U, \tilde{\lambda}^\tau, \tilde{\lambda}^\sigma, \tilde{\lambda}^\omega, \tilde{\lambda}^\rho) \prec_{LU} \vartheta(\tilde{\xi}),$$

which contradicts with (19). $\qquad\qquad\square$

5 The Mond–Weir type duality

The Wolfe dual of the primal problem, which we discussed in the last section, says that all functions must be convex. Wolfe duality does not work for functions, where the objective function is only pseudoconvex and the constraints are only quasiconvex in the primal problem MIVVC (see, Mond [24]). So, in this section, we propose a Mond–Weir type dual to the primal problem MIVVC to weaken the convexity assumptions.

Consider $\xi_0 \in \mathbb{F}_{\mathbb{VC}}, (u, \alpha^L, \alpha^U, \lambda^\tau, \lambda^\sigma, \lambda^\omega, \lambda^\rho) \in R^n \times R_+^m \times R_+^m \times R^p \times R^q \times R^r \times R^r$ with $\sum_{i\in T}(\alpha_i^L + \alpha_i^U) = 1, \lambda_{T_+(\xi_0)}^\rho \ge 0, \lambda_{T_{0+}(\xi_0)}^\omega \le 0$, and $\lambda_{T_{+-}(\xi_0)\cup T_{0-}(\xi_0)}^\omega \ge 0$. We consider the Mond–Weir type dual problem as follows:

$(MWD_M(\xi_0))$ $R_+^m - \max \vartheta(u)$

subject to

$$\sum_{i \in T} \alpha_i^L \nabla \vartheta_i^L(u) + \sum_{i \in T} \alpha_i^U \nabla \vartheta_i^U(u) + \sum_{i \in T_\tau} \lambda_i^\tau \nabla \tau_i(u) +$$

$$\sum_{i \in T_\sigma} \lambda_i^\sigma \nabla \sigma_i(u) - \sum_{i \in T_r} \lambda_i^\rho \nabla \rho_i(u) + \sum_{i \in T_r} \lambda_i^\omega \nabla \omega_i(u) = 0,$$

$$\lambda_i^\tau \tau_i(u) \geq 0 \ (i \in T_\tau), \lambda_i^\sigma \sigma_i(u) = 0 \ (i \in T_\sigma), -\lambda_i^\rho \rho_i(u) \geq 0$$

$$(i \in T_r), \lambda_i^\omega \omega_i(u) \geq 0 \ (i \in T_r), \sum_{i \in T} \left(\alpha_i^L + \alpha_i^U \right) = 1,$$

$$\lambda_{T_+(\xi_0)}^\rho \geq 0, \lambda_{T_{0+}(\xi_0)}^\omega \leq 0 \text{ and } \lambda_{T_{+-}(\xi_0) \cup T_{0-}(\xi_0)}^\omega \geq 0, (u, \alpha^L, \alpha^U,$$

$$\lambda^\tau, \lambda^\sigma, \lambda^\omega, \lambda^\rho) \in R^n \times R_+^m \times R_+^m \times R^p \times R^q \times R^r \times R^r.$$

The feasible set of $(MWD_M(\xi_0))$ is defined by

$$\mathbb{F}_{\mathbb{VC}M}(\xi_0) := \Big\{ (u, \alpha^L, \alpha^U, \lambda^\tau, \lambda^\sigma, \lambda^\omega, \lambda^\rho) \in R^n \times R_+^m \times R_+^m \times R^p \times R^q$$

$$\times R^r \times R^r \mid \lambda_i^\tau \tau_i(u) \geq 0 \ (i \in T_\tau), \lambda_i^\sigma \sigma_i(u) = 0 \ (i \in T_\sigma),$$

$$- \lambda_i^\rho \rho_i(u) \geq 0 \ (i \in T_r), \lambda_i^\omega \omega_i(u) \geq 0 \ (i \in T_r),$$

$$\sum_{i \in T} \left(\alpha_i^L + \alpha_i^U \right) = 1, \lambda_{T_+(\xi_0)}^\rho \geq 0, \lambda_{T_{0+}(\xi_0)}^\omega \leq 0, \text{ and}$$

$$\lambda_{T_{+-}(\xi_0) \cup T_{0-}(\xi_0)}^\omega \geq 0, \sum_{i \in T} \alpha_i^L \nabla \vartheta_i^L(u) + \sum_{i \in T} \alpha_i^U \nabla \vartheta_i^U(u)$$

$$+ \sum_{i \in T_\tau} \lambda_i^\tau \nabla \tau_i(u) + \sum_{i \in T_\sigma} \lambda_i^\sigma \nabla \sigma_i(u) - \sum_{i \in T_r} \lambda_i^\rho \nabla \rho_i(u)$$

$$+ \sum_{i \in T_r} \lambda_i^\omega \nabla \omega_i(u) = 0 \Big\}.$$

Furthermore, let us denote by Γ_M the projection of $\mathbb{F}_{\mathbb{VC}M}$ on $\mathbb{R}^n$; that is,

$$\Gamma_M(\xi_0) := \left\{ u \in \mathbb{R}^n \mid (u, \alpha^L, \alpha^U, \lambda^\tau, \lambda^\sigma, \lambda^\omega, \lambda^\rho) \in \Gamma_M(\xi_0) \right\}.$$

The other Mond–Weir type duality problem of MIVVC, which is not dependent on ξ_0, is

$$(MWD_M): \quad R_+^m - \max \ \vartheta(\psi)$$
$$\text{subject to}$$
$$(\psi, \alpha^L, \alpha^U, \lambda^\tau, \lambda^\sigma, \lambda^\omega, \lambda^\rho) \in \Gamma_M := \bigcap_{\xi_0 \in \Gamma} \Gamma_M(\xi_0).$$

Definition 9. Let $\xi_0 \in \mathbb{F}_{\mathbb{VC}}$. Then $(\bar{u}, \bar{\alpha}^L, \bar{\alpha}^U, \bar{\lambda}^\tau, \bar{\lambda}^\sigma, \bar{\lambda}^\omega, \bar{\lambda}^\rho) \in \mathbb{F}_{\mathbb{VC}M}(\xi_0)$ is a locally LU-efficient solution of $(MWD_M(\xi_0))$ (locally weakly LU-efficient solution of $(MWD_M(\xi_0))$) if there exists $U \in \Theta(\bar{u})$ such that there is no $(u, \alpha^L, \alpha^U, \lambda^\tau, \lambda^\sigma, \lambda^\omega, \lambda^\rho) \in \mathbb{F}_{\mathbb{VC}M}(\xi_0) \cap U$ satisfying

$$\vartheta(\bar{u}) \preceq_{LU} \vartheta(u)$$
$$(\vartheta(\bar{u}) \prec_{LU} \vartheta(u))$$

Theorem 7 (Weak duality). Let $\xi \in \mathbb{F}_{\mathbb{VC}}$ and $(\psi, \alpha^L, \alpha^U, \lambda^\tau, \lambda^\sigma, \lambda^\omega, \lambda^\rho) \in \mathbb{F}_{\mathbb{VC}M}$. Suppose that τ_i $(i \in T_\tau^+(\xi)), \sigma_i$ $(i \in T_\sigma^+(\xi)), -\sigma_i$ $(i \in T_\sigma^-(\xi)), \rho_i$ $(i \in \hat{T}_0^-(\xi)), -\rho_i$ $(i \in \hat{T}_+^+(\xi) \cup \hat{T}_0^+(\xi)), \omega_i$ $(i \in T_{+0}^+(\xi) \cup T_{+-}^+(\xi) \cup T_{00}^+(\xi) \cup T_{0-}^+(\xi)), -\omega_i$ $(i \in T_{+0}^-(\xi) \cup T_{0+}^-(\xi) \cup T_{00}^-(\xi))$ are quasiconvex functions at ψ on $\mathbb{F}_{\mathbb{VC}M} \cup \Gamma_M$. If $\vartheta_i^L, \vartheta_i^U (i \in T)$ are strictly pseudoconvex functions at ψ on $\mathbb{F}_{\mathbb{VC}M} \cup \Gamma_M$, then $\vartheta(\xi) \not\preceq_{LU} \vartheta(\psi)$.

Proof. For $\xi \in \mathbb{F}_{\mathbb{VC}M}$ and

$$(\psi, \alpha^L, \alpha^U, \lambda^\tau, \lambda^\sigma, \lambda^\omega, \lambda^\rho) \in \mathbb{F}_{\mathbb{VC}M} = \bigcap_{\xi_0 \in \mathbb{F}_{\mathbb{VC}M}} \mathbb{F}_{\mathbb{VC}M}(\xi_0),$$

we have

$$\tau_i(\xi) \leq 0 \, (i \in T_\tau), \quad \sigma_i(\xi) = 0 \, (i \in T_\sigma), \quad \rho_i(\xi) \geq 0 \, (i \in T_r), \quad \omega_i(\xi)\rho_i(\xi) \leq 0 \, (i \in T_r), \tag{22}$$

$$\sum_{i \in T} \alpha_i^L \nabla \vartheta_i^L(\psi) + \sum_{i \in T} \alpha_i^U \nabla \vartheta_i^U(\psi) + \sum_{i \in T_\tau} \lambda_i^\tau \nabla \tau_i(\psi) + \sum_{i \in T_\sigma} \lambda_i^\sigma \nabla \sigma_i(\psi)$$
$$- \sum_{i \in T_r} \lambda_i^\rho \nabla \rho_i(\psi) + \sum_{i \in T_r} \lambda_i^\omega \nabla \omega_i(\psi) = 0, \tag{23}$$

and

$$\lambda_i^\tau \tau_i(\psi) \geq 0 \, (i \in T_\tau), \quad \lambda_i^\sigma \sigma_i(\psi) = 0 \, (i \in T_\sigma),$$
$$-\lambda_i^\rho \rho_i(\psi) \geq 0 \, (i \in T_r), \quad \lambda_i^\omega \omega_i(\psi) \geq 0 \, (i \in T_r), \tag{24}$$

with

$$\sum_{i \in T} \left(\alpha_i^L + \alpha_i^U \right) = 1, \quad \lambda_{T_+(\xi)}^\rho \geq 0, \quad \lambda_{T_{0+}(\xi)}^\omega \leq 0, \quad \lambda_{T_{+-}(\xi) \cup T_{0-}(\xi)}^\omega \geq 0. \tag{25}$$

It follows from the above inequalities that

$$\tau_i(\xi) \leq 0 \leq \tau_i(\psi) \leq 0, \quad \text{for all } i \in T_\tau^+(\xi),$$

$$\sigma_i(\xi) = \sigma_i(\psi) = 0, \quad \text{for all } i \in T_\sigma^+(\xi) \cup T_\sigma^-(\xi),$$

$$\rho_i(\xi) = 0 \leq \rho_i(\psi), \quad \text{for all } i \in \hat{T}_0^-(\xi),$$

$$-\rho_i(\xi) \leq 0 \leq -\rho_i(\psi), \quad \text{for all } i \in \hat{T}_+^+(\xi) \cup \hat{T}_0^+(\xi),$$

$$\omega_i(\xi) \leq 0 \leq \omega_i(\psi), \quad \text{for all } i \in T_{+0}^+(\xi) \cup T_{+-}^+(\xi) \cup T_{00}^+(\xi) \cup T_{0-}^+(\xi),$$

$$-\omega_i(\xi) \leq 0 \leq -\omega_i(\psi) = 0, \quad \text{for all } i \in T_{+0}^-(\xi) \cup T_{0+}^-(\xi) \cup T_{00}^-(\xi).$$

Thus, we deduce from the quasiconvexity of τ_i $(i \in T_\tau^+(\xi)), \sigma_i$ $(i \in T_\sigma^+(\xi)), -\sigma_i$ $(i \in T_\sigma^-(\xi)), \rho_i$ $(i \in \hat{T}_0^-(\xi)), -\rho_i$ $(i \in \hat{T}_+^+(\xi) \cup \hat{T}_0^+(\xi)),$

ω_i ($i \in T_{+0}^+(\xi) \cup T_{+-}^+(\xi) \cup T_{00}^+(\xi) \cup T_{0-}^+(\xi)$), $-\omega_i$ ($i \in T_{+0}^-(\xi) \cup T_{0+}^-(\xi) \cup T_{00}^-(\xi)$) at ψ and the definitions of index sets that

$$\langle \nabla \tau_i(\psi), \xi - \psi \rangle \leq 0, \lambda_i^\tau > 0, \quad \text{for all } i \in T_\tau^+(\xi),$$

$$\langle \nabla \sigma_i(\psi), \xi - \psi \rangle \leq 0, \lambda_i^\sigma > 0, \quad \text{for all } i \in T_\sigma^+(\xi),$$

$$\langle -\nabla \sigma_i(\psi), \xi - \psi \rangle \leq 0, \lambda_i^\sigma < 0, \quad \text{for all } i \in T_\sigma^-(\xi),$$

$$\langle \nabla \rho_i(\psi), \xi - \psi \rangle \leq 0, \lambda_i^\rho < 0, \quad \text{for all } i \in \hat{T}_0^-(\xi),$$

$$\langle -\nabla \rho_i(\psi), \xi - \psi \rangle \leq 0, \lambda_i^\rho > 0, \quad \text{for all } i \in \hat{T}_+^+(\xi) \cup \hat{T}_0^+(\xi),$$

$$\langle \nabla \omega_i(\psi), \xi - \psi \rangle \leq 0, \lambda_i^\omega > 0, \quad \text{for all } i \in T_{+0}^+(\xi) \cup T_{+-}^+(\xi) \cup T_{00}^+(\xi) \cup T_{0-}^+(\xi),$$

$$\langle -\nabla \omega_i(\psi), \xi - \psi \rangle \leq 0, \lambda_i^\omega < 0, \quad \text{for all } i \in T_{+0}^-(\xi) \cup T_{0+}^-(\xi) \cup T_{00}^-(\xi),$$

Employing this together with (23) gives us the inequality

$$\left\langle \sum_{i \in T} \alpha_i^L \nabla \vartheta_i^L(\psi) + \sum_{i \in T} \alpha_i^U \nabla \vartheta_i^U(\psi), \xi - \psi \right\rangle$$

$$= -\left\langle \sum_{i \in T_\tau} \lambda_i^\tau \nabla \tau_i(\psi) + \sum_{i \in T_\sigma} \lambda_i^\sigma \nabla \sigma_i(\psi) - \sum_{i \in T_r} \lambda_i^\rho \nabla \rho_i(\psi) + \sum_{i \in T_r} \lambda_i^\omega \nabla \omega_i(\psi), \xi - \psi \right\rangle$$

$$\geq 0. \tag{26}$$

Assume by contradiction that

$$\vartheta(\xi) \preceq_{LU} \vartheta(\psi).$$

This is equivalent to

$$\begin{cases} \vartheta^L(\xi) < \vartheta^L(\psi) \\ \vartheta^U(\xi) \leq \vartheta^U(\psi) \end{cases}, \quad \text{or} \quad \begin{cases} \vartheta^L(\xi) \leq \vartheta^L(\psi) \\ \vartheta^U(\xi) < \vartheta^U(\psi) \end{cases}, \quad \text{or} \quad \begin{cases} \vartheta^L(\xi) < \vartheta^L(\psi) \\ \vartheta^U(\xi) < \vartheta^U(\psi) \end{cases}.$$

Since $\vartheta_i^L, \vartheta_i^U$ ($i \in T$) are strictly pseudoconvex functions at ψ, we have

$$\langle \nabla \vartheta_i^L(\psi), \xi - \psi \rangle < 0, \text{ for all } i \in T,$$
$$\langle \nabla \vartheta_i^U(\psi), \xi - \psi \rangle < 0, \text{ for all } i \in T.$$

Taking into account $\alpha^L \in R_+^m, \alpha^U \in R_+^m$ and from $\sum_{i=1}^m (\alpha_i^L + \alpha_i^U) = 1$, we have

$$\left\langle \sum_{i=1}^m \alpha_i^L \nabla \vartheta_i^L(\psi) + \sum_{i=1}^m \alpha_i^U \nabla \vartheta_i^U(\psi), \xi - \psi \right\rangle < 0,$$

contradicting to (26). $\qquad \qquad \square$

Example 3. Let $m = n = 1$, let $p = q = 0$ and let $r = 1$. Let us investigate the following $(MIVVC - 3)$:

$MIVVC-3 \qquad \mathbb{R}_+ - \min \ \vartheta(\xi) = (\vartheta_1(\xi), \vartheta_2(\xi))$

$$= \left([4\xi^2 - \xi, 4\xi^2 + \xi + 1], \ [\xi^2 - 2\xi, \xi^4 + 2\xi]\right)$$

subject to

$$\rho_1(\xi) = \xi \geq 0,$$
$$\omega_1(\xi)\rho_1(\xi) = (-1 - \xi)\xi \leq 0.$$

Then, $\mathbb{F}_{\mathbb{VC}3} = \{\xi \in R \mid \rho_1(\xi) \geq 0, \ \omega_1(\xi)\rho_1(\xi) \leq 0\}$. For any $\xi_0 \in \mathbb{F}_{\mathbb{VC}3}$, the corresponding Mond–Weir dual problem to MIVVC-3 is given by

$(MWD_M - 1) \ R_+^m - \max \ \vartheta(u)$

$$= \left([4u^2 - u, 4u^2 + u + 1], [u^2 - 2u, u^4 + 2u]\right)$$

subject to

$$\alpha_1^L(8u - 1) + \alpha_1^U(8u + 1) + \alpha_2^L(2u - 2) + \alpha_2^U(4u^3 + 2)$$
$$- \lambda_1^\rho(1) + \lambda_1^\omega(-1) = 0, -\lambda_1^\rho(u) \geq 0, \lambda_1^\omega(-1 - u) \geq 0,$$
$$\alpha_1^L + \alpha_1^U = 1, \alpha_2^L + \alpha_2^U = 1,$$

$$\lambda_1^\rho \begin{cases} \geq 0, \ if 1 \in T_+(\xi_0), \\ \in R, \ if 1 \in T_0(\xi_0), \end{cases} \quad \lambda_1^\omega \begin{cases} \leq 0, \ if 1 \in T_{0+}(\xi_0), \\ \geq 0, \ if 1 \in T_{+-}(\xi_0) \cup T_{0-}(\xi_0), \\ \in R, \ if 1 \in T_{+0}(\xi_0) \cup T_{00}(\xi_0), \end{cases}$$

$where \ (u, \alpha_1^L, \alpha_1^U, \alpha_2^L, \alpha_2^U, \lambda_1^\omega, \lambda_1^\rho) \in R \times R_+ \times R_+ \times R_+ \times R_+ \times R \times R.$

Therefore, we get the following feasible set of problem $(MWD_M(\xi_0) - 1)$:

$$(\mathbb{F}_{\mathbb{VC}M}(\xi_0) - 1) := \left\{ (u, \alpha_1^L, \alpha_1^U, \alpha_2^L, \alpha_2^U, \lambda^\omega, \lambda^\rho) \in R^n \times R_+^m \times R_+^m \times R_+^m \right.$$
$$\times R_+^m \times R^r \times R^r \mid -\lambda_1^\rho(u) \geq 0, \lambda_1^\omega(-1 - u) \geq 0,$$
$$\alpha_1^L + \alpha_1^U = 1, \alpha_2^L + \alpha_2^U = 1, \lambda_1^\rho \in R, \lambda_1^\omega \in R,$$
$$\alpha_1^L \nabla \vartheta_1^L(u) + \alpha_1^U \nabla \vartheta_1^U(u) + \alpha_2^L \nabla \vartheta_2^L(u)$$
$$\left. + \alpha_2^U \nabla \vartheta_2^U(u) - \lambda_1^\rho \nabla \rho_1(u) + \lambda_1^\omega \nabla \omega_1(u) = 0 \right\}.$$

By taking $\xi_0 = 0 \in \mathbb{F}_{\mathbb{VC}3}$, we evidence from Examples 1 and 2 that all suppositions of Theorem 1 are fulfilled. Now, by choosing $\alpha_1^L = \alpha_1^U = \frac{1}{2}, \alpha_2^L = \alpha_2^U = \frac{1}{2}, \lambda_1^\omega = 0, \lambda_1^\rho = 0$,
we have

$$-\lambda_1^\rho(\xi_0) \geq 0, \qquad \lambda_1^\omega(-1 - \xi_0) \geq 0,$$
$$\frac{1}{2}(-1) + \frac{1}{2}(1) + \frac{1}{2}(-2) + \frac{1}{2}(2) - \lambda_1^\rho(1) + \lambda_1^\omega(-1) = 0.$$

Finally, by the strict pseudoconvexity of $\vartheta_i^L, \vartheta_i^U (i \in T)$ at ψ on $\mathbb{F}_{\mathbb{VC}M} \cup \Gamma_M$ and by simple calculations, we get $\vartheta(\xi) \npreceq_{LU} \vartheta(\psi)$.

Theorem 8 (Strong duality)**.** Let $\xi_0 \in \mathbb{F}_{\mathbb{VC}}$ be a locally weakly efficient solution of MIVVC. If MIVVC-VACQ holds at ξ_0 and the set Δ_1 is closed, then there exists $(\bar{\alpha}^L, \bar{\alpha}^U, \bar{\lambda}^\tau, \bar{\lambda}^\sigma, \bar{\lambda}^\omega, \bar{\lambda}^\rho) \in R_+^m \times R_+^m \times R^p \times R^q \times R^r \times R^r$ with $\sum_{i=1}^m \left(\bar{\alpha}_i^L + \bar{\alpha}_i^U \right) = 1, \bar{\lambda}_{T_+(\xi_0)}^\rho = 0, \bar{\lambda}_{T_{00}(\xi_0) \cup T_{0-}(\xi_0)}^\rho \geq 0, \bar{\lambda}_{T_{+-}(\xi_0) \cup T_{0+}(\xi_0) \cup T_{0-}(\xi_0)}^\omega = 0$ and $\bar{\lambda}_{T_{+0}(\xi_0) \cup T_{00}(\xi_0)}^\omega \geq 0$ such that $(\xi_0, \bar{\alpha}^L, \bar{\alpha}^U, \bar{\lambda}^\tau, \bar{\lambda}^\sigma, \bar{\lambda}^\omega, \bar{\lambda}^\rho) \in \mathbb{F}_{\mathbb{VC}M}(\xi_0)$. Furthermore, assume that $\tau_i(i \in T_\tau^+(\xi_0)), \sigma_i(i \in T_\sigma^+(\xi_0)), -\sigma_i(i \in T_\sigma^-(\xi_0)), \rho_i(i \in \hat{T}_0^-(\xi_0)), -\rho_i(i \in \hat{T}_+^+(\xi_0) \cup \hat{T}_0^+(\xi_0)), \omega_i(i \in T_{+0}^+(\xi_0) \cup T_{+-}^+(\xi_0) \cup T_{00}^+(\xi_0) \cup T_{0-}^+(\xi_0)), -\omega_i(i \in T_{+0}^-(\xi_0) \cup T_{0+}^-(\xi_0) \cup T_{00}^-(\xi_0))$ are quasiconvex functions at ξ_0. If $\vartheta_i^L, \vartheta_i^U(i \in T)$ are strictly pseudoconvex functions at ξ_0, then $(\xi_0, \bar{\alpha}^L, \bar{\alpha}^U, \bar{\lambda}^\tau, \bar{\lambda}^\sigma, \bar{\lambda}^\omega, \bar{\lambda}^\rho)$ is an LU-efficient solution of $MWD_M(\xi_0)$.

Proof. By Theorem (1), there exists $(\bar{\alpha}^L, \bar{\alpha}^U, \bar{\lambda}^\tau, \bar{\lambda}^\sigma, \bar{\lambda}^\omega, \bar{\lambda}^\rho) \in R_+^m \times R_+^m \times R^p \times R^q \times R^r \times R^r$ with $\sum_{i=1}^m \left(\alpha_i^L + \alpha_i^U \right) = 1, \bar{\lambda}_{T_+(\xi_0)}^\rho = 0, \bar{\lambda}_{T_{00}(\xi_0) \cup T_{0-}(\xi_0)}^\rho \geq 0, \bar{\lambda}_{T_{+-}(\xi_0) \cup T_{0+}(\xi_0) \cup T_{0-}(\xi_0)}^\omega = 0$ and $\bar{\lambda}_{T_{+0}(\xi_0) \cup T_{00}(\xi_0)}^\omega \geq 0$ such that

$$\sum_{i \in T} \alpha_i^L \nabla \vartheta_i^L(\psi) + \sum_{i \in T} \alpha_i^U \nabla \vartheta_i^U(\psi) + \sum_{i \in T_\tau} \lambda_i^\tau \nabla \tau_i(\psi) + \sum_{i \in T_\sigma} \lambda_i^\sigma \nabla \sigma_i(\psi)$$

$$- \sum_{i \in T_r} \lambda_i^\rho \nabla \rho_i(\psi) + \sum_{i \in T_r} \lambda_i^\omega \nabla \omega_i(\psi) = 0.$$

Since $\bar{\lambda}^\tau \in R^p$, one has $\bar{\lambda}_i^\tau \tau_i(\xi_0) = 0$ for all $i \in T_\tau$. The fact that $\xi_0 \in \mathbb{F}_{\mathbb{VC}}$ guarantees that $\bar{\lambda}_i^\sigma \sigma_i(\xi_0) = 0$. Furthermore, we deduce from $\bar{\lambda}_{T_+(\xi_0)}^\rho = 0$ and $\rho_i(\xi_0) = 0$ for all $i \in T_0(\xi_0)$ that $-\bar{\lambda}_i^\rho \rho_i(\xi_0) = 0$ for all $i \in T_r$. In addition, we get from $\bar{\lambda}_{T_{+-}(\xi_0) \cup T_{0+}(\xi_0) \cup T_{0-}(\xi_0)}^\omega = 0$ and $\omega_i(\xi_0) = 0$ for all $i \in T_{+0}(\xi_0) \cup T_{00}(\xi_0)$, that $\bar{\lambda}_i^\omega \omega_i(\xi_0) = 0$ for all $i \in T_r$. Thus, $(\xi_0, \bar{\alpha}^L, \bar{\alpha}^U, \bar{\lambda}^\tau, \bar{\lambda}^\sigma, \bar{\lambda}^\omega, \bar{\lambda}^\rho) \in \mathbb{F}_{\mathbb{VC}M}(\xi_0)$.

(i). Now, arguing by contradiction, let us suppose that $(\xi_0, \bar{\alpha}^L, \bar{\alpha}^U, \bar{\lambda}^\tau, \bar{\lambda}^\sigma, \bar{\lambda}^\omega, \bar{\lambda}^\rho)$ is not a weakly LU-efficient solution of $MWD_M(\xi_0)$. By the definition, there exists $(u, \alpha^L, \alpha^U, \lambda^\tau, \lambda^\sigma, \lambda^\omega, \lambda^\rho) \in \mathbb{F}_{\mathbb{VC}M}(\xi_0)$ such that

$$\vartheta(\xi_0) \prec_{LU} \vartheta(u),$$

which contradicts with Theorem 4(i).

(ii). Reasoning to the contrary, Let us assume that $(\xi_0, \bar{\alpha}^L, \bar{\alpha}^U, \bar{\lambda}^\tau, \bar{\lambda}^\sigma, \bar{\lambda}^\omega, \bar{\lambda}^\rho)$ is not an LU-efficient solution of $MWD_M(\xi_0)$. Then, there exists $(u, \alpha^L, \alpha^U, \lambda^\tau, \lambda^\sigma, \lambda^\omega, \lambda^\rho) \in \mathbb{F}_{\mathbb{VC}M}(\xi_0)$ such that

$$\vartheta(\xi_0) \preceq_{LU} \vartheta(u)$$

which contradicts with Theorem 4(ii), and thus, completes the proof. $\square$

Theorem 9 (Strict converse duality)**.** Let $\tilde{\xi} \in \mathbb{F}_{\mathbb{VC}}$ be a locally weakly efficient solution of MIVVC such that MIVVC-VACQ holds at $\tilde{\xi}$ and the strong

duality between the MIVVC and the $(MWD_M)(\tilde\xi)$ as in Theorem 8 holds. Also, let $(\tilde\psi, \tilde\alpha^L, \tilde\alpha^U, \tilde\lambda^\tau, \tilde\lambda^\sigma, \tilde\lambda^\omega, \tilde\lambda^\rho) \in \mathbb{F}_{\mathbb{VC}M}$ be an LU-efficient solution of $(MWD_M)(\tilde\xi)$. Moreover, Suppose that $\vartheta_i^L, \vartheta_i^U (i \in T)$ are strictly pseudoconvex functions and that $\tau_i(i \in T_\tau^+(\tilde\xi)), \sigma_i(i \in T_\sigma^+(\tilde\xi)), -\sigma_i(i \in T_\sigma^-(\tilde\xi)), \rho_i(i \in \hat{T}_0^-(\tilde\xi)), -\rho_i(i \in \hat{T}_+^+(\tilde\xi) \cup \hat{T}_0^+(\tilde\xi)), \omega_i(i \in T_{+0}^+(\tilde\xi) \cup T_{+-}^+(\tilde\xi) \cup T_{00}^+(\tilde\xi) \cup T_{0-}^+(\tilde\xi)), -\omega_i(i \in T_{+0}^-(\tilde\xi) \cup T_{0+}^-(\tilde\xi) \cup T_{00}^-(\tilde\xi))$ are quasiconvex functions at $\tilde\psi$ on $\mathbb{F}_{\mathbb{VC}M} \cup \Gamma_M$, respectively.

Proof. Suppose, contrary to the result, that $\tilde\xi \neq \tilde\psi$. Then, by the strong duality theorem, there exist $(\alpha^L, \alpha^U, \lambda^\tau, \lambda^\sigma, \lambda^\omega, \lambda^\rho) \in R_+^m \times R_+^m \times R^p \times R^q \times R^r \times R^r$ such that $(\tilde\psi, \alpha^L, \alpha^U, \lambda^\tau, \lambda^\sigma, \lambda^\omega, \lambda^\rho)$ is an LU-efficient solution of $MWD_M(\tilde\xi)$, and hence

$$\vartheta(\tilde\xi) = \vartheta(\tilde\psi). \tag{27}$$

By the strict pseudoconvexity of $\vartheta_i^L, \vartheta_i^U (i \in T)$ at $\tilde\psi$ on $\mathbb{F}_{\mathbb{VC}M} \cup \Gamma_M$, we have

$$\left\langle \sum_{i=1}^m \alpha_i^L \nabla\vartheta_i^L(\tilde\psi) + \sum_{i=1}^m \alpha_i^U \nabla\vartheta_i^U(\tilde\psi), \tilde\xi - \tilde\psi \right\rangle < 0. \tag{28}$$

By the quasiconvexity of $\tau_i(i \in T_\tau^+(\tilde\xi)), \sigma_i(i \in T_\sigma^+(\tilde\xi)), -\sigma_i(i \in T_\sigma^-(\tilde\xi)), \rho_i(i \in \hat{T}_0^-(\tilde\xi)), -\rho_i(i \in \hat{T}_+^+(\tilde\xi) \cup \hat{T}_0^+(\tilde\xi)), \omega_i(i \in T_{+0}^+(\tilde\xi) \cup T_{+-}^+(\tilde\xi) \cup T_{00}^+(\tilde\xi) \cup T_{0-}^+(\tilde\xi)), -\omega_i(i \in T_{+0}^-(\tilde\xi) \cup T_{0+}^-(\tilde\xi) \cup T_{00}^-(\tilde\xi))$ at $\tilde\psi$ on $\mathbb{F}_{\mathbb{VC}} \cup \Gamma_{MWD}$ and by the definitions of index sets, we have

$$\left\langle \nabla\tau_i(\tilde\psi), \tilde\xi - \tilde\psi \right\rangle \leq 0, \lambda_i^\tau > 0, \quad \text{for all } i \in T_\tau^+(\tilde\xi),$$

$$\left\langle \nabla\sigma_i(\tilde\psi), \tilde\xi - \tilde\psi \right\rangle \leq 0, \lambda_i^\sigma > 0, \quad \text{for all } i \in T_\sigma^+(\tilde\xi),$$

$$\left\langle -\nabla\sigma_i(\tilde\psi), \tilde\xi - \tilde\psi \right\rangle \leq 0, \lambda_i^\sigma < 0, \quad \text{for all } i \in T_\sigma^-(\tilde\xi),$$

$$\left\langle \nabla\rho_i(\tilde\psi), \tilde\xi - \tilde\psi \right\rangle \leq 0, \lambda_i^\rho < 0, \quad \text{for all } i \in \hat{T}_0^-(\tilde\xi),$$

$$\left\langle -\nabla\rho_i(\tilde\psi), \tilde\xi - \tilde\psi \right\rangle \leq 0, \lambda_i^\rho > 0, \quad \text{for all } i \in \hat{T}_+^+(\tilde\xi) \cup \hat{T}_0^+(\tilde\xi),$$

$$\left\langle \nabla\omega_i(\tilde\psi), \tilde\xi - \tilde\psi \right\rangle \leq 0, \lambda_i^\omega > 0, \quad \text{for all } i \in T_{+0}^+(\tilde\xi) \cup T_{+-}^+(\tilde\xi) \cup T_{00}^+(\tilde\xi) \cup T_{0-}^+(\tilde\xi),$$

$$\left\langle -\nabla\omega_i(\tilde\psi), \tilde\xi - \tilde\psi \right\rangle \leq 0, \lambda_i^\omega < 0, \quad \text{for all } i \in T_{+0}^-(\tilde\xi) \cup T_{0+}^-(\tilde\xi) \cup T_{00}^-(\tilde\xi),$$

which implies that

$$\left\langle \sum_{i \in T_\tau} \lambda_i^\tau \nabla\tau_i(\tilde\psi) + \sum_{i \in T_\sigma} \lambda_i^\sigma \nabla\sigma_i(\tilde\psi) - \sum_{i \in T_r} \lambda_i^\rho \nabla\rho_i(\tilde\psi) + \sum_{i \in T_r} \lambda_i^\omega \nabla\omega_i(\tilde\psi), \tilde\xi - \tilde\psi \right\rangle \leq 0. \tag{29}$$

On adding the inequalities (28) and (29) and by using the duality constraint of $(MWD_M(\xi_0))$, we have

$$\vartheta(\tilde{\psi}) \prec_{LU} \vartheta(\tilde{\xi}).$$

which contradicts with (27). $\square$

6 Special cases

(i). If $\vartheta_1(\xi) = \vartheta_2(\xi) = \cdots = \vartheta_m(\xi)$ then the MIVVC problem reduces to the following (IVVC) problem of Ahmad et al. [2]:

$$\text{(P-1)} \qquad \min \quad \vartheta(\xi) = (\vartheta_1(\xi)) = [\vartheta_1^L(\xi), \vartheta_1^U(\xi)]$$
$$\text{subject to}$$
$$\tau_i(\xi) \leq 0, \quad \text{for all } i = 1, 2, \ldots, p,$$
$$\sigma_i(\xi) = 0, \quad \text{for all } i = 1, 2, \ldots, q,$$
$$\rho_i(\xi) \geq 0, \quad \text{for all } i = 1, 2, \ldots, r,$$
$$\omega_i(\xi)\rho_i(\xi) \leq 0, \quad \text{for all } i = 1, 2, \ldots, r.$$

(ii). If $\vartheta_1(\xi) = \vartheta_2(\xi) = \cdots = \vartheta_m(\xi)$ and $\vartheta_1^L(\xi) = \vartheta_1^U(\xi)$ then the MIVVC problem reduces to the following (MPVC) problem of Hoheisel and Kanzow [12] and the (MPVC) problem of Ahmad, Kummari, and Al-Homidan [3]:

$$\text{(P-2)} \qquad \min \quad \vartheta(\xi)$$
$$\text{subject to}$$
$$\tau_i(\xi) \leq 0, \quad \text{for all } i = 1, 2, \ldots, p,$$
$$\sigma_i(\xi) = 0, \quad \text{for all } i = 1, 2, \ldots, q,$$
$$\rho_i(\xi) \geq 0, \quad \text{for all } i = 1, 2, \ldots, r,$$
$$\omega_i(\xi)\rho_i(\xi) \leq 0, \quad \text{for all } i = 1, 2, \ldots, r.$$

(iii). If $\rho_i(\xi) = 0 = \omega_i(\xi)$, for all $i = 1, 2, \ldots, r$, then MIVVC problem reduces to the following IVP problem of Antczak and Michalak [5]:

$$\text{(P-3)} \qquad \min \quad \vartheta(\xi) = (\vartheta_1(\xi), \vartheta_2(\xi), \ldots, \vartheta_m(\xi))$$
$$\text{subject to}$$
$$\tau_i(\xi) \leq 0, \quad \text{for all } i = 1, 2, \ldots, p,$$
$$\sigma_i(\xi) = 0, \quad \text{for all } i = 1, 2, \ldots, q.$$

As a result of the above special cases, it is evident that the problem MIVVC presented in this article is more generalized.

7 Conclusion

In this paper, we have considered a multiobjective interval-valued programming problem involving vanishing constraints. Based on generalized convexity assumptions, the sufficiency of the Karush–Khun–Tucker necessary optimality conditions has been established. Furthermore, we have anticipated Wolfe and Mond–Weir dual problems for the considered multiobjective programming problem with interval-valued objective function and delved into several duality results under convexity assumptions. The results established in the paper were exemplified by an example.

Acknowledgements

References

[1] Achtziger, W. and Kanzow, C. *Mathematical programs with vanishing constraints: Optimality conditions and constraint qualifications*, Math. Program. 114(1) (2008), 69–99.

[2] Ahmad, I., Kummari, K. and Al-Homidan, S. *Sufficiency and duality for interval-valued optimization problems with vanishing constraints using weak constraint qualifications*, Int. J. Anal. Appl. 18(5) (2020), 784–798.

[3] Ahmad, I., Kummari, K. and Al-Homidan, S. *Unified duality for mathematical programming problems with vanishing constraints*, Int. J. Nonlinear Anal. Appl. 13(2) (2022), 3191–3201.

[4] Antczak, T. *Optimality conditions and Mond–Weir duality for a class of differentiable semi-infinite multiobjective programming problems with vanishing constraints*, 4OR-Q J Oper. Res. 20 (2022), 417–442.

[5] Antczak, T. and Michalak, A. *Optimality conditions and duality results for a class of differentiable vector optimization problems with the multiple interval-valued objective function*, In: International Conference on Control, Artificial Intelligence, Robotics and Optimization (ICCAIRO) (2017, May), 207–218. IEEE.

[6] Chanas, S. and Kuchta, D. *Multiobjective programming in optimization of interval objective functions - a generalized approach*, Eur. J. Oper. Res. 94 (1996), 594–598.

[7] Chen, G., Huang, X. and Yang, X. *Vector optimization, set-valued and variational analysis*, Lecture Notes in Economics and Mathematical Systems, Springer, Berlin, 541, 2005.

[8] Christian, K., Andreas, P., Hans George, B. and Sebastian, S. *A parametric active set method for quadratic programs with vanishing constraints*, Technical Report, 2012.

[9] Dorsch, D., Shikhman, V. and Stein, O. *Mathematical programs with vanishing constraints: critical point theory*, J. Glob. Optim. 52(3) (2012), 591–605.

[10] Dussault, J. P., Haddou, M. and Migot, T. *Mathematical programs with vanishing constraints: constraint qualifications, their applications, and a new regularization method*, Optimization. 68(2-3) (2019), 509–538.

[11] Hoheisel, T. *Mathematical programs with vanishing constraints*, PhD thesis, Department of Mathematics, University of Würzburg, 2009.

[12] Hoheisel, T. and Kanzow, C. *Stationary conditions for mathematical programs with vanishing constraints using weak constraint qualifications*, J. Math. Anal. Appl. 337(1) (2008), 292–310.

[13] Hoheisel, T., Kanzow, C. and Schwartz, A. *Convergence of a local regularization approach for mathematical programmes with complementarity or vanishing constraints*, Optim. Methods Softw. 27(3) (2012), 483–512.

[14] Hoheisel, T., Pablos, B., Pooladian, A., Schwartz, A. and Steverango, L. *A Study of One-Parameter Regularization Methods for Mathematical Programs with vanishing Constraints*, Optim. Methods Softw. (2020), 1–43.

[15] Hosseinzade, E. and Hassanpour, H. *The Karush-Kuhn-Tucker optimality conditions in interval-valued multiobjective programming problems*, J. Appl. Math. Inform. 29(5-6) (2011), 1157–1165.

[16] Ishibuchi, H. and Tanaka, M. *Multiobjective programming in optimization of the interval objective function*, Eur. J. Oper. Res. 48 (1990), 219–225.

[17] Izmailov, A. F. and Pogosyan, A. *Optimality conditions and Newton-type methods for mathematical programs with vanishing constraints*, Comput. Math. Math. Phys. 49(7) (2009), 1128–1140.

[18] Jayswal, A. and Singh, V. *The characterization of efficiency and saddle point criteria for multiobjective optimization problem with vanishing constraints*, Acta. Math. Sci. 39 (2019), 382–394.

[19] Jean-Claude, L. *Robot motion planning*, Kluwer Academic Publishers, Norwell, MA, 1991.

[20] Jung, M. N., Kirches, C. and Sager, S. *On perspective functions and vanishing constraints in mixed-integer nonlinear optimal control*, Facets of Combinatorial Optimization: Festschrift for Martin Grötschel. Ed. by M. Jünger and G. Reinelt, Springer, Heidelberg, 387–417, 2013.

[21] Karmakar, S. and Bhunia, A. K. *An alternative optimization technique for interval objective constrained optimization problems via multiobjective programming*, J. Egypt. Math. Soc. 22 (2014), 292–303.

[22] Luc, DT. *Theory of vector optimization*, Lecture Notes in Economics and Mathematical Systems, Springer-Verlag, Berlin, (319) 1989.

[23] Mishra, S. K., Singh, V. and Laha, V. *On duality for mathematical programs with vanishing constraints*, Ann. Oper. Res. 243 (2016), 249–272.

[24] Mond, B. *Mond–Weir duality*, Optimization, 157–165, Springer Optim. Appl., 32, Springer, New York, 2009.

[25] Tung, L. T. *Karush–Kuhn–Tucker optimality conditions and duality for multiobjective semi-infinite programming with vanishing constraints*, Ann. Oper. Res. 311(2) (2022), 1307–1334.

[26] Urli, B. and Nadeau, R. *An interactive method to multiobjective linear programming problems with interval coefficients*, INFOR Inf. Syst. Oper. Res. 30(2) (1992), 127–137.

[27] Wu, H. C. *The Karush-Kuhn-Tucker optimality conditions in multiobjective programming problems with interval-valued objective functions*, Eur. J. Oper. Res. 196(1) (2009), 49–60.

[28] Wu, H. C. *Duality theory for optimization problems with interval-valued objective functions*, J. Optim. Theory. Appl. 144 (2010), 615–628.

How to cite this article

Error estimates for approximating fixed points and best proximity points for noncyclic and cyclic contraction mappings

A. Safari-Hafshejani

Abstract

In this article, we find a priori and a posteriori error estimates of the fixed point for the Picard iteration associated with a noncyclic contraction map, which is defined on a uniformly convex Banach space with a modulus of convexity of power type. As a result, we obtain priori and posteriori error estimates of Zlatanov for approximating the best proximity points of cyclic contraction maps on this type of space.

1 Introduction

A basic result in fixed point theory is the Banach contraction principle. Fixed point theory is an important tool to solve the equation $Tx = x$ for mappings T is defined on subsets of metric or normed spaces. One of the advantages of Banach's fixed point theorem is the estimation of the error of successive iterations and the rate of convergence. There are equations $Tx = x$ for which the exact solution is not easy to find or even is not possible to find. The error estimate is very useful in these cases. An extensive study about approximations of fixed points for self-maps can be found in [2]. In 2016, Zlatanov [17] obtained error estimates for approximating the best proximity

points for cyclic contraction maps as generalization of the Banach contraction principle. More cases can be found in [10, 11, 16] and references therein.

One other kind of a generalization of the Banach contraction principle is the notation of noncyclical maps; that is, $T : A \cup B \to A \cup B$ such that $T(A) \subseteq A$ and $T(B) \subseteq B$. Also, a sufficient condition for the existence and the uniqueness of fixed points in uniformly convex Banach spaces are given in [15].

In this article, we obtain "a priori error estimates" and "a posteriori error estimates" for approximating the fixed point of noncyclic contractions. As a result, we obtain "a priori error estimates" and "a posteriori error estimates" of Zlatanov for approximating the best proximity point of cyclic contractions.

2 Preliminaries

In this section, we recall some definitions and facts, which will be used hereafter. Let A and B be nonempty subsets of a metric space (X, d). The map $T : A \cup B \to A \cup B$ is called a noncyclic map if $T(A) \subseteq A$ and $T(B) \subseteq B$. The noncyclic map $T : A \cup B \to A \cup B$ is called a noncyclic contraction map if there holds the inequality $d(Tx, Ty) \leq kd(x, y) + (1 - k)d(A, B)$ for some $k \in (0, 1)$ and all $x \in A$ and $y \in B$, where $d(A, B) := \inf\{d(x, y) : x \in A, y \in B\}$. We say that $(\xi, \eta) \in A \times B$ is an optimal pair of fixed points of the noncyclic mapping T provided that

$$T\xi = \xi, \quad T\eta = \eta \quad \text{and} \quad d(\xi, \eta) = d(A, B),$$

The definition for noncyclic contraction was introduced in [8].

The map $T : A \cup B \to A \cup B$ is called a cyclic map if $T(A) \subseteq B$ and $T(B) \subseteq A$. The cyclic map $T : A \cup B \to A \cup B$ is called a cyclic contraction map if there holds the inequality $d(Tx, Ty) \leq kd(x, y) + (1 - k)d(A, B)$ for some $k \in (0, 1)$ and all $x \in A$ and $y \in B$. A point $\xi \in A \cup B$ is called a best proximity point for T if $d(\xi, T\xi) = d(A, B)$; see [4, 6, 7] and references therein. If sets A and B have a nonempty intersection, then every best proximity point of T is a fixed point of T.

Definition 1. [9] The modulus of convexity of a Banach space X is the function $\delta_X : [0, 2] \to [0, 1]$ defined by

$$\delta_X(\epsilon) = \inf\left\{1 - \|\frac{x + y}{2}\| : \|x\| \leq 1, \|y\| \leq 1, \|x - y\| \geq \epsilon\right\}.$$

The norm is called uniformly convex if $\delta_X(\epsilon) > 0$ for all $\epsilon > 0$. The space $(X, \|.\|)$ is called a uniformly convex space.

As a result of [15, Lemma 2.2 and Theorem 2.7], we have the next theorem.

Theorem 1. [15] Let A and B be nonempty, closed, and convex subsets of a uniformly convex Banach space $(X, \| \cdot \|)$ and let $T : A \cup B \to A \cup B$

be a noncyclic contraction map. Then T has a unique optimal pair of fixed points (ξ, η) such that for every $x_0 \in A$ and $y_0 \in B$ the sequences $\{T^n x_0\}$ and $\{T^n y_0\}$ converge to ξ and η, respectively.

Definition 2. [9] A Banach space X is said to be uniformly convex if there exists a strictly increasing function $\delta : [0, 2] \to [0, 1]$ such that the following implication holds for all $x, y, p \in X$, $R > 0$ and $r \in [0, 2R]$:

$$\left.\begin{array}{l} \|x - p\| \leq R \\ \|y - p\| \leq R \\ \|x - y\| \geq r \end{array}\right\} \quad \Rightarrow \quad \|\frac{x+y}{2} - p\| \leq \left(1 - \delta(\frac{r}{R})\right)R. \tag{1}$$

If $(X, \|\cdot\|)$ is a uniformly convex Banach space, then $\delta_X(\epsilon)$ is strictly increasing function. Therefore if $(X, \|\cdot\|)$ is a uniformly convex Banach space, then there exists the inverse function δ^{-1} of the modulus of convexity. If there exist constants $C > 0$ and $q > 0$ such that the inequality $\delta_X(\epsilon) \geq C\epsilon^q$ holds for every $\epsilon \in (0, 2]$, then we say that the modulus of convexity is of power type q. It is well known that the modulus of convexity with respect to the canonical norm $\|\cdot\|_p$ in l_p or L_p is of power type, and there holds the inequalities $\delta_X(\epsilon) \geq \frac{\epsilon^p}{p2^p}$ for $p \geq 2$ and $\delta_X(\epsilon) \geq \frac{(p-1)\epsilon^2}{8}$ for $p \in (1, 2)$; see [13]. An extensive study of the geometry of Banach spaces can be found in [1, 3, 5].

3 Main results

In this section, we begin with the following lemma as a result of [15, Lemma 2.2], which will be used later.

Lemma 1. Let A and B be nonempty subsets of a metric space (X, d) and let $T : A \cup B \to A \cup B$ be a noncyclic contraction map. Then, for every $x \in A$ and $y \in B$, there holds the inequality

$$d(T^n x, T^n y) - d(A, B) \leq k^n(d(x, y) - d(A, B)). \tag{2}$$

In the following result, we obtain our main result in this section.

Theorem 2. Suppose that A and B are nonempty, closed, and convex subsets of a uniformly convex Banach space $(X, \|\cdot\|)$ such that $d := d(A, B) > 0$, and that $T : A \cup B \to A \cup B$ is a noncyclic contraction map. Let $\delta_X(\epsilon) \geq C\epsilon^q$ for some $C > 0$, $q \geq 2$ and every $\epsilon \in (0, 2]$. Then

(i) T has a unique optimal pair of fixed points $(\xi, \eta) \in A \times B$;

(ii) for every $x_0 \in A$ and $y_0 \in B$ the sequences $\{T^n x_0\}$ and $\{T^n y_0\}$ converge to ξ and η, respectively;

(iii) a priori error estimate holds

$$\|\xi - T^m x_0\| \le \frac{M_{x_0,y_0}}{1 - \sqrt[q]{k}} \sqrt[q]{\frac{M_{x_0,y_0} - d}{Cd}} (\sqrt[q]{k})^m;$$

(iv) a posteriori error estimate holds

$$\|T^n x_0 - \xi\| \le \frac{M_{x_n,y_n}}{1 - \sqrt[q]{k}} \sqrt[q]{\frac{M_{x_n,y_n} - d}{Cd}};$$

where for every $x \in A$ and $y \in B$, $M_{x,y} := \max\{\|x - y\|, \|Tx - y\|\}$.

Proof. The proof of (i) and (ii) follows from Theorem 1. (iii) For every $n \in \mathbb{N}$ let $x_n = T^n x_0$ and let $y_n = T^n y_0$. From Lemma 1, we have the inequalities

$$\|x_n - y_n\| \le k^n(\|x_0 - y_0\| - d) + d \le k^n(M_{x_0,y_0} - d) + d,$$

$$\|x_{n+1} - y_n\| \le k^n(\|Tx_0 - y_0\| - d) + d \le k^n(M_{x_0,y_0} - d) + d,$$

and

$$\|x_n - x_{n+1}\| \le 2(k^n(M_{x_0,y_0} - d) + d).$$

Now, from (1) with $x = x_n$, $y = x_{n+1}$, $z = y_n$, $r = \|x_n - x_{n+1}\|$, $R = k^n(M_{x_0,y_0} - d) + d$, and using the convexity of the set A, we get the chain of inequalities

$$\begin{aligned} d &\le \|\frac{x_n + x_{n+1}}{2} - y_n\| \\ &\le \left(1 - \delta\left(\frac{\|x_n - x_{n+1}\|}{d + k^n(M_{x_0,y_0} - d)}\right)\right)\left(d + k^n(M_{x_0,y_0} - d)\right). \end{aligned} \quad (3)$$

Using (3), we obtain the inequality

$$\delta\left(\frac{\|x_n - x_{n+1}\|}{d + k^n(M_{x_0,y_0} - d)}\right) \le \frac{k^n(M_{x_0,y_0} - d)}{d + k^n(M_{x_0,y_0} - d)}. \quad (4)$$

From the uniform convexity of X, it follows that δ is strictly increasing, and therefore there exists its inverse function δ^{-1}, which is strictly increasing. From (4), we get

$$\|x_n - x_{n+1}\| \le \left(d + k^n(M_{x_0,y_0} - d)\right)\delta^{-1}\left(\frac{k^n(M_{x_0,y_0} - d)}{d + k^n(M_{x_0,y_0} - d)}\right). \quad (5)$$

It follows from the inequality $\delta_X(t) \ge Ct^q$ that $\delta_X^{-1}(t) \le \left(\frac{t}{C}\right)^{\frac{1}{q}}$. Using (5), we obtain

$$\|x_n - x_{n+1}\| \le M_{x_0,y_0} \sqrt[q]{\frac{k^n(M_{x_0,y_0} - d)}{C(d + k^n(M_{x_0,y_0} - d))}} \tag{6}$$

$$\le M_{x_0,y_0} \sqrt[q]{\frac{M_{x_0,y_0} - d}{Cd}} (\sqrt[q]{k})^n.$$

So, from (6), we obtain

$$\|x_n - x_{n+1}\| \le M_{x_0,y_0} \sqrt[q]{\frac{M_{x_0,y_0} - d}{Cd}} (\sqrt[q]{k})^n. \tag{7}$$

From (i) and (ii), there exists a unique fixed point $\xi \in A$ such that for every $x_0 \in A$, the sequence $\{T^n x_0\}$ converges to ξ. After substitution in (7), we get the inequality

$$\sum_{n=1}^{\infty} \|x_n - x_{n+1}\| \le M_{x_0,y_0} \sqrt[q]{\frac{M_{x_0,y_0} - d}{Cd}} \frac{\sqrt[q]{k}}{1 - \sqrt[q]{k}}.$$

Consequently, the series $\sum_{n=1}^{\infty} \|x_n - x_{n+1}\|$ is absolutely convergent. Thus, for any $m \ge 1$, there holds $\xi = x_m - \sum_{n=m}^{\infty}(x_n - x_{n+1})$, and we get the inequality

$$\|\xi - x_m\| \le \sum_{n=m}^{\infty} \|x_n - x_{n+1}\| \le M_{x_0,y_0} \sqrt[q]{\frac{M_{x_0,y_0} - d}{Cd}} \frac{(\sqrt[q]{k})^m}{1 - \sqrt[q]{k}}.$$

Hence,

$$\|\xi - T^m x_0\| \le \frac{M_{x_0,y_0}}{1 - \sqrt[q]{k}} \sqrt[q]{\frac{M_{x_0,y_0} - d}{Cd}} (\sqrt[q]{k})^m.$$

(iv) In a similar way (7), we have

$$\|x_{n+i} - x_{n+i+1}\| \le M_{x_n,y_n} \sqrt[q]{\frac{M_{x_n,y_n} - d}{Cd}} (\sqrt[q]{k})^i.$$

So,

$$\|x_n - x_{n+m}\| \le \sum_{i=0}^{m-1} \|x_{n+i} - x_{n+i+1}\|$$

$$\le M_{x_n,y_n} \sqrt[q]{\frac{M_{x_n,y_n} - d}{Cd}} \sum_{i=0}^{m-1} (\sqrt[q]{k})^i.$$

Hence,

$$\|x_n - x_{n+m}\| \leq \frac{M_{x_n,y_n}}{1 - \sqrt[q]{k}} \sqrt[q]{\frac{M_{x_n,y_n} - d}{Cd}} (1 - (\sqrt[q]{k})^m). \tag{8}$$

After letting $m \to \infty$ in (8), we obtain the inequality

$$\|T^n x_0 - \xi\| \leq \frac{M_{x_n,y_n}}{1 - \sqrt[q]{k}} \sqrt[q]{\frac{M_{x_n,y_n} - d}{Cd}}.$$

□

In the sequence, we obtain the main result of [17] as a special case of Theorem 2.

Corollary 1. [17, Theorem 3.2] Suppose that A and B are nonempty, closed and convex subsets of a uniformly convex Banach space $(X, \|.\|)$ such that $d := d(A, B) > 0$, and that $T : A \cup B \to A \cup B$ is a cyclic contraction map. Let $\delta_X(\epsilon) \geq C\epsilon^q$ for some $C > 0$, $q \geq 2$, and every $\epsilon \in (0, 2]$. Then

(i) there exists a unique best proximity point ξ of T in A, $T\xi$ is a unique best proximity point of T in B and $\xi = T^2\xi$;

(ii) for every $x_0 \in A$, the sequence $\{T^{2n}x_0\}$ converges to ξ and $\{T^{2n+1}x_0\}$ converges to $T\xi$.

(iii) a priori error estimate holds

$$\|\xi - T^{2n}x_0\| \leq \frac{\|x_0 - Tx_0\|}{1 - \sqrt[q]{k^2}} \sqrt[q]{\frac{\|x_0 - Tx_0\| - d}{Cd}} (\sqrt[q]{k})^{2n};$$

(iv) a posteriori error estimate holds

$$\|T^{2n}x_0 - \xi\| \leq \frac{\|T^{2n-1}x_0 - T^{2n}x_0\|}{1 - \sqrt[q]{k^2}} \sqrt[q]{\frac{\|T^{2n-1}x_0 - T^{2n}x_0\| - d}{Cd}} \sqrt[q]{k}.$$

Proof. The proof of (i) and (ii) follows from [17, Theorem 2.1].

Because T is a cyclic contraction map, it is clear that T^2 is a noncyclic contraction map and

$$d(T^2x, T^2y) \leq k^2 d(x, y) + (1 - k^2)d(A, B).$$

(iii) As T is a cyclic contraction map, we have

$$\|T^2x_0 - Tx_0\| \leq k\|Tx_0 - x_0\| + (1 - k)d(A, B) \leq \|Tx_0 - x_0\|.$$

So,

$$\max \{\|x_0 - Tx_0\|, \|T^2x_0 - Tx_0\|\} = \|x_0 - Tx_0\|.$$

Hence,

$$M_{x_0, Tx_0} = \|x_0 - Tx_0\|.$$

Applying Theorem 2(iii) for noncyclic contraction T^2, we obtain

$$\|\xi - T^{2m}x_0\| \leq \frac{M_{x_0, Tx_0}}{1 - \sqrt[q]{k^2}} \sqrt[q]{\frac{M_{x_0, Tx_0} - d}{Cd}} (\sqrt[q]{k^2})^m$$

$$= \frac{\|x_0 - Tx_0\|}{1 - \sqrt[q]{k^2}} \sqrt[q]{\frac{\|x_0 - Tx_0\| - d}{Cd}} (\sqrt[q]{k})^{2m}.$$

(iv) Since T is a cyclic contraction map, we get

$$\|T^{2n+2}x_0 - T^{2n+1}x_0\| \leq k\|T^{2n+1}x_0 - T^{2n}x_0\| + (1 - k)d(A, B)$$

$$\leq \|T^{2n+1}x_0 - T^{2n}x_0\|,$$

for every $n \in \mathbb{N}$. So,

$$\max\left\{\|T^{2n}x_0 - T^{2n+1}x_0\|, \|T^{2n+2}x_0 - T^{2n+1}x_0\|\right\} = \|T^{2n}x_0 - T^{2n+1}x_0\|.$$

Hence, we have relations

$$M_{T^{2n}x_0, T^{2n+1}x_0} = \|T^{2n}x_0 - T^{2n+1}x_0\|, \tag{9}$$

$$M_{T^{2n}x_0, T^{2n+1}x_0} \leq \|T^{2n-1}x_0 - T^{2n}x_0\|, \tag{10}$$

$$M_{T^{2n}x_0, T^{2n+1}x_0} - d \leq k(\|T^{2n-1}x_0 - T^{2n}x_0\| - d). \tag{11}$$

Applying Theorem 2(iv) for noncyclic contraction T^2, (9), (10), and (11), we obtain

$$\|T^{2n}x_0 - \xi\| \leq \frac{M_{T^{2n}x_0, T^{2n+1}x_0}}{1 - \sqrt[q]{k^2}} \sqrt[q]{\frac{M_{T^{2n}x_0, T^{2n+1}x_0} - d}{Cd}}$$

$$\leq \frac{\|T^{2n}x_0 - T^{2n+1}x_0\|}{1 - \sqrt[q]{k^2}} \sqrt[q]{k} \sqrt[q]{\frac{\|T^{2n-1}x_0 - T^{2n}x_0\| - d}{Cd}}$$

$$\leq \frac{\|T^{2n-1}x_0 - T^{2n}x_0\|}{1 - \sqrt[q]{k^2}} \sqrt[q]{\frac{\|T^{2n-1}x_0 - T^{2n}x_0\| - d}{Cd}} \sqrt[q]{k}.$$

$$\square$$

Let A and B be nonempty, closed, and convex subsets of a uniformly convex Banach space $(X, \|\cdot\|)$ with a modulus of convexity of power type. Theorem 2 shows that if noncyclic contraction T has a fixed point $\xi \in A$ such that $\{T^n x_0\}$ converges to ξ for some $x_0 \in A$ and (2) holds for every $x \in A$

and $y \in B$, then priori and posteriori errors estimates hold in relations (iii) and (iv) of Theorem 2, respectively. Also, Zlatanov [17] showed that if the cyclic contraction T has the best proximity point $\xi \in A$ such that $\{T^{2n}x_0\}$ converges to ξ for some $x_0 \in A$ and

$$d(T^n x, T^{n+1} x) - d(A, B) \le k^n(d(x, Tx) - d(A, B)) \tag{12}$$

for every $x \in A \cup B$, then priori and posteriori errors estimates hold in relations (iii) and (iv) of Corollary 1, respectively. In fact, these results can be generalized to contractions that satisfy these conditions. For instance, consider the generalized cyclic quasi-contraction $T : A \cup B \to A \cup B$ introduced in [14]. The author proved that if A and B are nonempty, closed, and convex subsets of a uniformly convex Banach space and $T : A \cup B \to A \cup B$ is a generalized cyclic quasi-contraction, that is, for which there exists $k \in [0, 1)$ such that

$$\|Tx - Ty\| \le k \max\left\{ \|x - y\|, \|x - Tx\|, \|y - Ty\|, \frac{\|x - Ty\| + \|Tx - y\|}{2} \right\}$$
$$+ (1 - c)d(A, B),$$

for all $x \in A$ and $y \in B$; then for every $x_0 \in A$ the sequence $\{T^{2n}x_0\}$ converges to some best proximity point $\xi \in A$ and (12) holds. So priori and posteriori errors estimates for each best proximity point of a generalized cyclic quasi-contraction hold in relations (iii) and (iv) of Corollary 1, respectively. Ilchev [11] used exactly this point to get the main results for the Kannan cyclic contractive maps.

4 A numerical example

We know that the space $(\mathbb{R}^p, \|\cdot\|_p)$ is uniformly convex with modulus of convexity of power type, provided that $p > 1$. The following example illustrates Theorem 2.

Example 1. Consider the space $\mathbb{R}^2$ endowed with the norms $\|(x, y)\|_2 = \sqrt[2]{|x|^2 + |y|^2}$. Let

$$A = \{(x, y) \in \mathbb{R}^2 : y - x + 1 \le 0, y + x - 1 \ge 0\}$$

and

$$B = \{(x, y) \in \mathbb{R}^2 : y - x - 1 \ge 0, y + x + 1 \le 0\}.$$

It is easy to calculate $d(A, B) = 2$. Suppose that $\lambda \in (0, 1)$. Let us define a map $T : \mathbb{R}_2^2 \to \mathbb{R}_2^2$ by

$$T(x, y) = \begin{cases} (1 - \lambda + \lambda x, \lambda y) & \text{if } (x, y) \in A, \\ (-1 + \lambda + \lambda x, \lambda y) & \text{if } (x, y) \in B. \end{cases}$$

We will show that the map $T : A \cup B \to A \cup B$ is a noncyclic contraction with $k = \lambda$. Consider $(x, y) \in A$, and let $(x', y') := T(x, y)$. Then

$$y' - x' + 1 = \lambda y - 1 + \lambda - \lambda x + 1 = \lambda(y - x + 1) \leq 0$$

and

$$y' + x' - 1 = \lambda y + 1 - \lambda + \lambda x - 1 = \lambda(y + x - 1) \geq 0.$$

Therefore, $T(A) \subseteq A$. The inclusion $T(B) \subseteq B$ is proved in a similar fashion. It is easy to observe that $(1, 0)$ is a fixed point of T in A, that $(-1, 0)$ is a fixed point of T in B, and that $\|(1, 0) - (-1, 0)\|_2 = 2$. Let $u_1 = (x, y)$ and let $u_2 = (x', y')$. Then

$$\begin{aligned} \|T(x, y) - T(x', y')\|_2 &= \|(2(1 - \lambda) + \lambda(x - x'), \lambda(y - y'))\|_2 \\ &= \sqrt{|2(1 - \lambda) + \lambda(x - x')|^2 + \lambda|y - y'|^2} \\ &= \|2(1 - \lambda)e_1 + \lambda(u_1 - u_2)\|_2 \\ &\leq \lambda\|u_1 - u_2\|_2 + (1 - \lambda)d(A, B). \end{aligned}$$

Thus we can apply Theorem 2 to get error estimates of the successive iterations $\{x_n\}$, where $x_{n+1} = Tx_n$. We will consider a numerical example with $\lambda = \frac{1}{16}$. From [13], we get $C = \frac{1}{8}$ and $q = 2$.

Applying Theorem 2(iv), we obtain

$$\|x_n - \xi\| \leq M_n,$$

for $n \geq 0$, where

$$M_n := \frac{8}{3} M_{x_n, y_n} \sqrt{M_{x_n, y_n} - 2}.$$

In the following table, we obtain the number n of iterations, needed by a posteriori estimate less than 0.005 with initial points $x_0 = (1000, 8)$ and $y_0 = (-500.5, -4)$, which is at least 8.

Applying Theorem 2(iii), we get

$$\|\xi - x_n\| \leq \frac{8}{3} M_{x_0, y_0} \sqrt{M_{x_0, y_0} - 2}(\frac{1}{4})^n,$$

The number n of iterations, needed by a priori error estimate less than 0.005 with an initial points $x_0 = (1000, 8)$ and $y_0 = (-500.5, -4)$, is at least 13.

Similarly, it is shown that the number n of iterations, needed by a posteriori estimate less than 0.005 for $\lambda = \frac{1}{4}$ with initial points $x_0 = (1000, 8)$ and

$\lambda = \frac{1}{16}$	x_n y_n	M_{x_n,y_n}	M_n
$n = 0$	$(1000, 8)$ $(-500.5, -4)$	1500.547983	154900.90193
$n = 1$	$(63.4375, 0.5)$ $(-32.21875, -0.25)$	95.65919017	2468.71315
$n = 2$	$(4.90234375, 3.125 \times 10^{-2})$ $(-2.951171875, -1.5625 \times 10^{-2})$	7.8536555	50.67037
$n = 3$	$(1.243896484, 1.953125 \times 10^{-3})$ $(-1.121948242, -9.765625 \times 10^{-4})$	2.3658465	3.81596
$n = 4$	$(1.01524353, 1.220703125 \times 10^{-4})$ $(-1.007621765, -6.103515625 \times 10^{-5})$	2.0228653	0.81568
$n = 5$	$(1.000952721, 7.629394531 \times 10^{-6})$ $(-1.00047636, -3.81469726 \times 10^{-6})$	2.0014290	0.20176
$n = 6$	$(1.000059545, 4.768371582 \times 10^{-7})$ $(-1.000029773, -2.384185791 \times 10^{-7})$	2.0000893	0.05040
$n = 7$	$(1.0000037215, 2.98023223 \times 10^{-8})$ $(-1.0000018607, -1.490116119 \times 10^{-8})$	2.0000055	0.01260
$n = 8$	$(1.0000002325, 1.8626451 \times 10^{-9})$ $(-1.0000001162, -9.31322574 \times 10^{-10})$	2.0000003	0.00315

$y_0 = (-500.5, -4)$, is at least 16. Also, the number n of iterations, needed by a priori error estimate less than 0.005, is at least 26.

5 Conclusion

In this article, we found a priori and a posteriori errors estimates for approximating fixed points for noncyclic contraction maps, which is defined on a uniformly convex Banach space with a modulus of convexity of power type. As seen in Example 1, a priori error estimate gives a larger number of iterations that are needed than a posteriori estimate. Therefore, it can be concluded that formula (iv) of Theorem 2 provides a better upper bound for error estimates.

Acknowledgements

References

[1] Beauzamy, B. *Introduction to Banach spaces and their geometry*, North–Holland Publishing Company, Amsterdam, 1979.

[2] Berinde, V. *Iterative approximation of fixed points*, Springer, Berlin, 2007.

[3] Deville, R., Godefroy, G. and Zizler, V. *Smoothness and renormings in Banach spaces*, Longman Scientific Technical, Harlow; copublished in the United States with John Wiley Sons, Inc., New York, 1993.

[4] Eldred, A. and Veeramani, P. *Existence and convergence of best proximity points*, J. Math. Anal. Appl. 323(2) (2006), 1001–1006.

[5] Fabian, M., Habala, P., Hajek, P., Montesinos, V., Pelant, J. and Zizler, V. *Functional analysis and infinite-dimensional geometry*, CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC, 8. Springer-Verlag, New York, 2001.

[6] Fallahi, K., Ghahramani, H. and Soleimani Rad, G. *Integral type contractions in partially ordered metric spaces and best proximity point*, Iran. J. Sci. Technol., Trans. A, Sci. 44 (2020), 177–183.

[7] Fallahi, K. and Soleimani Rad, G. *Best proximity points theorem in b-metric spaces endowed with a graph*, Fixed Point Theory, 21(2) (2020), 465–474.

[8] Gabeleh, M. and Shahzad, N. *Best proximity pair and fixed point results for noncyclic mappings in convex metric spaces*, Filomat, 30(12) (2016), 3149–3158.

[9] Goebel, K. and Kirk, W.A. *Topics in metric fixed point theory*, Cambridge University Press, 1990.

[10] Hristov, M., Ilchev, A. and Zlatanov, B. *On the best proximity points for p-cyclic summing contractions*, Mathematics, 8(7) (2020), 1–11.

[11] Ilchev, A. *Error estimates for approximating best proximity points for Kannan cyclic contractive maps*, AIP Conference Proceedings, 2048(1) (2018), 050002.

[12] Kirk, W., Srinivasan, P. and Veeramani, P. *Fixed points for mappings satisfying cyclical contractive conditions*, Fixed Point Theory, 4(1) (2003), 79–89.

[13] Meir, A. *On the Uniform Convexity of L_p Spaces $1 < p < 2$*, Illinois J. Math. 28(3) (1984), 420–424.

[14] Safari-Hafshejani, A. *The existence of best proximity points for generalized cyclic quasi-contractions in metric spaces with the UC and ultrametric properties*, Fixed Point Theory, 23(2) (2022), 507–518.

[15] Safari-Hafshejani, A., Amini-Harandi, A. and Fakhar, M. *Best proximity points and fixed points results for noncyclic and cyclic Fisher quasi-contractions*, Numer. Funct. Anal. Optim. 40(5) (2019), 603–619.

[16] Scheu, G. *On the optimization of the error estimate of a fixed point theorem*, J. Math. Anal. Appl. 57 (1977), 298–322.

[17] Zlatanov, B. *Error estimates for approximating best proximity points for cyclic contractive maps*, Carpathian J. Math. 32(2) (2016), 265–270.

How to cite this article

Solving two-dimensional coupled Burgers equations via a stable hybridized discontinuous Galerkin method

S. Baharlouei, R. Mokhtari*, [ID] and N. Chegini

Abstract

The purpose of this paper is to design a fully discrete hybridized discontinuous Galerkin (HDG) method for solving a system of two-dimensional (2D) coupled Burgers equations over a specified spatial domain. The semi-discrete HDG method is designed for a nonlinear variational formulation on the spatial domain. By exploiting broken Sobolev approximation spaces in the HDG scheme, numerical fluxes are defined properly. It is shown that the proposed method is stable under specific mild conditions on the stabilization parameters to solve a well-posed (in the sense of energy method) 2D coupled Burgers equations, which is imposed by Dirichlet boundary conditions. The fully discrete HDG scheme is designed by exploiting the Crank–Nicolson method for time discretization. Also, the Newton–Raphson method that has the order of at least two is nominated for solving the obtained nonlinear system of coupled Burgers equations over the rectangular domain. To reduce the complexity of the proposed method and the size of the linear system, we exploit the Schur complement idea. Numerical results declare that the best possible rates of convergence are achieved for approximate solutions of the 2D coupled Burgers equations and their first-order derivatives. Moreover, the proposed HDG method is examined for two other types of systems, that is, a system with high Reynolds numbers and a system with an unavailable exact solution. The acceptable results

*Corresponding author

Received 2 February 2023; revised 5 April 2023; accepted 10 April 2023

Shima Baharlouei
Department of Mathematical Sciences, Isfahan University of Technology, Isfahan 84156-83111, Iran.. e-mail: s.baharloui@math.iut.ac.ir

Reza Mokhtari
Department of Mathematical Sciences, Isfahan University of Technology, Isfahan 84156-83111, Iran. e-mail: mokhtari@iut.ac.ir

Nabi Chegini
Department of Mathematics, Tafresh University, Tafresh 39518-79611, Iran. e-mail: nabichegini@tafreshu.ac.ir

of examples show the flexibility of the proposed method in solving various problems.

AMS subject classifications (2020): 65M60, 65M12

Keywords: Coupled Burgers equations; hybridized discontinuous Galerkin method; stability analysis.

1 Introduction

Throughout the history of science, finding the analytical and especially numerical solutions of nonlinear evolution equations such as Burgers and coupled Burgers equations [3, 28, 31, 35], KdV type equations [2, 4, 27], Navier–Stokes equations [34], and nonlinear Schrödinger equations [7] play crucial roles in various fields of science and engineering for the detection of physical phenomena. The system of two-dimensional (2D) coupled Burgers equations, as a simplified form of some complex and practical equations in engineering such as the incompressible Navier-Stokes equation, is widely used in fluid dynamics such as modeling of the shock waves moving in viscous liquid [17], shallow water waves [18, 26], turbulent medium [5], and diffusion processes [1]. According to the new works that are done in the literature [24, 38, 39], we realize that providing methods of finding the numerical solutions of Burgers and coupled Burgers equations still have their importance. Moreover, numerical scientists consider Burgers and coupled Burgers equations as test problems to introduce and experiment with new numerical methods. In other words, these equations are used to compare different numerical methods in various aspects to choose and extend the most appropriate one to a specialized subject. This paper proposes a stable scheme for solving the 2D nonlinear coupled Burgers equations over rectangular domains numerically.

The general form of the 2D system of coupled Burgers equations reads as

$$
\begin{cases}
\mathfrak{u}_t + \mathfrak{u}\mathfrak{u}_x + \mathfrak{v}\mathfrak{u}_y - \dfrac{1}{\mathrm{Re}}(\mathfrak{u}_{xx} + \mathfrak{u}_{yy}) = 0, \\
\mathfrak{v}_t + \mathfrak{u}\mathfrak{v}_x + \mathfrak{v}\mathfrak{v}_y - \dfrac{1}{\mathrm{Re}}(\mathfrak{v}_{xx} + \mathfrak{v}_{yy}) = 0,
\end{cases}
$$

or equivalently

$$
\begin{cases}
\mathfrak{u}_t + \mathbf{U} \cdot \nabla \mathfrak{u} - \dfrac{1}{\mathrm{Re}}\Delta \mathfrak{u} = 0, \\
\mathfrak{v}_t + \mathbf{U} \cdot \nabla \mathfrak{v} - \dfrac{1}{\mathrm{Re}}\Delta \mathfrak{u} = 0,
\end{cases}
\tag{1}
$$

where $\mathrm{Re} > 0$ is the Reynolds number, $\mathbf{U} = (\mathfrak{u}, \mathfrak{v})^{\mathsf{T}}$, and $\mathbf{x} = (x, y) \in \Omega = (a, b) \times (c, d) \subset \mathbb{R}^2$. In this paper, system (1) is equipped by the Dirichlet boundary conditions and suitable initial conditions.

Analytical solution of system (1) can be obtained, for instance, by the Hopf–Cole transformation; see [20]. Providing an explicit analytical solution for system (1) is not trivial. However, if, by any chance, an explicit form becomes available, then evaluating the analytical solution requires high computational costs with a considerable amount of time, which may be accompanied by uncontrollable errors regarding the discretization of the analytical solution. Based on these reasons, it is requested design be stable and effective numerical methods for computing numerical solutions. For solving system (1), many numerical methods have been proposed, such as the decomposition method [19], Chebyshev spectral collocation method [23], and some others; for instance, see [24, 38, 39].

Since the main approach of this paper is directly related to the discontinuous Galerkin (DG) method and is considered a continuation of the local discontinuous Galerkin (LDG) method, it is necessary to briefly review the history and background of DG and LDG methods. The first DG method was proposed by Reed and Hill in 1973 for a time-independent linear hyperbolic equation [6], and then it was utilized and developed for time-dependent partial differential equations (PDEs); see [10, 16]. Provable cell-entropy inequality for L^2 stability, h-p adaptivity, and flexibility to handle complicated geometry for arbitrary order of accuracy with local in-data communication, and other abilities lead to applying the DG method to various types of differential equations. To dominate the limitations of the DG method for solving high-order partial differential equations, an LDG method was proposed. This method was used for the first time for solving a second-order time-dependent convection-diffusion equation [15]. The main idea of the LDG method is the transformation of a high-order equation into a first-order system of equations before solving the new system by the DG method. Due to eliminating all of the auxiliary variables locally, the LDG method inherits all flexibilities of the DG method. Recent applications of the LDG method for higher-order nonlinear PDEs can be found, for instance, in [8, 25, 31].

The usage of the hybridization technique in the context of the finite element method goes back many years ago, while its application in the context of DG methods has a recent history and goes back to 2004. In fact, the hybridized discontinuous Galerkin (HDG) method was proposed for the first time by combining the DG method and continuous Galerkin (CG) method to solve the steady-state problems [11], and then it was generalized by Cockburn et al. [12, 13, 14]. Recently, HDG methods have been widely used to solve evolution equations numerically, in particular for compressible flow problems [22, 30, 33, 36, 37], Stokes flow [9, 21], continuum mechanics problems [29], and linear elasticity problems [32]. The HDG methods inherit the optimal convergence rate from the DG methods for approximate solutions and their derivatives with respect to spatial variables. HDG methods have two kinds of unknowns; global unknowns that are used in the definition of numerical traces (or in numerical fluxes) and obtained from the global system, and local unknowns that can be eliminated locally and are obtained by weak formula-

tion. Local and global unknowns are approximated by piecewise polynomials of degree k, respectively, in $\mathbb{R}^d$ and $\mathbb{R}^{d-1}$, where d is the dimension of the spatial domain. Due to the consideration of global unknowns, one can infer that the degree of freedom in the HDG method is reduced compared to the traditional implicit DG methods. The key to the success of the HDG method is the way of defining numerical fluxes that are based on global unknowns and stabilization parameters. The numerical fluxes of the HDG method are not defined uniquely in most situations, but those have to be defined in such a way that the desired definitions of numerical fluxes ensure the stability of the scheme. Also, the definitions of the numerical fluxes cause significantly smaller bandwidth than the corresponding matrices of the traditional CG method, and therefore lower computational cost is accessible in any HDG method. In solving a problem with nonsmooth solutions, the HDG method as a kind of DG method is a suitable scheme. This advantage is based on the fact that the HDG method produces numerical approximations using discontinuous trial functions over the entire given domain. In summary, it is worth pointing out that the HDG method has unique properties, which make this method superior, such as reducing the degree of freedom compared to the traditional implicit DG methods, making smaller bandwidth compared to the corresponding matrices of traditional CG and DG methods, and having less computational time; see [4]. In this paper, we intend to use a kind of HDG method for discretizing the 2D coupled Burgers equations (1) in the spatial domain.

The rest of the paper is organized as follows. In Section 2, some prerequisites such as notations, discretization of temporal and spatial domains, and approximation spaces, are expressed in dimension two. Section 3 is dedicated to the employment of the HDG method to the 2D coupled Burgers equations. In fact, in this section, a semi-discrete scheme is presented for the 2D coupled Burgers equations with suitable definitions of numerical fluxes and stabilization parameters. In addition, the stability of the proposed semi-discrete HDG scheme is investigated in this section. In other words, we prove that the method is stable in the L^2 norm under certain conditions on the stabilization parameters. Then, a full discretization approach is designed in Section 4 by exploiting the Crank–Nicolson method for time discretization and Newton–Raphson as a nonlinear solver. Numerical experiments in Section 5 show that the optimal order of accuracy is derived by the proposed method. Also, by performing some experiments, the numerical solutions of system (1) are investigated for large Reynolds numbers. Moreover, a 2D problem with different values of Reynolds numbers is investigated such that its exact solution is unavailable. The conclusion is given in Section 6. The paper is ended with an Appendix.

2 Prerequisites

Order to set up a system of weak formulation of coupled Burgers equations, it requires defining necessary notations and relevant approximation spaces for a desired HDG method. With T as a final time and for all $t \in (0, T]$, we consider a given bounded spatial domain $\Omega \subset \mathbb{R}^2$ with suitable partitioning. Suppose that the domain $\Omega = (a, b) \times (c, d)$ is split into conforming and uniform finite element meshes with N triangles such that in this mesh generation, all triangles have no intersection except in common edges or vertices. In general, each of these triangles is denoted by $\mathcal{K}$. By considering h as the longest edge among triangles, the finite collection of disjoint elements, and the set of the boundaries of elements, respectively, are denoted by

$$\mathscr{K}_h := \{\mathcal{K}\}, \qquad \partial \mathscr{K}_h = \{\partial \mathcal{K}\},$$

where $\Omega = \bigcup_{\mathcal{K} \in \mathscr{K}_h} \mathcal{K}$, and $\partial \mathcal{K}$ denotes the boundary of element $\mathcal{K}$. The collection $\mathcal{F}_h = \mathcal{F}_h^\partial \cup \mathcal{F}_h^0$ is the set of all faces such that $\mathcal{F}_h^0$ and $\mathcal{F}_h^\partial$ represent, respectively, the set of interior and boundary faces. More precisely, the set of faces contains all edges of triangles. Let us consider two elements $\mathcal{K}^-$ and $\mathcal{K}^+$ and their common face $e = \partial \mathcal{K}^- \cap \partial \mathcal{K}^+ \in \mathcal{F}_h^0$. As illustrated in Figure 1, $\mathbf{n}^-$ and $\mathbf{n}^+$ are, respectively, the corresponding outward unit normal vectors of face e with respect to $\mathcal{K}^-$ and $\mathcal{K}^+$. Let v^- and v^+ be the limits of the function v at face e associated with $\partial \mathcal{K}^+$ and $\partial \mathcal{K}^-$, respectively. Thus the mean and jump values of an arbitrary real valued function v on the given face e are, respectively, defined as

$$\{\{v\}\} = \frac{1}{2}(v^- + v^+), \quad [[v]] = v^- \mathbf{n}^- + v^+ \mathbf{n}^+.$$

We note that the mean and jump values of function v at boundary face
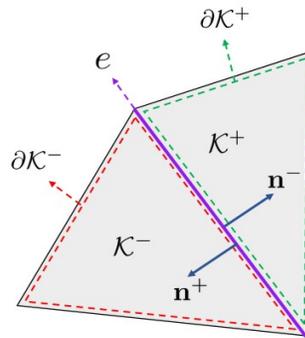
Figure 1: Common face e of two elements $\mathcal{K}^+, \mathcal{K}^-$ with outward unit vectors.

$e = \partial\Omega \cap \mathcal{K} \in \mathcal{F}_h^\partial$ are determined as $\{\{v\}\} = v$ and $[[v]] = v\mathbf{n}$, respectively. So, the mean and jump of function v can be rewritten as

$$\{\{v\}\} = \begin{cases} (v^+ + v^-)/2, & e \in \mathcal{F}_h^0, \\ v, & e \in \mathcal{F}_h^\partial, \end{cases} \qquad [[v\mathbf{n}]] = \begin{cases} v^+\mathbf{n}^+ + v^-\mathbf{n}^-, & e \in \mathcal{F}_h^0, \\ v\mathbf{n}, & e \in \mathcal{F}_h^\partial. \end{cases}$$

To obtain weak formulations of the 2D coupled Burgers equations, one needs to define appropriate approximation spaces. Regarding the nature of any DG method, broken Sobolev spaces are relevant spaces for approximating the solutions of system (1) via the HDG method. The corresponding broken Sobolev space, associated with the partition $\mathcal{K}_h$ is defined as

$$H^1(\mathcal{K}_h) = \{v : \Omega \to \mathbb{R} : v\,|_\mathcal{K} \in H^1(\mathcal{K}), \text{for all } \mathcal{K} \in \mathcal{K}_h\},$$

and associated with the set $\mathcal{F}_h$ is defined as

$$\mathcal{M}^1(\mathcal{F}_h) = \{\mu : \mathcal{F}_h \to \mathbb{R} : \mu\,|_e \in H^1(e), \text{for all } e \in \mathcal{F}_h\}.$$

Discontinuous finite element spaces for scalar and vector valued functions, as subspaces of broken Sobolev space $H^1(\mathcal{K}_h)$ are, respectively, defined by

$$S_{h,k} = \left\{ w \in H^1(\mathcal{K}_h) : w\,|_\mathcal{K} \in \mathcal{P}_k(\mathcal{K}), \text{for all } \mathcal{K} \in \mathcal{K}_h \right\},$$
$$\mathbf{S}_{h,k} = \left\{ w \in (H^1(\mathcal{K}_h))^2 : w\,|_\mathcal{K} \in (\mathcal{P}_k(\mathcal{K}))^2, \text{for all } \mathcal{K} \in \mathcal{K}_h \right\},$$

where $\mathcal{P}_k(\mathcal{K})$ is the set of polynomials of degree at most k on the element $\mathcal{K} \in \mathcal{K}_h$. The approximation space of the broken Sobolev space over $\mathcal{F}_h$ (or skeleton space) is defined as

$$M_{h,k} = \{\mu \in \mathcal{M}^1(\mathcal{F}_h) : \mu\,|_e \in \mathcal{P}_k(e), \text{for all } e \in \mathcal{F}_h\}.$$

Regarding the boundary conditions, it is needed to define the appropriate subspace of the skeleton space. Consider Dirichlet boundary conditions and the boundary data $b_\mathfrak{u}$ and $b_\mathfrak{v}$ on $\partial\Omega$, which are associated with $\mathfrak{u}$ and $\mathfrak{v}$, respectively. Let $\Gamma_\mathfrak{u}$ and $\Gamma_\mathfrak{v}$ be collections of boundary faces in which boundary data $b_\mathfrak{u}$ and $b_\mathfrak{v}$ are specified over $\Gamma_\mathfrak{u}$ and $\Gamma_\mathfrak{v}$, respectively. Based on the given boundary conditions, we define

$$M_{h,k}(l, \Gamma) := \{\mu \in M_{h,k} : \mu(\mathbf{x}) = \Pi l(\mathbf{x}),\ \mathbf{x} \in \Gamma\},$$

where $\Gamma \in \{\Gamma_\mathfrak{u}, \Gamma_\mathfrak{v}\}$, and Π is the L^2 projection with respect to the skeleton space of the boundary of the domain Ω. The approximation spaces $S_{h,k}$, $\mathbf{S}_{h,k}$, and $M_{h,k}$ are equipped by the following inner products, respectively,

$$(w_1, w_2)_{\mathcal{K}_h} = \sum_{\mathcal{K} \in \mathcal{K}_h} (w_1, w_2)_\mathcal{K}, \quad \langle \mu_1, \mu_2 \rangle_{\partial \mathcal{K}_h} = \sum_{\mathcal{K} \in \mathcal{K}_h} \langle \mu_1, \mu_2 \rangle_{\partial \mathcal{K}},$$

where

$$(w_1, w_2)_{\mathcal{K}} = \int_{\mathcal{K}} w_1(\mathbf{x}) \cdot w_2(\mathbf{x}) \ d\mathbf{x}, \quad \langle \mu_1, \mu_2 \rangle_{\partial \mathcal{K}} = \int_{\partial \mathcal{K}} \mu_1 \cdot \mu_2 \ ds,$$

in which w_1, w_2 are defined on $\mathcal{K}_h$ and μ_1, μ_2 are defined on $\partial \mathcal{K}_h$. By considering vector functions $\boldsymbol{w} = (w_1, w_2)^\mathsf{T}$, $\boldsymbol{z} = (z_1, z_2)^\mathsf{T}$, $\boldsymbol{\mu} = (\mu_1, \mu_2)^\mathsf{T}$, and $\boldsymbol{\eta} = (\eta_1, \eta_2)^\mathsf{T}$, the following inner products are needed

$$(\boldsymbol{w}, \boldsymbol{z})_{\mathcal{K}_h} = \sum_{\mathcal{K} \in \mathcal{K}_h} (\boldsymbol{w}, \boldsymbol{z})_{\mathcal{K}}, \quad \langle \boldsymbol{\mu}, \boldsymbol{\eta} \rangle_{\partial \mathcal{K}_h} = \sum_{\mathcal{K} \in \mathcal{K}_h} \langle \boldsymbol{\mu}, \boldsymbol{\eta} \rangle_{\partial \mathcal{K}},$$

where w_1, w_2, z_1, and z_2 are defined on $\mathcal{K}_h$, and μ_1, μ_2, η_1, and η_2 are defined on $\partial \mathcal{K}_h$. Besides, we have

$$(\boldsymbol{w}, \boldsymbol{z})_{\mathcal{K}} = ((w_1, z_1)_{\mathcal{K}}, \ (w_2, z_2)_{\mathcal{K}})^\mathsf{T}, \qquad \langle \boldsymbol{\mu}, \boldsymbol{\eta} \rangle_{\partial \mathcal{K}} = (\langle \mu_1, \eta_1 \rangle_{\partial \mathcal{K}}, \langle \mu_2, \eta_2 \rangle_{\partial \mathcal{K}})^\mathsf{T}.$$

3 Construction of the semi-discrete HDG method

As mentioned, we assume that system (1) is equipped by the Dirichlet boundary conditions over the rectangular domain Ω. The initial step is to reformulate the 2D coupled Burgers equations (1) into a first-order system of equations. By defining the auxiliary variables $\mathbf{P} = (\mathfrak{p}_1, \mathfrak{p}_2)^\mathsf{T} = (\nabla \mathfrak{u})^\mathsf{T}$ and $\mathbf{Q} = (\mathfrak{q}_1, \mathfrak{q}_2)^\mathsf{T} = (\nabla \mathfrak{v})^\mathsf{T}$, the corresponding first-order system of (1) reads as

$$\begin{cases} \mathfrak{u}_t + \mathbf{U} \cdot \nabla \mathfrak{u} - \dfrac{1}{\mathrm{Re}} \nabla \cdot \mathbf{P} = 0, \\ \qquad\qquad\qquad \mathbf{P} - \nabla \mathfrak{u} = 0, \\ \mathfrak{v}_t + \mathbf{U} \cdot \nabla \mathfrak{v} - \dfrac{1}{\mathrm{Re}} \nabla \cdot \mathbf{Q} = 0, \\ \qquad\qquad\qquad \mathbf{Q} - \nabla \mathfrak{v} = 0. \end{cases} \tag{2}$$

By establishing the corresponding semi-discrete HDG method of the system (2), the stability of the semi-discrete method over the temporal interval $[0, t]$ for $t \in (0, T]$, is explained in the next subsection.

To have a corresponding conditionally well-posed problem of the system (2), it is worth pointing out that this system should be equipped with initial and boundary conditions. Weak formulation of the system (2) can be formed by multiplying each equation of (2) by an appropriate test function, integrating over each element $\mathcal{K} \in \mathcal{K}_h$, and using the Green's first identity. Consequently, the aim is to find numerical approximations $(u, v, P, Q) \in S_{h,k}^2 \times \mathbf{S}_{h,k}^2$ such that for all test functions $(w_1, w_2, \mathbf{w}_1, \mathbf{w}_2) \in S_{h,k}^2 \times \mathbf{S}_{h,k}^2$ and $\mathcal{K} \in \mathcal{K}_h$, it holds that

$$
\begin{cases}
(u_t, w_1)_{\mathcal{K}} + (\mathbf{U} \cdot \nabla \mathfrak{u}, w_1)_{\mathcal{K}} + (\frac{1}{\text{Re}}P, \nabla w_1)_{\mathcal{K}} + \langle -\widehat{\frac{1}{\text{Re}}P}\mathbf{n}, w_1 \rangle_{\partial \mathcal{K}} = 0, \\
\qquad ((P, \mathbf{w}_1))_{\mathcal{K}} + ((u, \nabla \mathbf{w}_1))_{\mathcal{K}} - \widehat{\langle\langle \hat{u}\mathbf{n}, \mathbf{w}_1 \rangle\rangle}_{\partial \mathcal{K}} = 0, \\
(v_t, w_2)_{\mathcal{K}} + (\mathbf{U} \cdot \nabla \mathfrak{v}, w_2)_{\mathcal{K}} + (\frac{1}{\text{Re}}Q, \nabla w_2)_{\mathcal{K}} + \langle -\widehat{\frac{1}{\text{Re}}Q}\mathbf{n}, w_2 \rangle_{\partial \mathcal{K}} = 0, \\
\qquad ((Q, \mathbf{w}_2))_{\mathcal{K}} + ((v, \nabla \mathbf{w}_2))_{\mathcal{K}} - \langle\langle \hat{v}\mathbf{n}, \mathbf{w}_2 \rangle\rangle_{\partial \mathcal{K}} = 0.
\end{cases} \tag{3}
$$

- Under imposed boundary conditions, numerical traces $\hat{u} \in M_{h,k}(b_{\mathfrak{u}}, \Gamma_{\mathfrak{u}})$ and $\hat{v} \in M_{h,k}(b_{\mathfrak{v}}, \Gamma_{\mathfrak{v}})$ are properly defined for all $z \in \mathcal{F}_h$ as

$$
\hat{u}(z) = \begin{cases} b_{\mathfrak{u}}, \, z \in \Gamma_{\mathfrak{u}}, \\ \xi, \;\; z \in \mathcal{F}_h \setminus \Gamma_{\mathfrak{u}}, \end{cases} \qquad \hat{v}(z) = \begin{cases} b_{\mathfrak{v}}, \, z \in \Gamma_{\mathfrak{v}}, \\ \zeta, \;\; z \in \mathcal{F}_h \setminus \Gamma_{\mathfrak{v}}, \end{cases} \tag{4}
$$

where $(\xi, \zeta) \in M_{h,k}(0, \Gamma_{\mathfrak{u}}) \times M_{h,k}(0, \Gamma_{\mathfrak{v}})$ is a global unknown pair. It can be observed that boundary data $b_{\mathfrak{u}}$ and $b_{\mathfrak{v}}$ are imposed in the definitions of the numerical traces $\hat{u}$ and $\hat{v}$, respectively, on $\Gamma_{\mathfrak{u}}$ and $\Gamma_{\mathfrak{v}}$. One can infer that $\hat{u}$ and $\hat{v}$ are global unknowns corresponding to the faces without a defined boundary data.

- In order to guarantee the stability of the semi-discrete method, numerical fluxes $-\widehat{\frac{1}{\text{Re}}P}$ and $-\widehat{\frac{1}{\text{Re}}Q}$ are defined as

$$
-\widehat{\frac{1}{\text{Re}}P} = -\frac{1}{\text{Re}}P + \tau(u - \hat{u})\mathbf{n}, \qquad -\widehat{\frac{1}{\text{Re}}Q} = -\frac{1}{\text{Re}}Q + \sigma(v - \hat{v})\mathbf{n},
$$
$$\tag{5}$$

where $\mathbf{n}$ is the outward unit normal vector with respect to the considered face. In (5), τ and σ are the stabilization parameters. The valid range of parameters τ and σ are determined in the stability theorem of the proceeding subsection. We note that the definitions of the numerical fluxes in (5) are not unique and depend on the form and physics of the problems.

Remark 1. It is noteworthy that numerical fluxes and stabilization parameters play a key role in the stability of the semi-discrete method. We emphasize that functions $-\frac{1}{\text{Re}}P$ and $-\frac{1}{\text{Re}}Q$ on each element edge are approximated by their corresponding numerical fluxes so that the numerical fluxes are single-valued continuous functions across the element edges. In HDG methods, numerical fluxes depend on the numerical traces while global unknowns in the definitions of numerical traces depend on the faces.

Due to the fact that $\hat{u}$ and $\hat{v}$ contain two global unknown variables over $[0, T] \times \Omega$, two extra global equations on each face should be added to the system (3). The required global equations can be gained by enforcing the conservation of the fluxes. Thus, the global unknowns are obtained with the following extra global equations:

$$[[\widehat{\frac{1}{\text{Re}}P \cdot \mathbf{n}}]] = 0, \ \text{ for } \ e \in \mathcal{F}_h^0, \qquad [[\widehat{\frac{1}{\text{Re}}Q \cdot \mathbf{n}}]] = 0, \ \text{ for } \ e \in \mathcal{F}_h^0. \qquad (6)$$

Then the local unknowns u, v, P, and Q, can be found by solving weak formulation (3) in each element $\mathcal{K} \in \mathscr{K}_h$.

3.1 Stability analysis

In this subsection, we verify the numerical stability of the weak formulation (3) over the time interval $[0, t]$, for all $t \in (0, T]$. To do this, let homogeneous Dirichlet boundary conditions imposed to the weak formulation (3). We start the analysis by multiplying the first equation of (1) by $\mathfrak{u}$ to get

$$\frac{1}{2}\frac{d}{dt}\mathfrak{u}^2 + \frac{1}{2}\mathbf{U} \cdot \nabla(\mathfrak{u}^2) - \frac{1}{\text{Re}}\mathfrak{u}\Delta\mathfrak{u} = 0. \qquad (7)$$

By integrating (7) over the given domain Ω and using the Green's first identity, we get

$$\frac{1}{2}\frac{d}{dt}\int_\Omega \mathfrak{u}^2 \, d\mathbf{x} + \frac{1}{2}\int_\Omega \mathbf{U}\cdot\nabla(\mathfrak{u}^2) \, d\mathbf{x} + \frac{1}{\text{Re}}\int_\Omega \nabla\mathfrak{u}\cdot\nabla\mathfrak{u} \, d\mathbf{x} - \frac{1}{\text{Re}}\int_\Gamma \mathfrak{u}\frac{\partial\mathfrak{u}}{\partial\mathbf{n}} \, ds = 0. \qquad (8)$$

By applying homogeneous Dirichlet boundary conditions to (8) and regarding $\int_\Omega \nabla\mathfrak{u} \cdot \nabla\mathfrak{u} \, d\mathbf{x} \geq 0$, one can conclude that (8) leads to

$$\frac{1}{2}\frac{d}{dt}\int_\Omega \mathfrak{u}^2 \, d\mathbf{x} + \frac{1}{2}\int_\Omega \mathbf{U} \cdot \nabla(\mathfrak{u}^2) \, d\mathbf{x} \leq 0. \qquad (9)$$

Integrating (9) over the time interval $[0, t]$, for $0 < t \leq T$, the following inequality holds:

$$\|\mathfrak{u}(\cdot, t)\|_\Omega^2 + \int_0^T \Phi(\mathfrak{u}, \Omega) \, d\mathbf{x} \leq \|\mathfrak{u}(\cdot, 0)\|_\Omega^2, \qquad (10)$$

with

$$\Phi(\mathfrak{r}, \Lambda) = \int_\Lambda \mathbf{U} \cdot \nabla(\mathfrak{r}^2) \, d\mathbf{x},$$

where $\mathfrak{r}$ is the function of $\mathbf{x}$ and t, and Λ is a subdomain of Ω. Also, Output of $\Phi(\mathfrak{r}, \Lambda)$ is a function of variable t. In the same approach, from the second equation of (1), we get

$$\|\mathfrak{v}(\cdot, t)\|_\Omega^2 + \int_0^T \Phi(\mathfrak{v}, \Omega) \, dt \leq \|\mathfrak{v}(\cdot, 0)\|_\Omega^2. \qquad (11)$$

Theorem 1. Let weak formulation (3) be equipped by the homogeneous Dirichlet boundary conditions over the domain Ω. By assuming $\tau > 0$ and $\sigma > 0$, it can be proved that the solution of weak formulation (3) satisfies the following inequalities for all $t \in (0, T]$:

$$\|u(\cdot, t)\|_{\mathscr{K}_h}^2 + \int_0^T \Phi(u, \mathscr{K}_h) \, dt \leq \|u(\cdot, 0)\|_{\mathscr{K}_h}^2,$$
$$\|v(\cdot, t)\|_{\mathscr{K}_h}^2 + \int_0^T \Phi(v, \mathscr{K}_h) \, dt \leq \|v(\cdot, 0)\|_{\mathscr{K}_h}^2.$$

Proof. By setting $w_1 = u$, $\mathbf{w}_1 = \frac{1}{\text{Re}} P$, $w_2 = v$, and $\mathbf{w}_2 = \frac{1}{\text{Re}} Q$, in the weak formulation (3) and summing the first three equations and the last three equations of (3) together, we get

$$\frac{1}{2} \frac{d}{dt} \|u\|_{\mathcal{K}}^2 + \frac{1}{\text{Re}} \|P\|_{\mathcal{K}}^2 + \bar{E}_{1,\mathcal{K}} + \frac{1}{2} \Phi(u, \mathcal{K}) = 0,$$
$$\frac{1}{2} \frac{d}{dt} \|v\|_{\mathcal{K}}^2 + \frac{1}{\text{Re}} \|Q\|_{\mathcal{K}}^2 + \bar{E}_{2,\mathcal{K}} + \frac{1}{2} \Phi(v, \mathcal{K}) = 0,$$

(12)

where

$$\bar{E}_{1,\mathcal{K}} = \frac{1}{\text{Re}} (P, \nabla u)_{\mathcal{K}} + \frac{1}{\text{Re}} (u, \nabla P)_{\mathcal{K}} + \langle \widehat{-\frac{1}{\text{Re}} P} \cdot \mathbf{n}, u \rangle_{\partial \mathcal{K}} - \frac{1}{\text{Re}} \langle \hat{u} \mathbf{n}, P \rangle_{\partial \mathcal{K}},$$
$$\bar{E}_{2,\mathcal{K}} = \frac{1}{\text{Re}} (Q, \nabla v)_{\mathcal{K}} + \frac{1}{\text{Re}} (v, \nabla Q)_{\mathcal{K}} + \langle -\frac{1}{\text{Re}} Q \cdot \mathbf{n}, v \rangle_{\partial \mathcal{K}} - \frac{1}{\text{Re}} \langle \hat{v} \mathbf{n}, Q \rangle_{\partial \mathcal{K}}.$$

Using the divergence theorem, the following relations are obtained:

$$\frac{1}{\text{Re}} (P, \nabla u)_{\mathcal{K}} + \frac{1}{\text{Re}} (u, \nabla P)_{\mathcal{K}} = \frac{1}{\text{Re}} \int_{\mathcal{K}} \nabla \cdot (Pu) \, dx = \frac{1}{\text{Re}} \int_{\partial \mathcal{K}} (Pu) \cdot \mathbf{n} \, ds$$
$$= \frac{1}{\text{Re}} \langle P \cdot \mathbf{n}, u \rangle_{\partial \mathcal{K}},$$
$$\frac{1}{\text{Re}} (Q, \nabla v)_{\mathcal{K}} + \frac{1}{\text{Re}} (v, \nabla Q)_{\mathcal{K}} = \frac{1}{\text{Re}} \int_{\mathcal{K}} \nabla \cdot (Qv) \, dx = \frac{1}{\text{Re}} \int_{\partial \mathcal{K}} (Qv) \cdot \mathbf{n} \, ds$$
$$= \frac{1}{\text{Re}} \langle Q \cdot \mathbf{n}, v \rangle_{\partial \mathcal{K}},$$

(13)

By applying (13) to $\bar{E}_{1,\mathcal{K}}$ and $\bar{E}_{2,\mathcal{K}}$, using

$$\langle \hat{u} \mathbf{n}, P \rangle_{\partial \mathcal{K}} = \langle P \cdot \mathbf{n}, \hat{u} \rangle_{\partial \mathcal{K}}, \qquad \langle \hat{v} \mathbf{n}, Q \rangle_{\partial \mathcal{K}} = \langle Q \cdot \mathbf{n}, \hat{v} \rangle_{\partial \mathcal{K}},$$

and adding

$$-\langle \widehat{-\frac{1}{\text{Re}} P} \cdot \mathbf{n}, \hat{u} \rangle_{\partial \mathcal{K}} = 0, \qquad -\langle \widehat{-\frac{1}{\text{Re}} Q} \cdot \mathbf{n}, \hat{v} \rangle_{\partial \mathcal{K}} = 0,$$

respectively, into $\bar{E}_{1,\mathcal{K}}$ and $\bar{E}_{2,\mathcal{K}}$, we obtain

$$\bar{E}_{1,\mathcal{K}} = \langle \widehat{-\frac{1}{\text{Re}} P} \cdot \mathbf{n}, u - \hat{u} \rangle_{\partial \mathcal{K}} + \langle \frac{1}{\text{Re}} P \cdot \mathbf{n}, u - \hat{u} \rangle_{\partial \mathcal{K}}$$
$$= \langle (\widehat{-\frac{1}{\text{Re}} P} + \frac{1}{\text{Re}} P) \cdot \mathbf{n}, u - \hat{u} \rangle_{\partial \mathcal{K}},$$
$$\bar{E}_{2,\mathcal{K}} = \langle \widehat{-\frac{1}{\text{Re}} Q} \cdot \mathbf{n}, v - \hat{v} \rangle_{\partial \mathcal{K}} + \langle \frac{1}{\text{Re}} Q \cdot \mathbf{n}, v - \hat{v} \rangle_{\partial \mathcal{K}}$$
$$= \langle (\widehat{-\frac{1}{\text{Re}} Q} + \frac{1}{\text{Re}} Q) \cdot \mathbf{n}, v - \hat{v} \rangle_{\partial \mathcal{K}}.$$

Using the definitions of $-\widehat{\frac{1}{\mathrm{Re}}P}$ and $-\widehat{\frac{1}{\mathrm{Re}}Q}$ from (5), we obtain

$$\bar{E}_{1,\mathcal{K}} = <\tau, (u-\hat{u})^2>_{\partial\mathcal{K}}, \qquad \bar{E}_{2,\mathcal{K}} = <\sigma, (v-\hat{v})^2>_{\partial\mathcal{K}}.$$

By summing $\bar{E}_{1,\mathcal{K}}$ and $\bar{E}_{2,\mathcal{K}}$ over all elements, we get

$$\sum_{\mathcal{K}\in\mathscr{K}_h}\bar{E}_{1,\mathcal{K}} = \sum_{\mathcal{K}\in\mathscr{K}_h}<\tau,(u-\hat{u})^2>_{\partial\mathcal{K}} = <\tau,(u-\hat{u})^2>_{\partial\mathscr{K}_h},$$
$$\sum_{\mathcal{K}\in\mathscr{K}_h}\bar{E}_{2,\mathcal{K}} = \sum_{\mathcal{K}\in\mathscr{K}_h}<\sigma,(v-\hat{v})^2>_{\partial\mathcal{K}} = <\sigma,(v-\hat{v})^2>_{\partial\mathscr{K}_h}.$$

According to the assumptions $\tau > 0$ and $\sigma > 0$, we can conclude $\sum_{\mathcal{K}\in\mathscr{K}_h}\bar{E}_{1,\mathcal{K}} \geq 0$ and $\sum_{\mathcal{K}\in\mathscr{K}_h}\bar{E}_{2,\mathcal{K}} \geq 0$. Finally, by summing (12) over all elements, using the obtained results, and $\|P\|^2_{\mathscr{K}_h}, \|Q\|^2_{\mathscr{K}_h} \geq 0$, we conclude

$$\frac{d}{dt}\|u\|^2_{\mathscr{K}_h} + \Phi(u, \mathscr{K}_h) \leq 0, \qquad \frac{d}{dt}\|v\|^2_{\mathscr{K}_h} + \Phi(v, \mathscr{K}_h) \leq 0.$$

By integrating above relations over $[0, t]$ for all $t \in (0, T]$, the assertion of the theorem is concluded. $\qquad\square$

Remark 2. According to (10)–(11), by assuming

$$\int_0^T \Phi(\mathfrak{u}, \Omega)\ dt \geq 0, \qquad \int_0^T \Phi(\mathfrak{v}, \Omega)\ dt \geq 0,$$

one can verify that the 2D coupled Burgers equations (1) is well-posed in the sense of the energy method. Therefore, in this case and based on Theorem 1, the proposed HDG method is stable with $\tau > 0$ and $\sigma > 0$.

Briefly, Theorem 1 and Remark 2 show that the proposed semi-discrete HDG method is stable for solving well-posed 2D coupled Burgers equations provided some specific mild conditions on the stabilization parameters. Moreover, this stability is unconditional because we have no condition on the step sizes.

4 Numerical algorithm and implementation issues

In order to design a fully discrete approximation method for solving the 2D nonlinear coupled Burgers equations (1), it is needed to apply a time-discretization approach to the weak formulation (3). To do this, we simply use the Crank–Nicolson method which is a method of order two. By considering time step $\Delta t = \frac{T}{J}$ with $J \in \mathbb{N}$ and time level $t_n = n\Delta t$, for $n = 0, \ldots, J$, the weak formulation (3) changes to

$$
\begin{cases}
\frac{1}{\Delta t}(u^n, w_1)_\mathcal{K} + \frac{1}{2}(U^n \cdot \nabla \mathfrak{u}^n, w_1)_\mathcal{K} + \frac{1}{2}(\frac{1}{\text{Re}}P^n, \nabla w_1)_\mathcal{K} + \frac{1}{2}\langle(\widehat{-\frac{1}{\text{Re}}P})^n\mathbf{n}, w_1\rangle_{\partial\mathcal{K}} = l_1(w_1), \\
\qquad\qquad\qquad ((P^n, \mathbf{w}_1))_\mathcal{K} + ((u^n, \nabla\mathbf{w}_1))_\mathcal{K} - \langle\langle\hat{u}^n\mathbf{n}, \mathbf{w}_1\rangle\rangle_{\partial\mathcal{K}} = 0, \\[2mm]
\frac{1}{\Delta t}(v^n, w_2)_\mathcal{K} + \frac{1}{2}(U^n \cdot \nabla \mathfrak{v}^n, w_2)_\mathcal{K} + \frac{1}{2}(\frac{1}{\text{Re}}Q^n, \nabla w_2)_\mathcal{K} + \frac{1}{2}\langle(\widehat{-\frac{1}{\text{Re}}Q})^n\mathbf{n}, w_2\rangle_{\partial\mathcal{K}} = l_2(w_2), \\
\qquad\qquad\qquad ((Q^n, \mathbf{w}_2))_\mathcal{K} + ((v^n, \nabla\mathbf{w}_2))_\mathcal{K} - \langle\langle\hat{v}^n\mathbf{n}, \mathbf{w}_2\rangle\rangle_{\partial\mathcal{K}} = 0,
\end{cases}
\tag{14}
$$

where $U^n = (u^n, v^n)^\intercal$, $P^n = (p_1^n, p_2^n)^\intercal$, and $Q^n = (q_1^n, q_1^n)^\intercal$, and

$$
\begin{aligned}
l_1(w_1) &= \frac{1}{\Delta t}(u^{n-1}, w_1)_\mathcal{K} + \frac{1}{2}(U^{n-1} \cdot \nabla \mathfrak{u}^{n-1}, w_1)_\mathcal{K} + \frac{1}{2}(\frac{1}{\text{Re}}P^{n-1}, \nabla w_1)_\mathcal{K} \\
&\quad + \frac{1}{2}\langle(\widehat{-\frac{1}{\text{Re}}P})^{n-1}\mathbf{n}, w_1\rangle_{\partial\mathcal{K}}, \\
l_2(w_2) &= \frac{1}{\Delta t}(v^n, w_2)_\mathcal{K} + \frac{1}{2}(U^{n-1} \cdot \nabla \mathfrak{v}^{n-1}, w_2)_\mathcal{K} + \frac{1}{2}(\frac{1}{\text{Re}}Q^{n-1}, \nabla w_2)_\mathcal{K} \\
&\quad + \frac{1}{2}\langle(\widehat{-\frac{1}{\text{Re}}Q})^{n-1}\mathbf{n}, w_2\rangle_{\partial\mathcal{K}}.
\end{aligned}
$$

The superscripts n and $n-1$ stand for the values at the time levels t_n and t_{n-1}, respectively. Likewise, the global equations should be considered at the time level t_n. By summing over all elements, inserting the flux definitions (5) into (6) and (14) at the time level t_n and also using boundary conditions (4), the algebraic system of equations or vector-matrix system can be obtained. The obtained system, steamed by exploiting the Crank–Nicolson method, is nonlinear, and we intend to solve it numerically so that preserves the second-order convergence in the temporal domain. Nevertheless, we exploit the Newton–Raphson method for solving the obtained nonlinear system. We set

$$
W^n = (u^n, v^n, p_1^n, p_2^n, q_1^n, q_2^n, \xi^n, \zeta^n) \in S_{h,k}^6 \times M_{h,k}(0, \Gamma_\mathfrak{u}) \times M_{h,k}(0, \Gamma_\mathfrak{v}),
$$

where $(u^n, v^n, p_1^n, p_2^n, q_1^n, q_2^n, \xi^n, \zeta^n)$ is the exact solution vector of system (14) and (6) at the time level t_n. With a suitable initial guess $W_{n,0}$, we are aiming to generate the following sequence of solution vectors

$$
W_{n,i} = W_{n,i-1} + \delta W_{n,i}, \quad i = 1, 2, \dots,
$$

where $W_{n,i}$ converges to the exact solution, namely, W^n, as i tends to infinity. We note that

$$
\delta W_{n,i} = (\delta u_{n,i}, \delta v_{n,i}, \delta p_{1,n,i}, \delta p_{2,n,i}, \delta q_{1,n,i}, \delta q_{2,n,i}, \delta\xi_{n,i}, \delta\zeta_{n,i}),
$$

is obtained via the Newton–Raphson method. In the other words, $\delta W_{n,i}$ is computed by solving the following linear variational formulation so that holds for all $(w_1, w_2, \mathbf{w}_1, \mathbf{w}_2) \in S_{h,k}^2 \times \mathbf{S}_{h,k}^2$ and $\mathcal{K} \in \mathscr{K}_h$ and $(\mu_1, \mu_2) \in M_{h,k}(0, \Gamma_\mathfrak{u}) \times M_{h,k}(0, \Gamma_\mathfrak{v})$:

$$\tilde{\mathsf{a}}_1(\delta u_{n,i}, w_1) + \tilde{\mathsf{a}}_2(\delta v_{n,i}, w_1) + \tilde{\mathsf{a}}_3(\delta p_{1,n,i}, w_1) + \tilde{\mathsf{a}}_4(\delta p_{2,n,i}, w_1)$$
$$+\tilde{\mathsf{a}}_5(\delta \xi_{n,i}, w_1) = \tilde{\mathsf{l}}_1(w_1),$$
$$\tilde{\mathsf{b}}_1(\delta u_{n,i}, w_{11}) + \tilde{\mathsf{b}}_2(\delta p_{1,n,i}, w_{11}) + \tilde{\mathsf{b}}_3(\delta \xi_{n,i}, w_{11}) = \tilde{\mathsf{l}}_2(w_{11}),$$
$$\tilde{\mathsf{b}}_4(\delta u_{n,i}, w_{12}) + \tilde{\mathsf{b}}_2(\delta p_{2,n,i}, w_{12}) + \tilde{\mathsf{b}}_5(\delta \xi_{n,i}, w_{12}) = \tilde{\mathsf{l}}_3(w_{12}),$$
$$\tilde{\mathsf{c}}_1(\delta u_{n,i}, w_2) + \tilde{\mathsf{c}}_2(\delta v_{n,i}, w_2) + \tilde{\mathsf{c}}_3(\delta q_{1,n,i}, w_2) + \tilde{\mathsf{c}}_4(\delta q_{2,n,i}, w_2)$$
$$+\tilde{\mathsf{c}}_5(\delta \xi_{n,i}, w_2) = \tilde{\mathsf{l}}_4(w_2),$$
$$\tilde{\mathsf{b}}_1(\delta v_{n,i}, w_{21}) + \tilde{\mathsf{b}}_2(\delta q_{1,n,i}, w_{21}) + \tilde{\mathsf{b}}_3(\delta \zeta_{n,i}, w_{21}) = \tilde{\mathsf{l}}_5(w_{11}),$$
$$\tilde{\mathsf{b}}_4(\delta v_{n,i}, w_{22}) + \tilde{\mathsf{b}}_2(\delta q_{2,n,i}, w_{22}) + \tilde{\mathsf{b}}_5(\delta \zeta_{n,i}, w_{22}) = \tilde{\mathsf{l}}_6(w_{12}),$$
$$\tau\tilde{\mathsf{d}}_1(\delta u_{n,i}, \mu_1) + \tilde{\mathsf{d}}_2(\delta p_{1,n,i}, \mu_1) + \tilde{\mathsf{d}}_3(\delta p_{2,n,i}, \mu_1) - \tau\tilde{\mathsf{d}}_4(\delta \xi_{n,i}, \mu_1) = \tilde{\mathsf{l}}_7(\mu_1),$$
$$\sigma\tilde{\mathsf{d}}_1(\delta v_{n,i}, \mu_2) + \tilde{\mathsf{d}}_2(\delta q_{1,n,i}, \mu_2) + \tilde{\mathsf{d}}_3(\delta q_{2,n,i}, \mu_2) - \sigma\tilde{\mathsf{d}}_4(\delta \zeta_{n,i}, \mu_2) = \tilde{\mathsf{l}}_8(\mu_2).$$
$$(15)$$

where $\mathbf{w}_1 = (w_{11}, w_{12})^{\mathsf{T}}$, and $\mathbf{w}_2 = (w_{21}, w_{22})^{\mathsf{T}}$. To observe the definition of multilinear forms and linear functionals in (15), we refer the reader to Appendix of the paper.

In order to solve the large and sparse linear variational formulation (15) more effectively, this system can be decomposed into two linear systems with smaller sizes by using the Schur complement idea. One can observe that (15) can be reformulated to the following vector-matrix equations:

$$\begin{cases} M_{11}X_{n,i} + M_{12}Y_{n,i} = \mathcal{R}_1, \\ M_{21}X_{n,i} + M_{22}Y_{n,i} = \mathcal{R}_2, \end{cases} \tag{16}$$

where $X_{n,i} = [\delta\bar{u}_{n,i} \ \delta\bar{v}_{n,i} \ \delta\bar{p}_{1,n,i} \ \delta\bar{p}_{2,n,i} \ \delta\bar{q}_{1,n,i} \ \delta\bar{q}_{2,n,i}]^{\mathsf{T}}$, $Y_{n,i} = [\delta\bar{\xi}_{n,i} \ \delta\bar{\zeta}_{n,i}]^{\mathsf{T}}$ are coefficients of approximate solutions with

$$M_{11} = \begin{bmatrix} \tilde{A}_1 & \tilde{A}_2^{\mathsf{T}} & \tilde{A}_3 & \tilde{A}_4 & 0 & 0 \\ \tilde{B}_1 & 0 & \tilde{B}_2 & 0 & 0 & 0 \\ \tilde{B}_1 & 0 & 0 & \tilde{B}_2 & 0 & 0 \\ \tilde{C}_1 & \tilde{C}_2 & 0 & 0 & \tilde{C}_3 & \tilde{C}_4 \\ 0 & \tilde{B}_1 & 0 & 0 & \tilde{B}_2 & 0 \\ 0 & \tilde{B}_1 & 0 & 0 & 0 & \tilde{B}_2 \end{bmatrix}, \qquad M_{12} = \begin{bmatrix} \tilde{A}_5 & 0 \\ \tilde{B}_3 & 0 \\ \tilde{B}_4 & 0 \\ 0 & \tilde{C}_5 \\ 0 & \tilde{B}_3 \\ 0 & \tilde{B}_4 \end{bmatrix},$$

$$M_{22} = \begin{bmatrix} -\tau\tilde{D}_4 & 0 \\ 0 & -\sigma\tilde{D}_4 \end{bmatrix}, \qquad M_{21} = \begin{bmatrix} \tau\tilde{D}_1 & 0 & \tilde{D}_2 & \tilde{D}_3 & 0 & 0 \\ 0 & \sigma\tilde{D}_1 & 0 & 0 & \tilde{D}_2 & \tilde{D}_3 \end{bmatrix},$$

$$\mathcal{R}_1 = \begin{bmatrix} \tilde{L}_1 \ \tilde{L}_2 \ \tilde{L}_3 \ \tilde{L}_4 \ \tilde{L}_5 \ \tilde{L}_6 \end{bmatrix}, \qquad \mathcal{R}_2 = \begin{bmatrix} \tilde{L}_7 \ \tilde{L}_8 \end{bmatrix},$$

In the above matrices and vectors, capital letters are interpreted as the matrix and vector representation of multi-linear forms and linear functionals defined in (15). Based on our experiences in the computer implementation of the HDG method, we have not benefited by not encountering non-invertible matrix M_{11}. Regarding this fact, we assume that M_{11} is invertible. Otherwise, it is not possible to propose the reduction of the complexity of the computations for solving (15), and so this system has to be solved directly. Based on the structure of the matrices in vector-matrix equations (16) and

the Schur complement issue, instead of solving (16), the following system of equations are solved in each iteration of the Newton–Raphson method:

$$(M_{22} - M_{21}M_{11}^{-1}M_{12})Y_{n,i} = \mathcal{R}_2 - M_{21}M_{11}^{-1}\mathcal{R}_1. \tag{17}$$

Thus $X_{n,k}$ can be computed by

$$X_{n,i} = M_{11}^{-1}\mathcal{R}_1 - M_{11}^{-1}M_{12}Y_{n,i}. \tag{18}$$

Based on the Newton–Raphson approach, and Schur complement decomposition, we finish this section by representing the details of the designed HDG scheme in the following algorithm.

Algorithm *HDG algorithm for 2D coupled Burgers equations* (1)

Input: Spatial domain Ω and number of elements, namely, N, time interval $[0, T]$ and number of time steps J, degree of approximate polynomials k, boundary data $\Gamma_{\mathfrak{u}}$ and $\Gamma_{\mathfrak{v}}$, initial data, tolerance $0 < \epsilon$, and stabilization parameters τ and σ.

Output: u^J, v^J, p_1^J, p_2^J, q_1^J, q_2^J, ξ^J, and ζ^J that are the approximate solutions of $\mathfrak{u}(x, y, T)$, $\mathfrak{v}(x, y, T)$, $\mathfrak{p}_1(x, y, T)$, $\mathfrak{p}_2(x, y, T)$, $\mathfrak{q}_1(x, y, T)$, $\mathfrak{q}_2(x, y, T)$, $\xi(x, y, T)$ and $\zeta(x, y, T)$.

Generate regular mesh for the domain Ω.

Set W_0 by given initial and boundary conditions.

For $n = 1, 2, \ldots, J$ do

   $W_{n,0} = W_{n-1}$, $\delta W_{n,0} = (\epsilon + 1)\vec{1}$, $i = 0$.

   While $\epsilon < \|\delta W_{n,i}\|$ do

      Compute $\delta W_{n,i+1}$ by Schur complement formulas (17) *and* (18)

      $W_{n,i+1} = W_{n,i} + \delta W_{n,i+1}$, $i = i + 1$

   end While

   $W^n = W_{n,i}$

end For

5 Numerical results

In this section, we aim to demonstrate the efficiency, validation, and applicability of the proposed fully discrete HDG method for system (1). We observe that the semi-discrete HDG method for system (1) is stable over the time interval $[0, t]$, for all $t \in (0, T]$ provided that system (1) is well-posed in

the sense of energy method. To design a fully discrete version of the HDG method, we proposed an approach with the order of at least two for time discretization, that is Crank–Nicolson. Also, the Newton–Raphson method that has the order of at least two is proposed for solving the obtained non-linear system, and therefore, the loss of accuracy will not appear. As seen, to reduce the complexity of the proposed method and the size of the linear system, we exploited the Schur complement idea. Numerical experiments of the proposed HDG method are reported in three examples that they are selected from [35].

In Example 1, the 2D system (1) is considered to investigate the spatial order of accuracy of the proposed HDG method. Also, the results are reported for different Reynolds numbers. In Example 2, the HDG solution is examined for very high Reynolds numbers in the system (1). In Example 3, a 2D coupled Burgers equation without having any exact solution is solved. In this example, the HDG results are compared with the numerical results in [35] and [3].

Example 1. [35] Consider the 2D coupled Burgers equations (1) with $\Omega = (0,1) \times (0,1)$, $T = 1$, and the following exact solutions:

$$\mathfrak{u} = \frac{3}{4} - \frac{1}{4(1 + \exp(\frac{\mathrm{Re}}{4}(-t - 4x + 4y)))},$$

$$\mathfrak{v} = \frac{3}{4} + \frac{1}{4(1 + \exp(\frac{\mathrm{Re}}{4}(-t - 4x + 4y)))}.$$

The initial and boundary conditions can be derived from the exact solutions. In Table 1, L^2 error norms and corresponding orders are reported for $\mathrm{Re} = 1$ and $\tau = \sigma = 0.5$. As seen, satisfactory and high accuracy errors in Table 1 indicate the good performance of our proposed method in solving system (1). Moreover, the results show the optimal convergence for approximate solutions u, v, and their first derivatives. As mentioned earlier, this optimal convergence is inherited from the DG method that is preserved well by our proposed method. In Table 2, the errors are reported for different Reynolds numbers $\mathrm{Re} = 0.1, 1, 10, 100, 200, 500$, approximate polynomials of degree $k = 2$ and $h = 0.2$. For this test, we set $\tau = \sigma = 0.5$ for $\mathrm{Re} = 0.1, 1, 10, 100$ and $\tau = \sigma = 2$ for $\mathrm{Re} = 200, 500$. Note that, by increasing the Reynolds number, the effectiveness of dissipative terms in the system (1) will be eliminated gradually, and so we will face an inviscid system. Therefore, we expect that the accuracy of the method decreases as the Reynolds number increases. According to Table 2, we can observe that the proposed HDG method produces acceptable approximate solutions even for high Reynolds numbers and the reduction of accuracy is acceptable.

Here, we intend to do a test and check the dependence on the accuracy of the numerical solutions on the stability parameters. In Table 3, L^2 error norms and corresponding orders are reported for $\mathrm{Re} = 1$ and $\tau = \sigma = -0.5$.

Table 1: L^2 error norms of approximate solutions u, v, p_1, p_2, q_1, and q_2 together with their corresponding spatial orders of accuracy for Example 1 with Re $= 1$, $\tau = \sigma = 0.5$ at $T = 1$.

k	h	$\|u - \mathfrak{u}\|_\Omega$	order	$\|p_1 - \mathfrak{p}_1\|_\Omega$	order	$\|p_2 - \mathfrak{p}_2\|_\Omega$	order
1	0.4	1.3059 E-6		3.7719 E-6		3.7620 E-6	
	0.2	3.2810 E-7	1.99	9.3005 E-7	2.02	9.3053 E-7	2.02
	0.1	7.6653 E-8	2.10	2.0728 E-7	2.017	2.0778 E-7	2.016
2	0.4	1.2004 E-6		3.8744 E-6		3.8762 E-6	
	0.2	1.5815 E-7	2.92	5.0946 E-7	2.93	5.0959 E-7	2.93
	0.1	1.9734 E-8	3.00	6.3571 E-8	3.00	6.3588 E-8	3.00
k	h	$\|v - \mathfrak{v}\|_\Omega$	order	$\|q_1 - \mathfrak{q}_1\|_\Omega$	order	$\|q_2 - \mathfrak{q}_2\|_\Omega$	order
1	0.4	1.3059 E-6		3.7719 E-6		3.7620 E-6	
	0.2	3.2810 E-7	1.99	9.3005 E-7	2.02	9.3053 E-7	2.02
	0.1	7.6653 E-8	2.10	2.0728 E-7	2.017	2.0778 E-7	2.016
2	0.4	1.2004 E-6		3.8744 E-6		3.8762 E-6	
	0.2	1.5815 E-7	2.92	5.0946 E-7	2.93	5.0959 E-7	2.93
	0.1	1.9734 E-8	3.00	6.3571 E-8	3.00	6.3588 E-8	3.00

Table 2: L^2 error norms for Example 1 with approximate polynomial of degree $k = 2$ and $h = 0.1$ for Re $= 0.1, 1, 10, 100, 250$, and 500, at the final time $T = 1$.

Re	$\|u - \mathfrak{u}\|_\Omega$r	$\|p_1 - \mathfrak{p}_1\|_\Omega$	$\|p_2 - \mathfrak{p}_2\|_\Omega$	$\|v - \mathfrak{v}\|_\Omega$	$\|q_1 - \mathfrak{q}_1\|_\Omega$	$\|q_2 - \mathfrak{q}_2\|_\Omega$
0.1	1.6141 E-11	1.9859 E-10	1.9929 E-10	1.8628 E-11	2.7554 E-10	2.7644 E-10
1	1.2378 E-8	3.9750 E-8	3.9774 E-8	1.2378 E-8	3.9751 E-8	3.9772 E-8
10	6.7030 E-6	2.5580 E-5	2.6527 E-5	6.7030 E-6	2.5580 E-5	2.6527 E-5
100	1.0638 E-3	1.9183 E-2	2.1982 E-2	1.0638 E-3	1.9183 E-2	2.1982 E-2
200	3.3772 E-3	1.2515 E-1	1.3351 E-1	3.3773 E-3	1.2514 E-1	1.3351 E-1
500	2.1209 E-2	7.1691 E-1	6.9343 E-1	2.1210 E-2	7.1691 E-1	6.9343 E-1

We can observe that the HDG method with negative stabilization parameters produces numerical results with high and unacceptable errors.

Example 2. [35] The aim of this example is to investigate the performance of the HDG method in solving system (1) with high Reynolds numbers. Consider system (1) with $\Omega = (0,1) \times (0,1)$, $T = 1$, and exact solutions

$$\mathfrak{u} = -\frac{2\pi \exp(\frac{-5\pi^2 t}{\text{Re}}) \cos(2\pi x) \sin(\pi y)}{\text{Re}(2 + \exp(\frac{-5\pi^2 t}{\text{Re}}) \sin(2\pi x) \sin(\pi y))},$$

$$\mathfrak{v} = -\frac{2\pi \exp(\frac{-5\pi^2 t}{\text{Re}}) \sin(2\pi x) \cos(\pi y)}{\text{Re}(2 + \exp(\frac{-5\pi^2 t}{\text{Re}}) \sin(2\pi x) \sin(\pi y))}.$$

The errors of numerical solutions u and v are shown in Figures 2 and 3, respectively, for $Re = 10000$ and 100000. Note that, the results have been obtained by setting $\tau = \sigma = 20$, $h = 0.1$, and $k = 1$. As mentioned in

Table 3: L^2 error norms of approximate solutions u, v, p_1, p_2, q_1, and q_2 together with their corresponding spatial orders of accuracy for Example 1 with $\mathrm{Re} = 1$, $\tau = \sigma = -0.5$ at $T = 1$.

k	h	$\|u - \mathfrak{u}\|_\Omega$	order	$\|p_1 - \mathfrak{p}_1\|_\Omega$	order	$\|p_2 - \mathfrak{p}_2\|_\Omega$	order
1	0.4	7.2644		2.4545 E+1		2.3701 E+1	
	0.2	6.2532	0.22	1.4973 E+1	0.71	1.4567 E+1	0.70
	0.1	6.1053	0.03	1.4241 E+1	0.07	1.4304 E+1	0.03
2	0.4	2.9309 E+6		1.3503 E+5		1.6600 E+5	
	0.2	8.4685 E+6	-1.53	8.4199 E+5	-2.6	8.0063 E+5	-2.27
	0.1	2.2730 E+5	-1.42	5.4620 E+4	-2.70	4.6931 E+4	-2.55
k	h	$\|v - \mathfrak{v}\|_\Omega$	order	$\|q_1 - \mathfrak{q}_1\|_\Omega$	order	$\|q_2 - \mathfrak{q}_2\|_\Omega$	order
1	0.4	7.3625		2.6346 E+1		2.5501 E+1	
	0.2	6.2372	0.24	1.4836 E+1	0.83	1.4870 E+1	0.78
	0.1	6.0686	0.04	1.4378 E+1	0.05	1.4228 E+1	0.06
2	0.4	1.1331 E+6		5.3889 E+6		6.3966 E+6	
	0.2	5.2046 E+6	-2.20	5.1255 E+5	-3.25	4.9186 E+5	-2.94
	0.1	9.5820 E+6	-0.88	2.3339 E+4	-2.19	1.9896 E+4	-2.20

Example 1, these high Reynolds numbers are going to omit the dissipative terms in system (1), but we can infer from Figures 2 and 3 that the behaviors of approximate solutions still follow the exact solutions very well. This shows the flexibility and superiority of the proposed HDG method for solving different types of system (1) numerically.

Example 3. [3, 35] In this example, a 2D problem with different values of Reynolds numbers will be investigated such that its exact solution is unavailable. Consider the 2D coupled Burgers equations (1) over the domain $\Omega = (0, 0.5) \times (0, 0.5)$ with the initial conditions

$$\mathfrak{u}(x, y, 0) = \sin(\pi x) + \cos(\pi x) \qquad \mathfrak{v}(x, y, 0) = x + y,$$

and the boundary conditions

$$\mathfrak{u}(0, y, t) = \cos(\pi y), \quad \mathfrak{u}(0.5, y, t) = 1 + \cos(\pi y),$$
$$\mathfrak{u}(x, 0, t) = 1 + \sin(\pi x), \quad \mathfrak{u}(x, 0.5, t) = \sin(\pi x),$$
$$\mathfrak{v}(0, y, t) = y, \quad \mathfrak{v}(0.5, y, t) = 0.5 + y, \quad \mathfrak{v}(x, 0, t) = x, \quad \mathfrak{v}(x, 0.5, t) = 0.5 + x.$$

In the proposed HDG scheme, we set $\sigma = \tau = 2$, $h = 0.05$, $\Delta t = 0.001$, and $T = 0.625$. According to this system that has no available exact solution, the only way to understand the correctness of the results is the comparison them with the results of other papers. In the following, the results are compared with the results of [3, 35]. The numerical approximations u and v are illustrated in Figure 4 with $k = 2$ and $\mathrm{Re} = 50$. Also, in Tables 4 and 5, the numerical results are reported at some selected mesh points for $\mathrm{Re} = 50, 500$. We find that, the results of the proposed HDG method are in

Table 4: Comparison of computed values of u and v for Re = 50 for Example 3. Results are reported for approximate polynomials of degree two, $h = 0.05$, and $\Delta t = 0.001$ at final time $T = 0.625$.

	u			v		
(x, y)	HDG method	[35]	[3]	HDG method	[35]	[3]
(0.1, 0.1)	0.96969	0.97146	0.96688	0.09817	0.09869	0.09824
(0.3, 0.1)	1.15072	1.15280	1.14827	0.14167	0.14158	0.14112
(0.2, 0.2)	0.86362	0.86307	0.85911	0.16915	0.16754	0.16681
(0.4, 0.2)	0.99136	0.97981	0.97637	0.18855	0.17109	0.17065
(0.1, 0.3)	0.66440	0.66316	0.66019	0.26491	0.26378	0.26261
(0.3, 0.3)	0.77587	0.77230	0.76932	0.24818	0.22654	0.22576
(0.2, 0.4)	0.59083	0.58180	0.57966	0.33124	0.32851	0.32745
(0.4, 0.4)	0.75273	0.75855	0.75678	0.38614	0.32499	0.32441

Table 5: Comparison of computed values of u and v for Re = 500 for Example 3. Results are reported for approximate polynomials of degree two, $h = 0.05$, and $\Delta t = 0.001$ at final time $T = 0.625$.

	u			v		
(x, y)	HDG method	[35]	[3]	HDG method	[35]	[3]
(0.15, 0.1)	0.96114	0.96151	0.96650	0.08662	0.09230	0.09020
(0.3, 0.1)	0.97324	1.03200	1.02970	0.07841	0.10728	0.10690
(0.1, 0.2)	0.84445	0.87814	0.84449	0.17889	0.16816	0.17972
(0.2, 0.2)	0.86926	1.06370	0.87631	0.16264	0.23690	0.16777
(0.1, 0.3)	0.67883	0.67920	0.67809	0.26177	0.26268	0.26222
(0.3, 0.3)	0.77557	0.79947	0.79792	0.21739	0.23550	0.23497
(0.15, 0.4)	0.54874	0.58959	0.54601	0.31817	0.30419	0.31753
(0.2, 0.4)	0.58850	0.78233	0.58874	0.30049	0.35294	0.30371

good agreements with the presented results in [3, 35]. Hence, the proposed HDG method copes well with equations without the exact solution.

6 Discussion and conclusion

Numerical simulation of the 2D coupled Burgers equations via the HDG method has been studied in this paper, so that this system is equipped with appropriate initial and boundary conditions. In general, HDG methods have less computational time compared to the other DG methods, especially the LDG methods, which are the nearest to the HDG. The main reason for this advantage is the way of defining numerical fluxes. In the HDG method, the definition of numerical fluxes is not unique and depends on the form and physics of the problem. On the other hand, the stability of the method is

Figure 2: The errors of approximate solutions u and v for Example 2 with $Re = 10000$ at $T = 1, 2$. The results are reported for $\tau = \sigma = 20$, approximate polynomial of degree one, and $h = 0.1$.

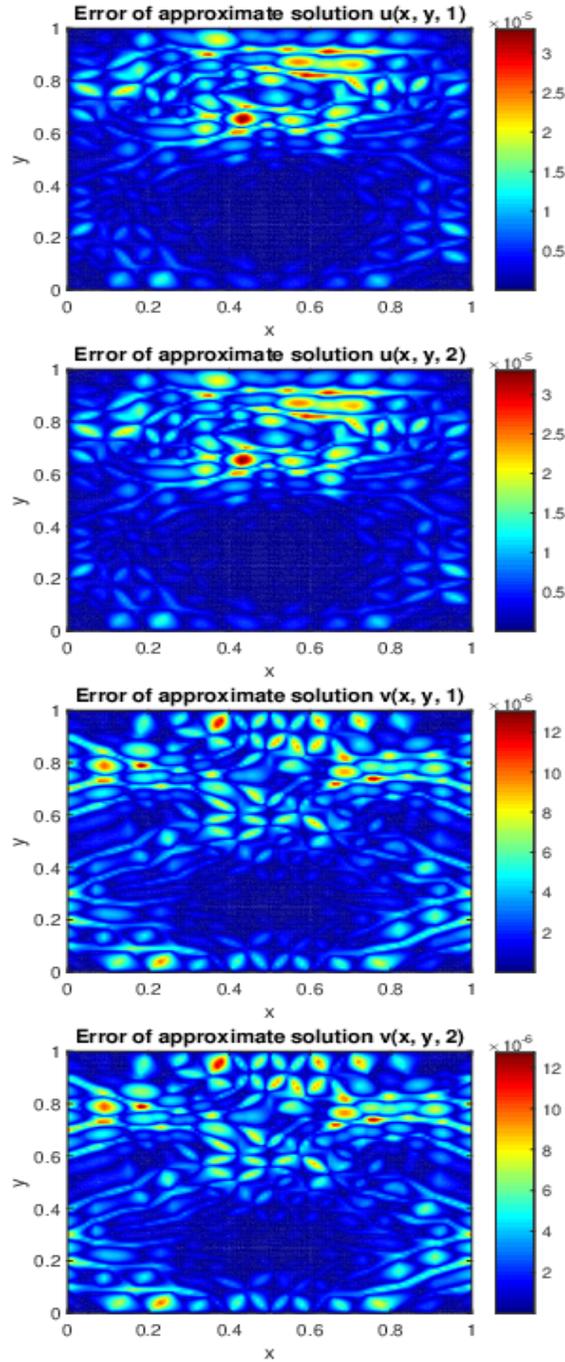Figure 3: The errors of approximate solutions u and v for Example 2 with $Re = 100000$ at $T = 1, 2$. The results are reported for $\tau = \sigma = 20$, approximate polynomial of degree one, and $h = 0.1$.
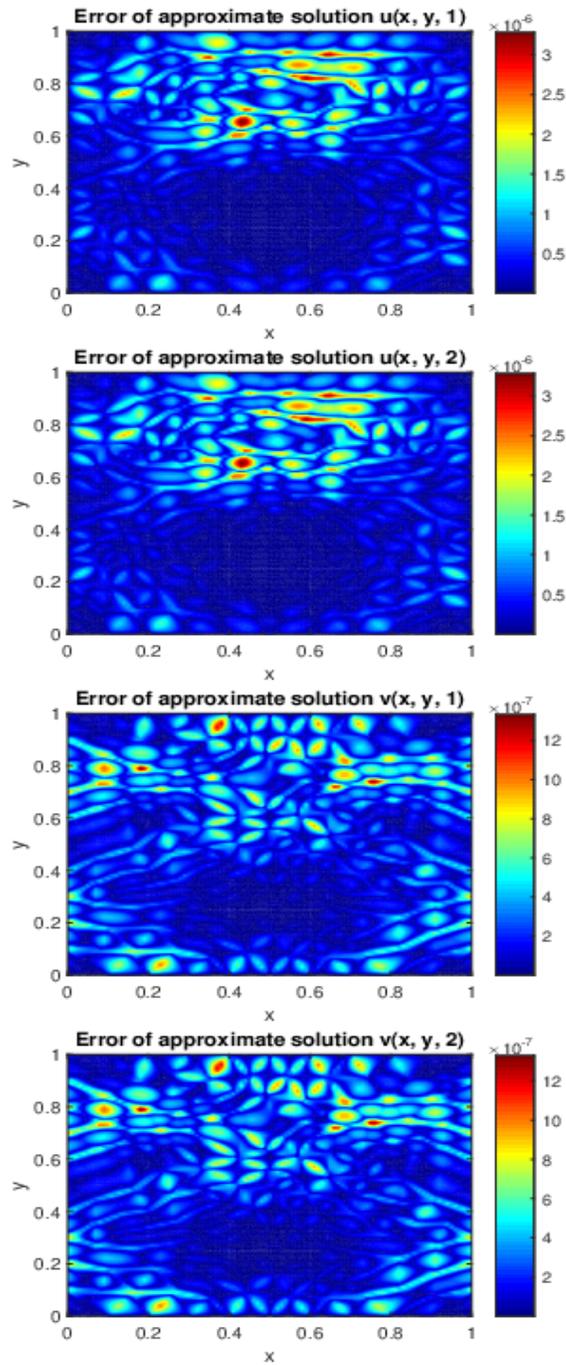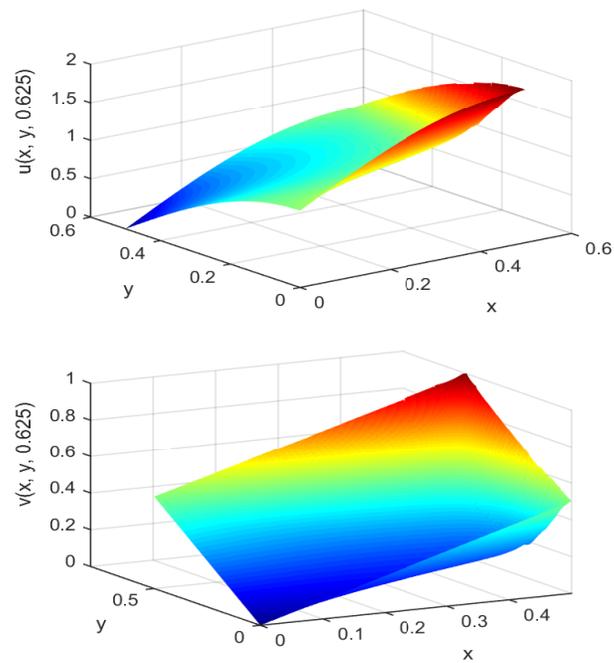
Figure 4: Approximate solutions u and v for Example 3 with $Re = 50$ at $T = 0.625$. The results are reported for $\tau = \sigma = 2$, approximate polynomial of degree two with $h = 0.05$ and $\Delta t = 0.001$.

completely dependent on the fluxes. So, one of the hardships of using the HDG method is finding appropriate definitions of numerical fluxes that guarantee stability. Fortunately, we presented a stable HDG method for solving system (1), while there is no stable (with a proven theorem) LDG method yet. Investigating the convergence of HDG methods for coupled and nonlinear problems is not easy. According to the authors' knowledge, already, the convergence of the HDG methods has been studied just for some simple and linear equations. Convergence of the proposed method can be considered as one of our future works.

The same as other HDG methods, by converting the initial system to a system of first-order equations and defining approximate broken Sobolev spaces associated with spatial partitioning, we set up the semi-discrete variational formulation of the coupled Burgers. Based on the structure of the HDG method, we have proposed numerical traces and fluxes for the variational formulation of the first-order equations. Numerical traces are supposed as global unknowns and depend on Dirichlet boundary data. Defining numerical flux in any HDG method plays a significant role in the stability of the semi-discrete HDG method over a time interval. After introducing appropriate numerical fluxes and imposing sufficient global equations over the spatial partitioning for system (1), the L^2 stability of the proposed semi-discrete HDG methods has been investigated under specific mild conditions on the stabilization parameters that are used in the definitions of numerical fluxes. With the intention of gaining a fully discrete scheme, the Crank–Nicolson method has been applied for time discretization. The choice of the Crank–Nicolson method was because of its unconditional stability and second-order accuracy. To preserve the second order of accuracy in time, the Newton–Raphson method has been nominated for solving the nonlinear system of equations. To solve the large and sparse linear variational systems, which is related to the Newton–Raphson method, the Schur complement idea has been used for reducing computational complexity and designing smaller systems of equations. To explain the details of the HDG method, an algorithm has been prepared. To verify the efficiency of the proposed HDG method, the method was applied to some model problems. In the presented examples, we showed that approximate solutions and their first derivatives of degree k have converged at order $k + 1$, which is an optimal order of convergence. Also, in another example, the ability of the proposed HDG method was checked for solving the 2D coupled Burgers equations with different and high Reynolds numbers. Finally, we tested this method to solve a system without the exact solution, and pleasant results were observed. Regarding the flexibility of the method and numeric experiences, one can infer that the HDG method is one of the outstanding methods that has been exploited for various types of evolution problems in higher dimensions.

Appendix

As mentioned in section 4, a fully discrete approximation method is obtained for solving the nonlinear coupled Burgers equations (1) by using the HDG and Crank–Nicolson methods for spatial and temporal discretization, respectively. Regarding the definitions of approximation spaces, the goal is to find $(n, v^n, p_1^n, p_2^n, q_1^n, q_2^n) \in S_{h,k}^6$ and $(\xi^n, \zeta^n) \in M_{h,k}(0, \Gamma_{\mathfrak{u}}) \times M_{h,k}(0, \Gamma_{\mathfrak{v}})$, such that all equations in (6) and (14) are satisfied for $n = 1, 2, \ldots, J$. As stated in section 4, by summing over all elements, inserting the flux definitions (5) into (6) and (14) at the time level t_n and also using boundary conditions (4), the following system of equations is obtained:

$$
\frac{1}{\Delta t}(u^n, w_1)_{\mathscr{K}_h} + \frac{1}{2}(u^n u_x^n, w_1)_{\mathscr{K}_h} + \frac{1}{2}(v^n u_y^n, w_1)_{\mathscr{K}_h} - \frac{1}{2\mathrm{Re}}(p_{1x}^n, w_1)_{\mathscr{K}_h}
$$

$$
-\frac{1}{2\mathrm{Re}}(p_{2y}^n, w_1)_{\mathscr{K}_h} + \frac{\tau}{2}\langle u^n, w_1\rangle_{\partial\mathscr{K}_h} - \frac{\tau}{2}\langle \xi^n, w_1\rangle_{\partial\mathscr{K}_h \backslash \Gamma_{\mathfrak{u}}} = l_1(w_1),
$$

$$
(p_1^n, w_{11})_{\mathscr{K}_h} + (u^n, (w_{11})_x)_{\mathscr{K}_h} - <\xi^n \mathbf{n}_x, w_{11}>_{\partial\mathscr{K}_h \backslash \Gamma_{\mathfrak{u}}} = l_2(w_{11}),
$$

$$
(p_2^n, w_{12})_{\mathscr{K}_h} + (u^n, (w_{12})_y)_{\mathscr{K}_h} - <\xi^n \mathbf{n}_y, w_{12}>_{\partial\mathscr{K}_h \backslash \Gamma_{\mathfrak{u}}} = l_3(w_{12}),
$$

$$
\frac{1}{\Delta t}(v^n, w_2)_{\mathscr{K}_h} + \frac{1}{2}(u^n v_x^n, w_2)_{\mathscr{K}} + \frac{1}{2}(v^n v_y^n, w_2)_{\mathscr{K}_h} - \frac{1}{2\mathrm{Re}}(q_{1x}^n, w_2)_{\mathscr{K}_h}
$$

$$
-\frac{1}{2\mathrm{Re}}(q_{2y}^n, w_2)_{\mathscr{K}_h} + \frac{\sigma}{2}\langle v^n, w_2\rangle_{\partial\mathscr{K}_h} - \frac{\sigma}{2}\langle \zeta^n, w_2\rangle_{\partial\mathscr{K}_h \backslash \Gamma_{\mathfrak{v}}} = l_4(w_2),
$$

$$
(q_1^n, w_{21})_{\mathscr{K}_h} + (v^n, (w_{21})_x)_{\mathscr{K}_h} - <\zeta^n \mathbf{n}_x, w_{21}>_{\partial\mathscr{K}_h \backslash \Gamma_{\mathfrak{v}}} = l_5(w_{21}),
$$

$$
(q_2^n, w_{22})_{\mathscr{K}_h} + (v^n, (w_{22})_y)_{\mathscr{K}_h} - <\zeta^n \mathbf{n}_y, w_{22}>_{\partial\mathscr{K}_h \backslash \Gamma_{\mathfrak{v}}} = l_6(w_{22}),
$$

$$
\tau\langle u^n, \mu_1\rangle_{\partial\mathscr{K}_h \backslash \mathcal{F}_h^{\partial}} - \tau\langle \xi^n, \mu_1\rangle_{\partial\mathscr{K}_h \backslash \mathcal{F}_h^{\partial}} - \frac{1}{\mathrm{Re}}\langle p_1^n \mathbf{n}_x, \mu_1\rangle_{\partial\mathscr{K}_h}
$$

$$
-\frac{1}{\mathrm{Re}}\langle p_2^n \mathbf{n}_y, \mu_1\rangle_{\partial\mathscr{K}_h} = \quad 0,
$$

$$
\sigma\langle v^n, \mu_2\rangle_{\partial\mathscr{K}_h \backslash \mathcal{F}_h^{\partial}} - \sigma\langle \zeta^n, \mu_2\rangle_{\partial\mathscr{K}_h \backslash \mathcal{F}_h^{\partial}} - \frac{1}{\mathrm{Re}}\langle q_1^n \cdot \mathbf{n}_x, \mu_2\rangle_{\partial\mathscr{K}_h}
$$

$$
-\frac{1}{\mathrm{Re}}\langle q_2^n \cdot \mathbf{n}_y, \mu_2\rangle_{\partial\mathscr{K}_h} = \quad 0,
$$

$$
\tag{19}
$$

where $\mathbf{n} = (\mathbf{n}_x, \mathbf{n}_y)^{\intercal}$, $\mu_1 \in M_{h,k}^{\mathfrak{u}}(0)$, $\mu_2 \in M_{h,k}^{\mathfrak{v}}(0)$, and

$$
l_1(w) = \frac{1}{\Delta t}(u^{n-1}, w)_{\mathscr{K}_h} - \frac{1}{2}(u^{n-1}u_x^{n-1}, w)_{\mathscr{K}_h} - \frac{1}{2}(v^{n-1}u_y^{n-1}, w)_{\mathscr{K}_h}
$$

$$
+\frac{1}{2\mathrm{Re}}(P_{1x}^{n-1}, w_1)_{\mathscr{K}_h} + \frac{1}{2\mathrm{Re}}(P_{2y}^{n-1}, w_1)_{\mathscr{K}_h} - \frac{\tau}{2}\langle u^{n-1}, w\rangle_{\partial\mathscr{K}_h}
$$

$$
+\frac{\tau}{2}\langle \hat{u}^{n-1}, w\rangle_{\partial\mathscr{K}_h} + \frac{\tau}{2}\langle b_{\mathfrak{u}}^n, w\rangle_{\partial\mathscr{K}_h \cap \Gamma_{\mathfrak{u}}}
$$

$$l_4(w) = \frac{1}{\Delta t}(v^{n-1}, w)_{\mathscr{K}_h} - \frac{1}{2}(u^{n-1}v_x^{n-1}, w)_{\mathscr{K}_h} - \frac{1}{2}(v^{n-1}v_y^{n-1}, w)_{\mathscr{K}_h}$$
$$+ \frac{1}{2\mathrm{Re}}(q_{1x}^{n-1}, w)_{\mathscr{K}_h} + \frac{1}{2\mathrm{Re}}(q_{2y}^{n-1}, w)_{\mathscr{K}_h} - \frac{\sigma}{2}\langle v^{n-1}, w\rangle_{\partial\mathscr{K}_h}$$
$$+ \frac{\sigma}{2}\langle \hat{v}^{n-1}, w_1\rangle_{\partial\mathscr{K}_h} + \frac{\sigma}{2}\langle b_{\mathfrak{v}}^n, w\rangle_{\partial\mathscr{K}_h \cap \Gamma_{\mathfrak{v}}}$$

$$l_2(w) = \langle b_{\mathfrak{u}}^n \mathbf{n}_x, w\rangle_{\partial\mathscr{K}_h \cap \Gamma_{\mathfrak{u}}}, \qquad l_3(w) = \langle b_{\mathfrak{u}}^n \mathbf{n}_y, w\rangle_{\partial\mathscr{K}_h \cap \Gamma_{\mathfrak{u}}},$$

$$l_5(w) = \langle b_{\mathfrak{v}}^n \mathbf{n}_x, w\rangle_{\partial\mathscr{K}_h \cap \Gamma_{\mathfrak{v}}}, l_6(w) = \langle b_{\mathfrak{v}}^n \mathbf{n}_y, w\rangle_{\partial\mathscr{K}_h \cap \Gamma_{\mathfrak{v}}}.$$

Regarding to the nonlinear weak formulation (19), it is needed to convert this weak form to a linear variational form by a suitable iterative method. As explained in section 4, by applying the Newton–Raphson method to the nonlinear variational formulation (19), we intend to find

$$\delta W_{n,i} = (\delta u_{n,i}, \delta v_{n,i}, \delta p_{1,n,i}, \delta p_{2,n,i}, \delta q_{1,n,i}, \delta q_{2,n,i}, \delta \xi_{n,i}, \delta \zeta_{n,i}),$$

such that for all $(w_1, w_2, w_{11}, w_{12}, w_{21}, w_{22}) \in S_{h,k}^6$ and $(\mu_1, \mu_2) \in M_{h,k}(0, \Gamma_{\mathfrak{u}}) \times M_{h,k}(0, \Gamma_{\mathfrak{v}})$, the bilinear form system (15) holds. We finish the Appendix by defining all multilinear forms and linear functionals, which are considered in (15),

$$\tilde{\mathsf{a}}_1(\delta u_{n,i}, w) = \frac{1}{\Delta t}(\delta u_{n,i}, w)_{\mathscr{K}_h} + \frac{1}{2}((u_x)_{n,i-1}\delta u_{n,i}, w)_{\mathscr{K}_h}$$
$$+ \frac{1}{2}(u_{n,i-1}(\delta u_x)_{n,i}, w)_{\mathscr{K}_h} + \frac{\tau}{2}\langle \delta u_{n,i}, w\rangle_{\partial\mathscr{K}_h}$$
$$+ \frac{1}{2}(v_{n,i-1}\delta(u_y)_{n,i}, w)_{\mathscr{K}_h},$$

$$\tilde{\mathsf{a}}_2(\delta v_{n,i}, w) = \frac{1}{2}((u_y)_{n,i-1}\delta v_{n,i}, w)_{\mathscr{K}_h},$$

$$\tilde{\mathsf{a}}_3(\delta p_{1,n,i}, w) = -\frac{1}{2\mathrm{Re}}((\delta p_{1,n,i})_x, w)_{\mathscr{K}_h},$$

$$\tilde{\mathsf{a}}_4(\delta p_{2,n,i}, w) = -\frac{1}{2\mathrm{Re}}((\delta p_{2,n,i})_x, w)_{\mathscr{K}_h},$$

$$\tilde{\mathsf{a}}_5(\delta \xi_{n,i}, w) = -\frac{\tau}{2}\langle \delta \xi_{n,i}, w\rangle_{\partial\mathscr{K}_h \setminus \Gamma_{\mathfrak{u}}},$$

$$\tilde{\mathsf{b}}_1(\delta u_{n,i}, w) = (\delta u_{n,i}, w_x)_{\mathscr{K}_h},$$

$$\tilde{\mathsf{b}}_2(\delta p_{1,n,i}, w) = (\delta p_{1,n,i}, w)_{\mathscr{K}_h},$$

$$\tilde{\mathsf{b}}_4(\delta u_{n,i}, w) = (\delta u_{n,i}, \mathbf{w}_y)_{\mathscr{K}_h},$$

$$\tilde{\mathsf{b}}_3(\delta \xi_{n,i}\mathbf{n}, w) = -\langle \delta \xi_{n,i}\mathbf{n}_x, w\rangle_{\partial\mathscr{K}_h \setminus \Gamma_{\mathfrak{u}}},$$

$$\tilde{\mathsf{b}}_5(\delta \xi_{n,i}, w) = -\langle \delta \xi_{n,i}\mathbf{n}_y, w\rangle_{\partial\mathscr{K}_h \setminus \Gamma_{\mathfrak{u}}},$$

$$\tilde{\mathsf{c}}_2(\delta v_{n,i}, w) = \frac{1}{\Delta t}(\delta v_{n,i}, w)_{\mathscr{K}_h} + \frac{1}{2}((v_y)_{n,i-1}\delta v_{n,i}, w)_{\mathscr{K}_h}$$

$$+ \frac{1}{2}(u_{n,i-1}\delta(v_x)_{n,i},w)_{\mathscr{K}_h} + \frac{\sigma}{2}\langle\delta v_{n,i},w\rangle_{\partial\mathscr{K}_h}$$

$$+ \frac{1}{2}(v_{n-1,i}(\delta v_y)_{n,i},w)_{\mathscr{K}_h},$$

$$\tilde{\mathsf{c}}_1(\delta u_{n,i},w) = \frac{1}{2}((v_x)_{n,i-1}\delta u_{n,i},w)_{\mathscr{K}_h},$$

$$\tilde{\mathsf{c}}_3(\delta q_{1,n,i},w) = -\frac{1}{2\mathrm{Re}}((\delta p_{1,n,i})_x,w)_{\mathscr{K}_h},$$

$$\tilde{\mathsf{c}}_4(\delta q_{2,n,i},w) = -\frac{1}{2\mathrm{Re}}((\delta q_{2,n,i})_x,w)_{\mathscr{K}_h},$$

$$\tilde{\mathsf{c}}_5(\delta\zeta_{n,i},w) = -\frac{\sigma}{2}\langle\delta\zeta_{n,i},w\rangle_{\partial\mathscr{K}_h\setminus\Gamma_{\mathfrak{v}}}$$

$$\tilde{\mathsf{d}}_1(\delta u_{n,i},\mu) = \langle\delta u_{n,i},\mu\rangle_{\partial\mathscr{K}_h\setminus\mathcal{F}_h^\partial},$$

$$\tilde{\mathsf{d}}_2(\delta p_{1,n,i},\mu) = -\frac{1}{\mathrm{Re}}\langle\delta p_{1,n,i}\mathbf{n}_x,\mu\rangle_{\partial\mathscr{K}_h},$$

$$\tilde{\mathsf{d}}_3(\delta p_{2,n,i},\mu) = -\frac{1}{\mathrm{Re}}\langle\delta p_{2,n,i}\mathbf{n}_y,\mu\rangle_{\partial\mathscr{K}_h},$$

$$\tilde{\mathsf{d}}_4(\delta\xi_{n,i},\mu) = \langle\delta\xi_{n,i},\mu\rangle_{\partial\mathscr{K}_h\setminus\mathcal{F}_h^\partial},$$

$$\tilde{\mathsf{l}}_1(w) = \frac{1}{\Delta t}(u^{n-1},w)_{\mathscr{K}_h} - \frac{1}{2}(u^{n-1}u_x^{n-1},w)_{\mathscr{K}_h} - \frac{1}{2}(v^{n-1}u_y^{n-1},w)_{\mathscr{K}_h}$$

$$+ \frac{1}{2\mathrm{Re}}(P_{1x}^{n-1},w_1)_{\mathscr{K}_h} + \frac{1}{2\mathrm{Re}}(P_{2y}^{n-1},w_1)_{\mathscr{K}_h} - \frac{\tau}{2}\langle u^{n-1},w\rangle_{\partial\mathscr{K}_h}$$

$$+ \frac{\tau}{2}\langle\hat{u}^{n-1},w\rangle_{\partial\mathscr{K}_h} + \frac{\tau}{2}\langle b_{\mathfrak{u}}^n,w\rangle_{\partial\mathscr{K}_h\cap\Gamma_{\mathfrak{u}}} - \frac{1}{\Delta t}(u_{n,i-1},w)_{\mathscr{K}_h}$$

$$- \frac{1}{2}(u_{n,i-1}p_{1,n,i-1},w)_{\mathscr{K}_h} - \frac{1}{2}(v_{n,i-1}p_{2,n,i-1},w)_{\mathscr{K}_h}$$

$$+ \frac{1}{2\mathrm{Re}}((p_{1,n,i-1})_x,w)_{\mathscr{K}_h} + \frac{1}{2\mathrm{Re}}((p_{2,n,i-1})_y,w)_{\mathscr{K}_h}$$

$$- \frac{\tau}{2}\langle u_{n,i-1},w\rangle_{\partial\mathscr{K}_h} + \frac{\tau}{2}\langle\xi_{n,i-1},w\rangle_{\partial\mathscr{K}_h\setminus\Gamma_{\mathfrak{u}}},$$

$$\tilde{\mathsf{l}}_2(w) = \langle b_{\mathfrak{u}}^n\mathbf{n}_x,w\rangle_{\partial\mathscr{K}_h\cap\Gamma_{\mathfrak{u}}} - (p_{1,n,i-1},w)_{\mathscr{K}_h} - (u_{n,i-1},(w)_x)_{\mathscr{K}_h}$$

$$+ <\xi_{n,i-1}\mathbf{n}_x,w>_{\partial\mathscr{K}_h\setminus\Gamma_{\mathfrak{u}}},$$

$$\tilde{\mathsf{l}}_3(w) = \langle b_{\mathfrak{u}}^n\mathbf{n}_y,w\rangle_{\partial\mathscr{K}_h\cap\Gamma_{\mathfrak{u}}} - (p_{2,n,i-1},w)_{\mathscr{K}_h} - (u_{n,i-1},(w)_y)_{\mathscr{K}_h}$$

$$+ <\xi_{n,i-1}\mathbf{n}_y,w>_{\partial\mathscr{K}_h\setminus\Gamma_{\mathfrak{u}}},$$

$$\tilde{\mathsf{l}}_4(w) = \frac{1}{\Delta t}(v^{n-1},w)_{\mathscr{K}_h} - \frac{1}{2}(u^{n-1}v_x^{n-1},w)_{\mathscr{K}_h} - \frac{1}{2}(v^{n-1}v_y^{n-1},w)_{\mathscr{K}_h}$$

$$+ \frac{1}{2\mathrm{Re}}(q_{1x}^{n-1},w_1)_{\mathscr{K}_h} + \frac{1}{2\mathrm{Re}}(q_{2y}^{n-1},w_1)_{\mathscr{K}_h} - \frac{\sigma}{2}\langle q^{n-1},w\rangle_{\partial\mathscr{K}_h}$$

$$+ \frac{\sigma}{2}\langle\hat{v}^{n-1},w\rangle_{\partial\mathscr{K}_h} + \frac{\sigma}{2}\langle b_{\mathfrak{v}}^n,w\rangle_{\partial\mathscr{K}_h\cap\Gamma_{\mathfrak{v}}} - \frac{1}{\Delta t}(v_{n,i-1},w)_{\mathscr{K}_h}$$

$$- \frac{1}{2}(u_{n,i-1}q_{1,n,i-1},w)_{\mathscr{K}_h} - \frac{1}{2}(v_{n,i-1}q_{2,n,i-1},w)_{\mathscr{K}_h}$$

$$+ \frac{1}{2\mathrm{Re}}((q_{1,n,i-1})_x, w)_{\mathscr{K}_h} + \frac{1}{2\mathrm{Re}}((q_{2,n,i-1})_y, w)_{\mathscr{K}_h}$$
$$- \frac{\sigma}{2}\langle v_{n,i-1}, w\rangle_{\partial\mathscr{K}_h} + \frac{\sigma}{2}\langle \zeta_{n,i-1}, w\rangle_{\partial\mathscr{K}_h\backslash\Gamma_{\mathfrak{v}}},$$
$$\tilde{l}_5(w) = \langle b_{\mathfrak{v}}^n \mathbf{n}_x, w\rangle_{\partial\mathscr{K}_h\cap\Gamma_{\mathfrak{v}}} - (q_{1,n,i-1}, w)_{\mathscr{K}_h} - (v_{n,i-1}, (w)_x)_{\mathscr{K}_h}$$
$$+ <\zeta_{n,i-1}\mathbf{n}_x, w>_{\partial\mathscr{K}_h\backslash\Gamma_{\mathfrak{v}}},$$
$$\tilde{l}_6(w) = \langle b_{\mathfrak{v}}^n \mathbf{n}_y, w\rangle_{\partial\mathscr{K}_h\cap\Gamma_{\mathfrak{v}}} - (q_{2,n,i-1}, w)_{\mathscr{K}_h} - (v_{n,i-1}, (w)_y)_{\mathscr{K}_h}$$
$$+ <\zeta_{n,i-1}\mathbf{n}_y, w>_{\partial\mathscr{K}_h\backslash\Gamma_{\mathfrak{v}}},$$
$$\tilde{l}_7(\mu) = -\tau\langle u_{n,i-1}, \mu\rangle_{\partial\mathscr{K}_h\backslash\mathcal{F}_h^\partial} + \tau\langle \xi_{n,i-1}, \mu\rangle_{\partial\mathscr{K}_h\backslash\mathcal{F}_h^\partial}$$
$$+ \frac{1}{\mathrm{Re}}\langle p_{1,n,i-1}\mathbf{n}_x, \mu\rangle_{\partial\mathscr{K}_h} + \frac{1}{\mathrm{Re}}\langle p_{2,n,i-1}\mathbf{n}_y, \mu\rangle_{\partial\mathscr{K}_h},$$
$$\tilde{l}_8(\mu) = -\sigma\langle v_{n,i-1}, \mu\rangle_{\partial\mathscr{K}_h\backslash\mathcal{F}_h^\partial} + \sigma\langle \zeta_{n,i-1}, \mu\rangle_{\partial\mathscr{K}_h\backslash\mathcal{F}_h^\partial}$$
$$+ \frac{1}{\mathrm{Re}}\langle q_{1,n,i-1}\mathbf{n}_x, \mu\rangle_{\partial\mathscr{K}_h} + \frac{1}{\mathrm{Re}}\langle q_{2,n,i-1}\mathbf{n}_y, \mu\rangle_{\partial\mathscr{K}_h}.$$

References

[1] Adomian, G. *The diffusion-Brusselator equation*, Comput. Math. with Appl. **29(5)** (1995), 1–3.

[2] Aliyi Koroche, K. and Muleta Chemeda, H. *Sixth-order compact finite difference method for solving KDV-Burger equation in the application of wave propagations*, Iran. J. Numer. Anal. Optim. 12(2) (2022), 277–300.

[3] Bahadır, A.R. *A fully implicit finite-difference scheme for two-dimensional Burgers' equations*, Appl. Math. Comput. 137(1) (2003), 131–137.

[4] Baharloui, S., Mokhtari, R. and Chegini, N. *A stable numerical scheme based on the hybridized discontinuous Galerkin method for the Ito-type coupled KdV system*, Commun. Appl. Math. Comput. 4(4) (2022), 1351–1373.

[5] Burgers, J.M. *A mathematical model illustrating the theory of turbulence*, Adv. Appl. Mech. (1948), 171–199.

[6] Castillo, P. *A review of the local discontinuous Galerkin (LDG) method applied to elliptic problems*, Appl. Numer. Math. 56(10-11) (2006), 1307–1313.

[7] Castillo, P. and Gómez, S. *Conservative super-convergent and hybrid discontinuous Galerkin methods applied to nonlinear Schrödinger equations*, Appl. Math. Comput. 371 (2020), 124950.

[8] Chand, A. and Saha Ray, S. *Numerical simulation of Allen–Cahn equation with nonperiodic boundary conditions by the local discontinuous Galerkin method*, Int. J. Mod. Phys. B. 37(02) (2023), 2350019.

[9] Cockburn, B., Fu, G. and Qiu, W. *A note on the devising of superconvergent HDG methods for Stokes flow by M-decompositions*, IMA J. Numer. Anal. 37(2) (2017), 730–749.

[10] Cockburn, B., Hou, S. and Shu, C.W. *The Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws. IV. The multidimensional case*, Math. Comput. 54(190) (1990), 545–581.

[11] Cockburn, B. and Gopalakrishnan, J. *A characterization of hybridized mixed methods for second order elliptic problems*, SIAM J. Numer. Anal. 42(1) (2004), 283–301.

[12] Cockburn, B., Gopalakrishnan, J. and Lazarov, R. *Unified hybridization of discontinuous Galerkin, mixed, and continuous Galerkin methods for second order elliptic problems*, SIAM J. Numer. Anal. 47(2) (2009), 1319–1365.

[13] Cockburn, B., Gopalakrishnan, J. and Sayas, F.J. *A projection-based error analysis of HDG methods*, Math. Comput. 79(271) (2010), 1351–1367.

[14] Cockburn, B., Guzmán, J., Soon, S.C. and Stolarski, H.K. *An analysis of the embedded discontinuous Galerkin method for second-order elliptic problems*, SIAM J. Numer. Anal. 47(4) (2009), 2686–2707.

[15] Cockburn, B. and Shu, C.W. *The local discontinuous Galerkin method for time-dependent convection-diffusion systems*, SIAM J. Numer. Anal. 35(6) (1998), 2440–2463.

[16] Cockburn, B. and Shu, C.W. *The Runge-Kutta discontinuous Galerkin method for conservation laws V: multidimensional systems*, J. Comput. Phys. 141(2) (1998), 199–224.

[17] Cole, J.D. *On a quasi-linear parabolic equation occurring in aerodynamics*, Q. Appl. Math. 9(3) (1951), 225-236.

[18] Debnath, L. *Nonlinear partial differential equations for scientists and engineers*, Third edition. Birkhäuser/Springer, New York, 2012.

[19] El-Sayed, S.M. and Kaya, D. *On the numerical solution of the system of two-dimensional Burgers' equations by the decomposition method*, Appl. Math. Comput. 158(1) (2004), 101–109.

[20] Fletcher, C.A. *Generating exact solutions of the two-dimensional Burgers' equations*, Int. J. Numer. Methods Fluids. 3 (1983), 213–216.

[21] Giacomini, M., Karkoulias, A., Sevilla, R. and Huerta, A. *A superconvergent HDG method for Stokes flow with strongly enforced symmetry of the stress tensor*, J. Sci. Comput. 77(3) (2018), 1679–1702.

[22] Giacomini, M., Sevilla, R. and Huerta, A. *Tutorial on Hybridizable Discontinuous Galerkin (HDG) formulation for incompressible flow problems*, Modeling in Engineering Using Innovative Numerical Methods for Solids and Fluids, (2020), 163–201.

[23] Khater, A.H., Temsah, R.S. and Hassan, M. *A Chebyshev spectral collocation method for solving Burgers'-type equations*, J. Comput. Appl. Math. 222(2) (2008), 333–350.

[24] Kumar, V., Singh, S. and Koksal, M.E. *A composite algorithm for numerical solutions of two-dimensional coupled Burgers equations*, J. Math. (2021), 1–13.

[25] Ling, D., Shu, C.W. and Yan, W. *Local discontinuous Galerkin methods for diffusive-viscous wave equations*, J. Comput. Appl. Math. 419 (2023), 114690.

[26] Logan, J.D. *An introduction to nonlinear partial differential equations*, John Wiley and Sons (2008).

[27] Mokhtari, R. and Mohseni, M. *A meshless method for solving mKdV equation*, Comput. Phys. Commun. 183(6) (2012), 1259–1268.

[28] Mokhtari, R., Toodar, A.S. and Chegini, N.G. *Application of the generalized differential quadrature method in solving Burgers' equations*, Commun. Theor. Phys. 56(6) (2011), 1009.

[29] Nguyen, N.C. and Peraire, J. *Hybridizable discontinuous Galerkin methods for partial differential equations in continuum mechanics*, J. Comput. Phys. 231(18) (2012), 5955–5988.

[30] Peraire, J., Nguyen, N. and Cockburn, B. *A hybridizable discontinuous Galerkin method for the compressible Euler and Navier-Stokes equations*, In 48th AIAA aerospace sciences meeting including the new horizons forum and aerospace exposition, 363 (2010).

[31] Rong-Pei, Z., Xi-Jun, Y. and Guo-Zhong, Z. *Local discontinuous Galerkin method for solving Burgers and coupled Burgers equations*, Chin. Phys. B. 20(11) (2011), 110205.

[32] Sevilla, R., Giacomini, M., Karkoulias, A. and Huerta, A. *A superconvergent hybridisable discontinuous Galerkin method for linear elasticity*, Int. J. Numer. Methods Eng. 116(2) (2018), 91–116.

[33] Sevilla, R. and Huerta, A. *Tutorial on hybridizable discontinuous Galerkin (HDG) for second-order elliptic problems*, Advanced Finite Element Technologies, (2016), 105–129.

[34] Shin, D., Jeon, Y. and Park, E.J. *Analysis of hybrid discontinuous Galerkin methods for linearized Navier–Stokes equations*, Numer. Methods Partial Differ. Equ. 39(1) (2023), 304–328.

[35] Srivastava, V.K. and Tamsir, M. *Crank-Nicolson semi implicit approach for numerical solutions of two-dimensional coupled nonlinear Burgers' equations*, Int. J. Appl.Mech. 17(2) (2012), 571.

[36] Vila-Pérez, J., Giacomini, M., Sevilla, R. and Huerta, A. *Hybridisable discontinuous Galerkin formulation of compressible flows*, Arch. Comput. Methods Eng. 28(2) (2021), 753–784.

[37] Williams, D. *An entropy stable, hybridizable discontinuous Galerkin method for the compressible Navier-Stokes equations*, Math. Comput. 87(309) (2018), 95–121.

[38] Yang, X., Ge, Y. and Lan, B. *A class of compact finite difference schemes for solving the 2D and 3D Burgers' equations*, Math. Comput. Simul. 185 (2021), 510–534.

[39] Zogheib, B., Tohidi, E., Baskonus, H.M. and Cattani, C. *Method of lines for multi-dimensional coupled viscous Burgers' equations via nodal Jacobi spectral collocation method*, Phys. Scr. 96(12) (2021), 124011.

How to cite this article

Evaluation of iterative methods for solving nonlinear scalar equations

M. Rezaiee-Pajand*, , A. Arabshahi and N. Gharaei-Moghaddam

Abstract

This study is aimed at performing a comprehensive numerical evaluation of the iterative solution techniques without memory for solving nonlinear scalar equations with simple real roots, in order to specify the most efficient and applicable methods for practical purposes. In this regard, the capabilities of the methods for applicable purposes are be evaluated, in which the ability of the methods to solve different types of nonlinear equations is be studied. First, 26 different iterative methods with the best performance are reviewed. These methods are selected based on performing more than 46000 analyses on 166 different available nonlinear solvers. For the easier application of the techniques, consistent mathematical notation is employed to present reviewed approaches. After presenting the diverse methodologies suggested for solving nonlinear equations, the performances of the reviewed methods are evaluated by solving 28 different nonlinear equations. The utilized test functions, which are selected from the reviewed research works, are solved by all schemes and by assuming different initial guesses. To select the initial guesses, endpoints of five neighboring intervals with different sizes around the root of test functions are used. Therefore, each problem is solved by ten different starting points. In order to calculate novel computational efficiency indices and rank them accurately, the results of the obtained solutions are used. These data include

Mohammad Rezaiee-Pajand
Professor of Civil Engineering, School of Engineering, Ferdowsi University of Mashhad.
e-mail: Rezaiee@um.ac.ir

Alireza Arabshahi
Ph.D. Student of Structural Engineering, School of Engineering, Ferdowsi University of Mashhad. e-mail: Alireza.Arabshahi68@gmail.com

Nima Gharaei-Moghaddam
Ph.D. of Structural Engineering, School of Engineering, Ferdowsi University of Mashhad.
e-mail: Nima.Gharaeimoghaddam@mail.um.st.ir

the number of iterations, number of function evaluations, and convergence times. In addition, the successful runs for each process are used to rank the evaluated schemes. Although, in general, the choice of the method depends on the problem in practice, but in practical applications, especially in engineering, changing the solution method for different problems is not feasible all the time, and accordingly, the findings of the present study can be used as a guide to specify the fastest and most appropriate solution technique for solving nonlinear problems.

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

Keywords: Nonlinear scalar equations; Iterative method; Efficiency index; Order of convergence; Initial guess; Function evaluation.

1 Introduction

Most of the practical problems in engineering and other fields of science can be modeled by mathematical functions, which are mostly nonlinear. For instance, in engineering applications, nonlinear structural analysis, or computation of three-dimensional stresses require to solve nonlinear equations. Another example of the case in civil engineering practice that requires solving a nonlinear scalar equation is the computation of the torsional-flexural buckling load of steel columns. Similarly, the final step in the mathematical modeling and formulation of many other fields of science is to solve a nonlinear equation. Therefore, a reliable and applicable method for solving nonlinear equations is a necessary tool for scientific research. This device is utilized in performing different science-based activities, such as analysis and design. This need was felt many years ago and consequently, various solution methods are proposed for solving nonlinear equations. From the early works in this field until now, many different schemes are proposed. Some of these techniques are analytical approaches that are limited to special cases of nonlinear equations, but most of them are numerical iterative schemes. An iterative solution technique, as its name indicates, computes the root of a nonlinear function through several iteration cycles by an initial guess. Most of these methods are modifications of the basic earlier techniques, like the Newton method.

The different iterative approach has different convergence order. Order of convergence is an important mathematical quantity that indicates the efficiency of the solver. However, despite this mathematical standpoint, from the practical view, a method with a higher order of convergence is not necessarily the best choice, and the performance of a solver depends on many different factors. On the other hand, the large number of existing iterative methods makes it more difficult to choose a suitable technique for a special applicable problem. Therefore, the main motivation of this study is to provide a clear understanding of the performance of many of these iterative schemes. For

this purpose, it is attempted to review many of the basic and well-known as well as newly proposed solution approaches in the first part of this study. Some big questions arise when facing this large number of iterative nonlinear solvers: Which method should be used for solving a given problem? Which approach is the fastest? Which one requires the least computational effort? Do they necessarily converge to the desired response?

The answer to these questions is not simple and certain and depends on many factors, including the problem at hand, the utilized initial guess, number of floating-point arithmetic, and the termination criterion. However, it is assured that no method can solve all the possible problems. The investigators who proposed the iterative solution schemes performed convergence analysis to demonstrate the ability of their methods to find the root of nonlinear functions. Therefore, they proposed a convergence order that, from the mathematical point of view, is an indication of the solution speed. In general, a solution method with a higher convergence order should converge faster to the response. However, in practice, the situation is not as easy as it seems. There is no guarantee that a certain solver can find the roots of a given nonlinear problem. Moreover, it is widely known that higher-order solvers converge faster when the initial guess is close enough to the root. In other words, increasing the order of convergence results in a smaller region of attraction for a certain number of iterations. Therefore, in practical problems where the initial guess may fall in a wide range around the response, a higher-order method is not necessarily superior. In cases, when the selected starting point is far from the root of the function, the higher order of convergence may lead to the inability of the method to find the response within a permissible number of iterations. Even in some cases, the method may diverge.

To the authors' best knowledge, despite a large number of available iterative methods for solving nonlinear scalar equations, there are very limited reviews about these techniques. One of the limited reviews in this field is performed by Babajee and Dauhoo [1]. They investigated the performance of the variants of the Newton method with cubic convergence. They also extended some of these methods to multivariate cases. In another similar study, Varona [27] performed a numerical and graphical comparison between some of the well-known solution methods. However, Varona utilized many different criteria for the evaluation of the solution methods, but his research work is mostly limited to traditional and well-known techniques, and their performance is evaluated by extensive applications during the past decades. Two more recent valuable review studies have also been performed by Cătinaş [2, 3]. Occasionally various researchers propose the same methods independently. This is due to the fact that there are so many iterative techniques available, and this quantity increases very fast every year. Therefore, there is a great need for studies like the present paper to provide useful information for the researchers in this regard to prevent the proposition of the same formulations by different investigators. Another merit of the present research

work is to study the effect of the initial guess on the performance of the methods and also evaluate the practical efficiency of different approaches.

The question of which solver is better remained unanswered. Due to the variable nature of different nonlinear problems in the various fields of science and application, giving a definite answer to this question is impossible. This study, it is comprehensively tried to provide a clearer image of the performance of a large number of iterative solution methods without memory. For this purpose, 26 different solution techniques (selected as the best-performing methods among 166 reviewed solvers which are not reported in this manuscript) are used to solve 28 different nonlinear functions by using ten diverse initial guesses for each function. Different initial guesses are used, to investigate their effect of them on the performance of the solution methods. It is worth mentioning that this important effect has been neglected in much of the previous research in this field. To compare the abilities of discussed approaches, a new computational efficiency index is proposed and utilized against the others which were used previously. All solvers are ranked based on the results of presented extensive numerical evaluations. The suggested index has a qualitative-quantitative base and can successfully rank the solution schemes. To indicate the most applicable solver, the results of the new index are compared with those attained by the traditional and well-known efficiency indices. Findings show that the suggested way can better distinguish the performance and efficiency of the nonlinear solvers for practical applications. Finally, according to the obtained results, the reviewed methods are ranked to specify the ones which are more efficient and applicable to be utilized, especially, in engineering practice.

2 Review of the available iterative solvers

In this section, various nonlinear solvers are reviewed briefly and presented in historical order. These methods fall in the category of iterative methods without memory; that is, only the results of the current iteration would be used to determine the next estimation. The iterative formula of each technique is provided for the $n + 1$th estimation of the root, assuming that the nth evaluation is available. The process commences by using an initial guess, x_0. Here, $f(x)$ indicate the nonlinear function needs to be solved. The reviewed methods are presented in Table 1, using uniform mathematical notations.

Table 1: Iterative solution methods

No	Method	Iterative Formula
1	Newton	$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$
2	Ostrowski [22]	$x_{n+1} = y_n - \frac{f(x_n)}{f(x_n)-2f(y_n)}\frac{f(y_n)}{f'(x_n)}$
3	Traub–Ostrowski [26]	$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}\left(\frac{f(x_n)-f(y_n)}{f(x_n)-2f(y_n)}\right)$
4	Jarrat relation [12]	$x_{n+1} = x_n - \frac{1}{2}\frac{f(x_n)}{f'(x_n)} + \frac{f(x_n)}{f'(x_n)-3f'(\widetilde{y_n})}$
5	4th order Newton [20]	$x_{n+1} = y_n - \frac{f(y_n)}{f'(y_n)}$
6	Three step Newton [4]	$x_{n+1} = x_n^* - \frac{f(x_n^*)}{f'(x_n)} \quad x_n^* = y_n - \frac{f(y_n)}{f'(x_n)}$
7	Hansen and Patrick [11]	$x_{n+1} = x_n - \frac{mf(x_n)}{\left(\frac{m+1}{2m}\right)f'(x_n)-\frac{f(x_n)f''(x_n)}{2f'(x_n)}}$
8	King method [13]	$x_{n+1} = y_n - \frac{f(y_n)}{f'(x_n)}\frac{f(x_n)}{f(x_n)-2f(y_n)}$
9	Kung–Traub [15]	$x_{n+1} = y_n - \frac{f(x_n)f(y_n)}{[f(x_n)-f(y_n)]^2}\frac{f(x_n)}{f'(x_n)}$
10	Potra and Ptak [23]	$x_{n+1} = x_n - \frac{f(x_n)+f(y_n)}{f'(x_n)}$
11	Halley [8]	$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)-\frac{1}{2}\frac{f''(x_n)f(x_n)}{f'(x_n)}}$
12	Dong method [6]	$x_{n+1} = x_n^* - \frac{\frac{m}{m+1}f(x_n)}{\left(1+\frac{1}{m}\right)^m f'(x_n^*)-f'(x_n)}$ $x_n^* = x_n - \frac{m}{m+1}\frac{f(x_n)}{f'(x_n)}$
13	Osada [21]	$x_{n+1} = x_n - \frac{1}{2}m(m+1)\frac{f(x_n)}{f'(x_n)}$ $+\frac{1}{2}(m-1)^2\frac{f'(x_n)}{f''(x_n)}$
14	Grau and Barrero method [9]	$x_{n+1} = x_n^* - \frac{x_n-y_n}{f(x_n)-2f(y_n)}f\left(x_n^*\right)$ $x_n^* = y_n - \frac{x_n-y_n}{f(x_n)-2f(y_n)}f\left(y_n\right)$
15	Noor, 1st method [19]	$x_{n+1} = x_n^* - \frac{f(x_n^*)}{f'(x_n^*)}$ $x_n^* = x_n - \frac{2f(x_n)}{f'(x_n)\pm\sqrt{f'^2(x_n)-4f^3(x_n)}}$
16	Noor, 2nd method [16]	$x_{n+1} = x_n + 4\left(x_n^*-x_n\right)\frac{f(x_n)}{3f(x_n)-2f(x_n^*)}$ $x_n^* = x_n - \frac{1}{2}\frac{2f(x_n)}{f'(x_n)\pm\sqrt{f'^2(x_n)+4f^2(x_n)}}$
17	Nedzhibov method [16]	$x_{n+1} = x_n - \frac{1}{2}\frac{f(x_n)}{f'(x_n)}\left(\frac{3f'(\widetilde{y_n})+f'(x_n)}{3f'(\widetilde{y_n})-f'(x_n)}\right)$
18	Kou et al. method [14]	$x_{n+1} = x_n - \left[1-\frac{3}{4}\right.$ $\left.\frac{\left(f'\widetilde{y_n}-f'(x_n)\right)\left(7f'(\widetilde{y_n})+f'(x_n)\right)}{(3f'(\widetilde{y_n})+5f'(x_n))(2f'(\widetilde{y_n})-f'(x_n))}\right]\frac{f(x_n)}{f'(x_n)}$
19	Sharma and Guha [25]	$x_{n+1} = x_n^* - \frac{f(x_n)+f(y_n)}{f(x_n)-f(y_n)}\frac{f(x_n^*)}{f'(x_n)}$ $x_n^* = y_n - \frac{f(x_n)}{f(x_n)-2f(y_n)}\frac{f(y_n)}{f'(x_n)}$
20	Yun [28]	$x_{n+1} = y_n - \frac{f(y_n)}{f'(x_n)} - \frac{f(x_n^*)}{f'(x_n)}$ $x_n^* = y_n - \frac{f(y_n)}{f'(x_n)}$
21	Fernandez and Aquino method [7]	$x_{n+1} = x_n + \frac{f^2(x_n)}{(f(y_n)-f(x_n))f'(x_n)}$ $-\frac{f^2(y_n)f(x_n)(f(y_n)-3f(x_n))}{(f(y_n)-f(x_n))^2(f(y_n)-2f(x_n))f'(x_n)}$
22	Noor, 3rd method [18]	$x_{n+1} = y_n - \frac{f(y_n)}{f'\left(\frac{x_n+y_n}{2}\right)}$
23	Noor, 4th method [18]	$x_{n+1} = x_n^* - \frac{f(x_n^*)}{f'\left(\frac{x_n+x_n^*}{2}\right)}$ $x_n^* = y_n - \frac{f(y_n)}{f'\left(\frac{x_n+y_n}{2}\right)}$
24	Noor, 5th method [18]	$x_{n+1} = x_n^* - \frac{4f(x_n^*)}{f'(x_n)+3f'\left(\frac{x_n+2x_n^*}{3}\right)}$ $x_n^* = y_n - \frac{4f(y_n)}{f'(x_n)+3f'\left(\frac{x_n+2y_n}{3}\right)}$
25	Shah and Noor 1th method [24]	$x_{n+1} = x_n^* - \frac{2f(x_n^*)f'(x_n)}{2f'^2(x_n)-f(x_n)f''(x_n)}$ $x_n^* = x_n - \frac{2f(x_n)f'(x_n)}{2f'^2(x_n)-f(x_n)f''(x_n)}$
26	Shah and Noor, 2nd method [24]	$x_{n+1} = x_n^* - \frac{2f(x_n^*)f'(x_n)}{2f'^2(x_n)-f(x_n)f''(x_n)}$ $x_n^* = x_n^{**} - \frac{2f(x_n^{**})f'(x_n)}{2f'^2(x_n)-f(x_n)f''(x_n)}$ $x_n^{**} = x_n - \frac{2f(x_n)f'(x_n)}{2f'^2(x_n)-f(x_n)f''(x_n)}$

In this table, m stands for the multiplicity of the roots. In addition, it should be noted that in the relations including $\pm$ sign in the denominator, the

sign should be selected so as to maximize absolute value of the denominator. The utilized intermediate variables used in the above-mentioned relations are defined as follows:

$$y_n = x_n - \frac{f\left(x_n\right)}{f'\left(x_n\right)}, \tag{1}$$

$$\widetilde{y_n} = x_n - \frac{2}{3}\frac{f\left(x_n\right)}{f'\left(x_n\right)}, \tag{2}$$

$$\overline{y_n} = x_n - \frac{1}{2}\frac{f\left(x_n\right)}{f'\left(x_n\right)}. \tag{3}$$

3 Efficiency and performance evaluations

The order of convergence is an important mathematical feature of a nonlinear solver, and the higher order of convergence is an indicator of the better performance of the methods from a mathematical standpoint. However, it is well known that increasing the order of convergence reduces the size of attraction intervals of Newton-type solution methods. The attraction interval is an interval around the root of the function that if the initial guess falls in this interval, the required iteration for converging to the root would be less than a specific number. In the higher-order methods, it is necessary to utilize initial guesses that are closer to the root of the function, which is practically difficult, because, in some cases, the range of possible responses is not known beforehand.

The comparison between different methods is not an easy task and depends on many factors, mostly the context in which the method is going to be used. However, the concentration of this study is on scalar nonlinear equations with real simple roots. In this regard, previous investigators proposed some efficiency indices for this purpose. The most well-known efficiency index is the one proposed by Traub [22]. Many of the reviewed researches utilized this index, which is defined by equation (4), to evaluate the performance of the proposed techniques:

$$EI = p^{\frac{1}{q}}. \tag{4}$$

In this relation, p is the convergence order and q is the number of function evaluations per step (NFE). There is also another common index, which is called the informational index, and is defined as follows:

$$EI_I = \frac{p}{q}. \tag{5}$$

In both indices, the higher value of the index is considered a sign of better performance. This is a widely accepted concept by mathematicians. However, from the practical point of view, for instance, for an engineer who aims to solve a nonlinear equation to find the response to a practical design problem,

it raises some questions. For example, two different methods with the same order of convergence and the total number of function evaluations have the same values of efficiency indices, but it is obvious that their performances are not necessarily similar. To show this complexity, the efficiency indices for the evaluated methods are calculated by the authors and shown in Table 2. These results present the convergence order and NFE, as well as, the values of the two introduced efficiency indices.

Table 2: Efficiency indices of the solution methods

No	Function Name	Convergence order	NFE	Informational Index($\frac{P}{q}$)	Efficiency Index($P^{\frac{1}{q}}$)
1	Newton	2	2	1.000	1.414
2	Ostrowski [22]	4	3	1.333	1.587
3	Traub-Ostrowski [26]	4	3	1.333	1.587
4	Jarrat relation [12]	4	3	1.333	1.587
5	4th order Newton [20]	4	4	1.000	1.414
6	Three step Newton [4]	4	4	1.000	1.414
7	Hansen and Patrick [11]	3	3	1.333	1.442
8	King method [13]	4	3	1.333	1.587
9	Kung–Traub [15]	8	3	2.667	2.000
10	Potra and Ptak [23]	3	3	1.000	1.442
11	Halley [8]	2	3	0.667	1.260
12	Dong method [6]	2	2	1.000	1.414
13	Osada [21]	3	3	1.000	1.442
14	Grau and Barrero method [9]	6	4	1.500	1.565
15	Noor, 1st method [19]	4	4	1.000	1.414
16	Noor, 2nd method [19]	4	3	1.333	1.587
17	Nedzhibov method [16]	3	3	1.000	1.442
18	Kou et al. method [14]	4	3	1.333	1.587
19	Sharma and Guha [25]	6	4	1.500	1.565
20	Yun [28]	4	3	1.333	1.587
21	Fernandez and Aquino method [7]	4	3	1.333	1.587
22	Noor, 3rd method [18]	3	4	0.750	1.316
23	Noor, 4th method [18]	3	4	0.750	1.316
24	Noor, 5th method [18]	3	4	0.750	1.316
25	Shah and Noor 1th method [24]	4	4	1.000	1.414
26	Shah and Noor, 2nd method [24]	5	5	1.000	1.380

It is concluded from this comprehensive study that many of the available methods have the same value as the efficiency indices, but as will be revealed in the coming sections, the solved problems show very different performances for these techniques. Moreover, the effect of the starting point on the per-

formance of a solution method is not included in these indices. Accordingly, it seems that these indices are not sufficient to judge about the applicability of them in solving different types of applied problems, such as engineering problems.

In most of the reviewed research, after the proposition of the method, mathematical proofs about the order of convergence of the suggested schemes are provided. In some of these studies, numerical evaluations were performed, in which; a limited number of test functions were solved, and the obtained results were compared with some other nonlinear solution techniques. The number of iterations, the total number of function evaluations, solution time, computational order of convergence, residual error of the function value, and difference of the last two estimations were the parameters, which were usually recorded in the numerical evaluations. Some of the available solution methods are able to calculate roots of nonlinear functions with very high accuracy, for example 10^{-100}. However, in practical applications, such high accuracy is not required. Instead, a robust method should be able to compute the response of different types of nonlinear functions within the least possible number of iterations. Moreover, the exact ranges of the responses for some practical purposes, such as nonlinear structural problems, are not known beforehand, or it is difficult to estimate such ranges in highly nonlinear problems in structural engineering. Therefore, a robust method must be able to solve a problem with a random initial guess. Therefore, some investigators attempted to study the effects of different initial guesses on the performance of the suggested methods by using assessing the basin of attractions [27, 5, 17, 10].

For the applicable purposes, a method is considered desirable if it can solve different types of nonlinear equations and by using diverse initial guesses. Moreover, a powerful solution technique requires a fewer number of function evaluations and computational time. Accordingly, to rank the reviewed methods and specify the most efficient techniques for applicable purposes, a thorough numerical evaluation program is necessary. Such a responsibility is defined and undertaken in the following of this study.

4 Test functions

To study the performance of these methods, 28 different scalar nonlinear functions are solved. These test equations, which are selected from the reviewed research works, are listed in Table 3. All of these nonlinear equations are famous benchmark problems. As was mentioned previously, the exact ranges of the responses for some practical purposes, such as nonlinear structural problems, are not known beforehand, or it is difficult to estimate such ranges in highly nonlinear problems in structural engineering. Therefore, a robust process must be able to find the response even in the cases that the initial guess is not close to the root. To investigate accu-

rately the effect of the selected starting point on the performance of the approaches, each test function is solved ten times by using ten different initial guesses. These starting points are selected as the endpoints of a symmetrical interval around the roots of the functions. Five intervals, namely, $[x^* - 0.1, x^* + 0.1], [x^* - 1, x^* + 1], [x^* - 10, x^* + 10], [x^* - 100, x^* + 100]$ and $[x^* - 1000, x^* + 1000]$ are selected. They are named very small, small, medium, large, and very large neighboring intervals. As mentioned in Table 3, x^* represents the root of the function.

Table 3: Test functions

No.	Test Functions	x^*
1	$f(x) = x^3 + 4x^2 - 15$	1.63198
2	$f(x) = xe^{x^2} - \sin^2(x) + 3\cos(x) + 5$	-1.20764
3	$f(x) = \cos(x) - x$	0.73908
4	$f(x) = x^3 + 1$	-1.00000
5	$f(x) = 2xe^{-5} + 1 - 2e^{-5x}$	0.13826
6	$f(x) = 2xe^{-10} + 1 - 2e^{-10x}$	0.06931
7	$f(x) = \sin^{-1}(x^2 - 1) - \frac{x}{2} + 1$	0.59481
8	$f(x) = x^5 + 23x - 6$	0.26082
9	$f(x) = x^3 + 4x^2 - 10$	1.36500
10	$f(x) = ln(x^2 + x + 2) - x + 1$	4.15200
11	$f(x) = e^{(-x^2 + x + 2)} - 1$	-1.00000
12	$f(x) = x^5 + x^4 + 4x^2 - 15$	1.34700
13	$f(x) = x^5 + x - 10000$	6.30800
14	$f(x) = (x - 1)^3 - 1$	2.00000
15	$f(x) = x^3 - 10$	2.15440
16	$f(x) = x^3 - 2x - 5$	2.09450
17	$f(x) = (x - 1)^3 - 2$	2.25992
18	$f(x) = e^{(x^2 + 7x - 30)} - 1$	3.00000
19	$f(x) = (x + 2)e^x - 1$	-0.44280
20	$x^3 - e^{-x}$	0.77290
21	$f(x) = e^{(-x^2 + x + 2)} - \cos(x + 1) + x^3 + 1$	-1.0000
22	$f(x) = x^3 + 4x^2 - 25$	2.03500
23	$f(x) = \sin^2(x) + x$	0.00000
24	$f(x) = \tan^{-1}(x) - 1$	1.55740
25	$f(x) = x^3 - \cos(x) + 2$	-1.17250
26	$f(x) = x^3 + 4x^2 + 8x + 8$	-2.00000
27	$f(x) = x^2 - (1 - x)^5$	0.34595
28	$f(x) = \mathrm{xlog}(\mathrm{x}) - 1.2$	2.74064

It must be noted that due to the described approach for the selection of the starting points of the iterative process, the nonlinear functions, with only one root, are selected for this study. The number of iterations, the total number of function evaluations, and convergence time are recorded for each run that reached the root within the admissible number of iterations. In this study, the permissible number of iterations is assumed to be 1000. It is noteworthy that the admissible tolerance for the convergence is selected

as equal to 10^{-10}. This value is mostly more than the necessary value for applicable engineering problems.

5 Proposed efficiency index

In addition to the ability to solve different nonlinear problems for dissimilar initial guesses, a powerful solution technique should be able to compute the response by utilizing a lesser number of function evaluations or computation time. Since the performance of a method for different types of nonlinear problems is not uniform, it is difficult to select more efficient methods directly based on the recorded parameters for each test function, including convergence time and the total number of function evaluations. Comprehensive numerical experiences inform the authors that more sophisticated approaches are necessary. In this study, a computational efficiency index in the following form is suggested:

$$EI_c = \alpha + \beta + 70\left(\frac{i_{\max} - i}{i_{\max} - i_{\min}}\right), \tag{6}$$

where EI_c is the computational index that can be computed for any of the recorded values, including convergence time or total number of function evaluations. The parameters α and β are calculated by the following relations:

$$\alpha = \begin{cases} 0 & \text{if the method diverged,} \\ 10 & \text{if the solver converged to the response,} \end{cases} \tag{7}$$

$$\beta = \begin{cases} 0 & \text{if the method didn't converge within the admissible} \\ & \text{number of iterations,} \\ 20 & \text{if the method converged within the admissible} \\ & \text{number of iterations.} \end{cases} \tag{8}$$

In equation (6), i stands for the selected parameter, which can be the number of function evaluations or the solution time. For a given starting point, $i_{\max}$ and $i_{\min}$ are the maximum and the minimum values of the selected parameter for the different solution methods. Because the number of the maximum allowable iterations is equal to 1000, the maximum and minimums are specified for the methods that have converged to the response in the admissible number of iterations. Therefore, the third term in the right-hand side of equation (6) is equal to 0 for the approaches that have diverged or cannot compute the response within the permissible iterations. It must be noted that the mentioned values for different parameters in the suggested efficiency index are found and proposed by the authors. In fact, the computa-

tional efficiency index and the related parameters are selected in the process of extensive numerical experiences as a tool for providing meaningful results. These are not based on any special mathematical concept. For various solution techniques, these indices are computed in a test function and for each initial guess. Then, the mean value of each neighboring interval is calculated by averaging the values of the initial guesses corresponding to the start and end point of the intervals. Finally, the averages of the derived values for all the test functions are computed.

It is evident that the suggested index, which varies between 0 and 100, has three different phases. These phases are specified by four boundary values of 0, 10, 30, and 100. The average efficiency index equal to 0 indicates that the solution method is not able to solve the nonlinear problems at all. Obviously, it is an extreme case that will not happen for any of the available solution techniques in the average results, because a solver, no matter how weak, is able to solve some type of nonlinear problems. The value of 10 demonstrates that a solution technique can converge to the response of the nonlinear problem but with iterations more than the admissible number. Therefore, if the average computational efficiency index for a solution method in a given neighboring interval falls between 0 and 10, it is an indication that the technique is probably not able to solve the problem on the condition that the initial guess is close to the endpoints of that interval. This probability is higher if the value is closer to 0. It must be noted that the term "probably" in the previous statement is of extreme importance because a solution method does not demonstrate the same performance for different types of nonlinear functions. On the other hand, the calculated indices in this study are derived based on solving a limited number of test functions. Therefore, it is neither possible nor logical to make a "certain" statement.

The next boundary value is 30, which characterizes the borderline between the probability of convergence to the response with fewer and more iterations than the permissible number of iterations. Therefore, the value of the efficiency index between 10 and 30 is the sign of the ability of the solution technique to converge to the root of the nonlinear function for a given neighboring interval. However, the number of required iterations is expected to be more than the admissible iterations. The values closer to 10 indicate a lower probability of convergence. Hence, the methods which have efficiency indices in the range of $(10, 30)$ are not numerically efficient, but there is an acceptable probability that they are able to solve nonlinear problems.

Finally, the last boundary value is 100, which shows that a solution method is the most efficient one, among the evaluated solution techniques. The values of efficiency indices between 30 and 100 show that there is a high probability of solving diverse nonlinear problems by the corresponding solution technique for the given neighboring. A technique is deemed more efficient if its average efficiency indices are closer to 100. It must be stated that the mentioned values are selected by the authors to provide a clear im-

age of the performance of the methods. Obviously, it is possible to choose different boundaries for the three phases of the proposed index.

6 The obtained results

The test functions presented in Table 2 are solved by the nonlinear solution methods listed in Table 1, and the previously mentioned parameters are recorded for each initial guess. To shorten the paper, the recorded values for each problem are not included in the text. The recorded values are used to calculate the performance criteria which were introduced in the previous section. Table 4 presents the total number of failed runs, as well as, the success ratio for each method.

Table 4: The success ratio of different solution methods

Rank	Function Name	Number of failed runs	Success ratio (%)
1	Halley	32	88.57
2	Traub–Ostrowski	33	88.21
2	Hansen and Patrick	33	88.21
3	Ostrowski	34	87.86
3	King method	34	87.86
3	Shah and Noor, 1st method	36	87.14
4	Shah and Noor, 2nd method	36	87.14
4	Grau and Barrero method	37	86.79
5	Noor, 2nd method	37	86.79
5	4th order Newton	38	86.43
6	Newton	39	86.07
7	Kou et al. method	39	86.07
7	Noor, 4th method	39	86.07
7	Jarrat relation	40	85.71
8	Osada	40	85.71
8	Noor, 5th method	40	85.71
8	Three step Newton method	41	85.36
9	Kung-Traub	41	85.36
9	Dong method	41	85.36
9	Yun	41	85.36
9	Noor, 3rd method	41	85.36
9	Potra and Ptak	42	85.00
10	Noor, 1st method	42	85.00
10	Nedzhibov method	42	85.00
10	Sharma and Guha	42	85.00
10	Fernandez and Aquino method	42	85.00

According to the total number of function evaluations, the average computational efficiency indices are listed in Table 5. These outcomes include separate results for each neighboring interval, and an overall average.

Table 5: The average computational efficiency indices based on the total number of function evaluations for different solution methods

No	Function Name	Neighboring interval					
		$[x^* - 1000, x^* + 1000]$	$[x^* - 100, x^* + 100]$	$[x^* - 10, x^* + 10]$	$[x^* - 1, x^* + 1]$	$[x^* - 0.1, x^* + 0.1]$	Average
1	Newton	73.101	76.014	83.273	87.158	96.147	83.139
2	Ostrowski [22]	73.071	77.045	88.475	89.494	96.845	84.986
3	Traub–Ostrowski [26]	71.914	77.424	89.328	89.494	96.845	85.001
4	Jarrat relation [12]	71.999	75.304	83.486	87.640	96.845	83.055
5	4th order Newton [20]	71.777	74.519	81.233	86.306	95.763	81.919
6	Three step Newton [4]	71.226	71.920	81.602	80.263	93.961	79.794
7	Hansen and Patrick [11]	70.477	74.770	91.034	92.024	96.400	84.941
8	King method [13]	73.071	77.045	88.475	89.494	96.845	84.986
9	Kung–Traub [15]	71.784	73.546	83.822	87.438	96.822	82.683
10	Potra and Ptak [23]	71.396	74.216	81.595	82.334	93.190	80.546
11	Halley [8]	70.503	73.860	92.453	92.457	96.845	85.223
12	Dong method [6]	71.204	75.192	84.251	90.540	97.812	83.800
13	Osada [21]	68.974	71.690	80.842	84.379	94.221	80.021
14	Grau and Barrero method [9]	73.621	74.829	84.297	87.235	96.235	83.243
15	Noor, 1st method [19]	66.743	67.428	80.845	88.717	90.966	78.940
16	Noor, 2nd method [19]	28.192	60.246	86.995	86.501	88.122	70.011
17	Nedzhibov method [16]	70.384	73.727	83.486	87.640	96.845	82.417
18	Kou et al. method [14]	72.420	78.026	85.926	86.146	96.863	83.876
19	Sharma and Guha [25]	71.552	75.038	82.283	83.326	96.227	81.685
20	Yun [28]	72.189	73.530	82.751	80.935	95.024	80.886
21	Fernandez and Aquino method [7]	72.002	73.790	81.968	87.158	96.488	82.281
22	Noor, 3rd method [18]	72.489	74.307	82.175	81.684	93.060	80.743
23	Noor, 4th method [18]	73.215	75.330	82.800	84.895	95.768	82.402
24	Noor, 5th method [18]	71.891	75.321	82.977	84.743	95.763	82.139
25	Shah and Noor, 1th method [24]	68.756	73.769	85.590	90.221	95.794	82.826
26	Shah and Noor, 2nd method [24]	66.049	70.799	88.613	89.720	94.820	82.000

Finally, the average computational efficiency indices based on the convergence time are presented in Table 6. The attained results will be discussed in the next section.

7 Discussion about the results

The first interesting finding according to the calculated success ratio is that the Halley approach, which is one of the basic solution techniques, performs better than all the other schemes, including those which are proposed newly and those which have a higher order of convergence. The other remarkable observation is the outstanding performance of the classical Newton method, which is ranked 6, among 26 reviewed procedures. Its success ratio is only about 2 percent less than the Halley technique! Based on the findings, most of the traditional solution techniques are ranked among the top solvers, according to the success ratio, while many of the recently proposed iterative approaches perform poorly.

The derived results for efficiency indices in Tables 5 and 6 provide the opportunity to study the effect of the starting point on the performance of reviewed methods. Before presenting a further discussion about the attained results, it seems necessary to rank the reviewed method based on the obtained outcomes. For the criteria of the success ratio, the presented results in Table 4 are ranked in descending order. As was expected, most of the reviewed techniques have computational efficiency indices higher than 90 for a very small neighboring interval. This is a sign of fast convergence because the starting point of the iterative process is very close to the response. The general trend of efficiency index variation shows that the efficiency of the methods reduces by increasing the distance of the initial guess from the root of the test functions. It is interesting to note that the traditional solution techniques, such as Newton and Traub–Ostrowski perform as well as the newly proposed higher-order schemes and are even better than many of them. As it was expected, the efficiency indices of convergence time and the total number of function evaluations are compatible with each other.

According to the computed indices, the Halley technique is one of the best approaches. An ironic and astonishing finding is that according to the three considered criteria, many of the basic and traditional nonlinear solution techniques, such as, Halley, Hansen and Patrick, Ostrowski, Newton, Traub–Ostrowski, and King are among the best solvers. Obviously more detailed discussion is possible about the performance of the reviewed methods according to the results of the comprehensive numerical evaluation undertaken in this study. For example, it is possible to study the effect of the type of nonlinear functions on the performance of the methods. It is noteworthy that this study, and the presented results can be useful means for the future investigator in the field of nonlinear solution techniques, as well as, the scientists and engineers who seek to select a powerful method for solving practical

Table 6: The average computational efficiency indices based on convergence time for different solution methods

No	Function Name	Neighboring interval					
		$[x^* - 1000, x^* + 1000]$	$x^* - 100, x^* + 100]$	$[x^* - 10, x^* + 10]$	$[x^* - 1, x^* + 1]$	$[x^* - 0.1, x^* + 0.1]$	Average
1	Newton	74.451	78.192	84.341	89.797	98.603	85.077
2	Ostrowski [22]	73.344	77.217	88.019	90.528	97.874	85.396
3	Traub-Ostrowski [26]	72.058	76.861	87.860	89.844	97.132	84.751
4	Jarrat rela-tion [12]	72.668	76.066	83.829	88.956	97.912	83.886
5	4th order Newton [20]	73.845	77.997	83.910	90.135	99.157	85.009
6	Three step Newton [4]	73.439	75.250	83.178	84.405	96.630	82.580
7	Hansen and Patrick [11]	71.501	76.920	92.294	94.485	98.558	86.751
8	King method [13]	73.139	77.217	87.849	90.193	97.622	85.204
9	Kung–Traub [15]	72.074	72.994	82.945	87.763	96.997	82.555
10	Potra and Ptak [23]	73.066	75.573	82.470	84.502	94.799	82.082
11	Halley [8]	71.503	76.377	93.845	94.867	98.964	87.111
12	Dong method [6]	71.666	75.657	84.771	91.766	98.810	84.534
13	Osada [21]	70.783	75.918	83.053	88.715	97.950	83.284
14	Grau and Barrero method [9]	74.486	76.892	85.450	89.366	98.458	84.930
15	Noor, 1st method [19]	68.174	68.515	80.645	90.791	92.626	80.150
16	Noor, 2nd method [19]	28.416	53.152	83.386	84.333	86.913	67.240
17	Nedzhibov method [16]	71.299	75.571	84.629	89.877	98.974	84.070
18	Kou et al. method [14]	73.062	79.502	86.988	88.036	98.742	85.266
19	Sharma and Guha [25]	71.808	74.764	81.532	83.318	96.303	81.545
20	Yun [28]	73.311	74.854	83.280	83.158	96.527	82.226
21	Fernandez and Aquino method [7]	71.627	71.338	79.678	84.718	94.360	80.344
22	Noor, 3rd method [18]	74.152	77.893	84.388	85.932	97.004	83.874
23	Noor, 4th method [18]	74.284	77.414	83.957	87.609	98.373	84.327
24	Noor, 5th method [18]	72.752	77.210	84.164	87.422	98.264	83.963
25	Shah and Noor, 1th method [24]	69.782	75.328	86.200	91.905	97.637	84.170
26	Shah and Noor, 2nd method [24]	66.809	71.563	88.311	90.772	96.203	82.732

nonlinear equations. In order to evaluate the performance of future solution techniques in a more realistic and applicable manner, the proposed efficiency index can be used along with the previous indices.

8 Conclusion

In this study, 26 different iterative nonlinear solution techniques for solving nonlinear scalar equations were reviewed and evaluated numerically. For this purpose, 28 different nonlinear problems, which were selected from a review of the previous studies in this field, were solved by the discussed techniques. To study the effects of the starting point on the performance of solvers, each problem was solved by assuming different initial guesses. The selected starting points for the iterative process were the endpoints of the symmetrical neighboring interval around the root of the solved functions. In each run of a solver, the recorded parameters include the total number of function evaluations and convergence time.

To compare the mentioned methods and rank them according to their performances, three different criteria, namely, success ratio, computational efficiency index of the total number of function evaluations, and also the solution time's efficiency index were defined. The first criterion was a simple ratio of the successful runs to the total number of runs for each scheme. The other two criteria were compared based on a new computational efficiency index, which was suggested by the authors in order to provide a clear picture of the procedure performances. The reason for proposing this new index was the inability of the classical efficiency indices, such as the well-known Traub index, in distinguishing between the performance of different solution methods that have the same order of convergence and the number of function evaluations per step. The solver is considered more efficient if its index is closer to 100. The comprehensive obtained results showed that the higher order of convergence is not necessarily a sign of better performance. Moreover, it is observed that many of the most powerful solution techniques are among the old and traditional approaches, such as Halley, Hansen and Patrick, Ostrowski, and even Newton. Astonishingly, it is found that some of the newly presented solvers are not as operational as the old ones. According to the performances of 26 different iterative techniques, the first four effective procedures for solving nonlinear equations can be ranked as follows: 1. Halley, 2. Traub–Ostrowski, 3. Ostrowski 4. Hansen and Patrick.

References

[1] Babajee, D. and Dauhoo, M. *An analysis of the properties of the variants of Newton's method with third order convergence*, Appl. Math. Comput. 183(1) (2006), 659–684.

[2] Cătinaş, E. *How many steps still left to x^*?*, SIAM Rev. 63(3) (2021), 585–624.

[3] Cătinaş, E. *A survey on the high convergence orders and computational convergence orders of sequences*, Appl. Math. Comput., 343 (2019), 1–20.

[4] Cheney, E.W. and Kincaid, D.R. *Numerical analysis: mathematics of scientific computing*, 1996: Brooks/Cole Publ.

[5] Chun, C. and Neta, B. *Comparison of several families of optimal eighth order methods*, Appl. Math. Comput., 274 (2016), 762–773.

[6] Dong, C. *A family of multiopoint iterative functions for finding multiple roots of equations*, Int. J. Comput. Math. 21(3-4) (1987), 363–367.

[7] Ferngndez-Torres, G. and Vgsquez-Aquino, J. *Three new optimal fourth-order iterative methods to solve nonlinear equations*, Adv. Numer. Anal. 2013, Art. ID 957496, 8 pp.

[8] Gander, W. *On Halley's iteration method*, Amer. Math. Monthly, 92(2) (1985), 131–134.

[9] Grau, M. and Daaz-Barrero, J.L. *An improvement to Ostrowski root-finding method*, Appl. Math. Comput., 173(1) (2006), 450–456.

[10] Gutiérrez, J.M., Magreg M.A. and Varona, J.L. *The "Gauss-Seidelization" of iterative methods for solving nonlinear equations in the complex plane*, Appl. Math. Comput., 218(6) (2011), 2467–2479.

[11] Hansen, E. and Patrick, M. *A family of root finding methods*, Numer. Math. 27(3) (1976/77), 257–269.

[12] Jarratt, P. *Some fourth order multipoint iterative methods for solving equations*, Math. Comput., 20(95) (1966), 434–437.

[13] King, R.F. *A family of fourth order methods for nonlinear equations*, SIAM J. Numer. Anal. 10(5) (1973), 876–879.

[14] 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.

[15] Kung, H. and Traub, J.F.*Optimal order of one-point and multipoint iteration*, J. Assoc. Comput. Mach. 21(4) (1974), 643–651.

[16] Nedzhibov, G.H. and Petkov, M.G. *On a family of iterative methods for simultaneous extraction of all roots of algebraic polynomial*, Appl. Math. Comput., 162(1) (2005), 427–433.

[17] Neta, B., Chun, C. and Scott, M. *Basins of attraction for optimal eighth order methods to find simple roots of nonlinear equations*, Appl. Math. Comput., 227 (2014), 567–592.

[18] Noor, M.A., Waseem, M., Noor, K.I., Ali, M.A. *New iterative technique for solving nonlinear equations*, Appl. Math. Comput., 256 (2015), 1115–1125.

[19] Noor, M.A., Ahmad, F. and Javeed, S. *Two-step iterative methods for nonlinear equations*, Appl. Math. Comput., 181(2) (2006), 1068–1075.

[20] Ortega, J.M. and Rheinboldt, W.C. *Iterative solution of nonlinear equations in several variables*, Reprint of the 1970 original. Classics in Applied Mathematics, 30. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2000.

[21] Osada, N. *An optimal multiple root-finding method of order three*, J. Comput. Appl. Math. 51(1) (1994), 131–133.

[22] Ostrowski, A.M. *Solution of Equations and Systems of Equations*, Pure and Applied Mathematics, Vol. IX. Academic Press, New York-London, 1960 ix+202 pp.

[23] Potra, F. and Ptak, V. *Nondiscrete induction and iterative processes*, Research Notes in Mathematics, 103. Pitman (Advanced Publishing Program), Boston, MA, 1984.

[24] Shah, F.A. and Noor, M.A. *Some numerical methods for solving nonlinear equations by using decomposition technique*, Appl. Math. Comput., 251 (2015), 378–386.

[25] Sharma, J.R. and Guha, R.K. *A family of modified Ostrowski methods with accelerated sixth order convergence*, Appl. Math. Comput., 190(1) (2007), 111–115.

[26] Traub, J. *Iterative Methods for the Solution of Equations*, Prentice-Hall, Englewood Cliffs, New Jersey, 1964.

[27] Varona, J.L. *Graphic and numerical comparison between iterative methods*, Math. Intell. 24(1) (2002), 37–47.

[28] Yun, J.H. *A note on three-step iterative method for nonlinear equations*, Appl. Math. Comput., 202(1) (2008), 401–405.

How to cite this article

Effective numerical methods for nonlinear singular two-point boundary value Fredholm integro-differential equations

S. Amiri

Abstract

We deal with some effective numerical methods for solving a class of nonlinear singular two-point boundary value Fredholm integro-differential equations. Using an appropriate interpolation and a q-order quadrature rule of integration, the original problem will be approximated by the nonlinear finite difference equations and so reduced to a nonlinear algebraic system that can be simply implemented. The convergence properties of the proposed method are discussed, and it is proved that its convergence order will be of $\mathcal{O}(h^{\min\{\frac{7}{2}, q-\frac{1}{2}\}})$. Ample numerical results are addressed to confirm the expected convergence order as well as the accuracy and efficiency of the proposed method.

1 Introduction

In this study, we consider the following nonlinear singular two-point boundary value Fredholm integro-differential equation (SFIDE):

$$(t^\alpha y'(t))' = f(t) + \int_0^1 v(t,s)u(y(s))ds, \quad t \in (0,1], \quad 0 < \alpha < 1, \quad (1)$$

$$y(0) = a, \quad y(1) = b, \quad (2)$$

Sadegh Amiri

Department of Basic Sciences, Shahid Sattari Aeronautical University of Science and Technology, P.O. Box: 13846-63113, Tehran, Iran. e-mail: s.amiri@ssau.ac.ir, amiri-math@yahoo.com

where $f(t), y(t)$ and kernels $v(s,t), u$ are known L_2 functions, and all of them are in $C^4((0,1])$. The nonlinear singular problems are extensively arisen in many applications in physics and astrophysics [11, 9, 12, 19, 20], chemical and mechanical engineering [18, 5, 15, 13], physiological process [14], population dynamics and epidemiology [23], fluid mechanics, electro hydrodynamics, nuclear physics, and chemical kinetics [2, 10].

Also, the nonlinear singular problems have many applications in stellar structure, thermal explosions, isothermal gas spheres, radiative cooling, thermionic currents, and the thermal behavior of a spherical cloud of gas [3, 17, 22].

The majority of engineering applications and various branches of science, such as financial mathematics, oceanography, population dynamics, fluid mechanics, plasma physics, electromagnetic theory, artificial neural networks, and biological processes, have been dominated by Fredholm integro-differential equations (FIDEs) [4, 6]. In general, the analytical solution of FIDEs is not available. As a result, various numerical techniques for determining approximate solutions of FIDEs have been created. The situation is significantly more complicated for FIDEs with singularities. A particular type of them called singularly perturbed Fredholm integro-differential equations (SPFIDEs), was discussed in [1, 6, 8, 7]. Numerical analysis of SPFIDEs has not yet been widely utilized. In this study, we focus on a specific case of nonlinear singular two-point boundary value Fredholm integro-differential equations of the form (1)–(2). Since solving problems of this type is very difficult, the main motivation of this study is to construct an efficient and useful numerical method with $\mathcal{O}(h^{\frac{7}{2}})$ accuracy in the L_2 norm for nonlinear singular problems of the form (1).

To formulate some accurate and effective methods for (1), we first apply a finite difference method to discretize the singular ordinary differential equation part and a suitable quadrature rule of integration for the singular two-point boundary value Fredholm integro-differential part of (1).

Then, the original problem is converted into a system of nonlinear algebraic equations. The numerical solution of the derived nonlinear system is computed by using some solver like the Newton method. Also, the convergence analysis of the present method is established.

The main features of the new method are as follows:

- It can be simply implemented by converting the singular problem into a system of nonlinear algebraic equations.

- The convergence rate of the proposed method is $\mathcal{O}(h^4)$ with respect to the L_∞ norm when applied to nonlinear singular problems.

- The proposed method is successful in solving some classes of singular problems, such as SFIDEs and SPFIDE.

- The provided comparative numerical simulations confirm that the proposed method is more accurate than the existing methods reported in the literature.

2 Formulation of the method

In this section, we formulate a novel numerical method for solving the two-point boundary value Fredholm integro-differential equation (1). At first if one takes $q(t) = t^\alpha y'(t)$, then (1) reduces to

$$q'(t) = f(t) + \int_0^1 v(t,s)u(y(s))ds. \tag{3}$$

Consider the partition $\{t_k = kh \ : \ k = 0, 1, \ldots, N\}$ of the interval $[0,1]$, where $t_0 = 0$ and $t_N = 1$ and $h = \frac{1}{N}$ denotes the step size. For $k = 0, 1, \ldots, N$, let Y_k and $V_{k,n}$ denote the approximate values of $y_k := y(t_k)$ and $v_{k,n} := v(t_k, t_n)$, respectively. For (3), we can conclude that

$$q(t) - q_k = \int_{t_k}^t f(\xi)d\xi + \int_{t_k}^t \int_0^1 v(\xi,s)u(y(s))dsd\xi. \tag{4}$$

Dividing both sides of (4) by t^α and then integrating over $[t_k, t_{k+1}]$ and $[t_{k-1}, t_k]$, we have

$$\int_{t_k}^{t_{k\pm1}} \left(y'(t) - \frac{q_k}{t^\alpha}\right)dt = \int_{t_k}^{t_{k\pm1}} \int_{t_k}^t \frac{f(\xi)}{t^\alpha}d\xi dt + \int_{t_k}^{t_{k\pm1}} \int_{t_k}^t \int_0^1 \frac{v(\xi,s)}{t^\alpha}u(y(s))dsd\xi dt.$$

By changing the order of integration, we get

$$y_{k\pm1} - y_k \mp q_k T_{k\pm\lfloor\frac{k\mp1}{k}\rfloor} = f_k^\pm + \int_0^1 u(y(s))v_k^\pm(s)ds, \tag{5}$$

where

$$T_{k-1} = \frac{t_k^{1-\alpha} - t_{k-1}^{1-\alpha}}{1-\alpha}, \qquad f_k^\pm = \int_{t_k}^{t_{k\pm1}} \frac{t_{k\pm1}^{1-\alpha} - \xi^{1-\alpha}}{1-\alpha}f(\xi)d\xi,$$

$$v_k^\pm(s) = \int_{t_k}^{t_{k\pm1}} \frac{t_{k\pm1}^{1-\alpha} - \xi^{1-\alpha}}{1-\alpha}v(\xi,s)d\xi, \tag{6}$$

and $k = 1, \ldots, N-1$. Eliminating q_k in (5) concludes that

$$\frac{1}{T_{k-1}}(y_k - y_{k-1} + f_k^-) + \frac{1}{T_k}(y_k - y_{k+1} + f_k^+)$$

$$+ \int_0^1 u(y(s))\left(\frac{1}{T_{k-1}}v_k^-(s) + \frac{1}{T_k}v_k^+(s)\right)ds = 0. \tag{7}$$

To solve (7), it is sufficient to utilize some suitable numerical integration methods to approximate $f^{\pm}$, $v^{\pm}$, and its integral part. By using the interpolating polynomials of $f(\xi)$ and $v(\xi, \cdot)$ at nodes t_k and $t_{k\pm 1}$, we can approximate the integrals given in (6) as follows:

$$
\begin{cases}
f_k^{\pm} = a_{0,k}^{\pm} f(t_k) + a_{1,k}^{\pm} f(t_{k\pm 1}) + a_{2,k}^{\pm} f''(t_k) + a_{3,k}^{\pm} f'''(\xi_k^{\pm}), \\
v_k^{\pm}(s) = a_{0,k}^{\pm} v(t_k, s) + a_{1,k}^{\pm} v(t_{k\pm 1}, s) + a_{2,k}^{\pm} \left. \frac{\partial^2 v}{\partial t^2}(t,s) \right|_{t=t_k} + a_{3,k}^{\pm} \left. \frac{\partial^3 v}{\partial t^3}(t,s) \right|_{t=\zeta_k^{\pm}},
\end{cases}
$$

in which $\xi_k^-, \zeta_k^- \in (t_{k-1}, t_k)$ and $\xi_k^+, \zeta_k^+ \in (t_k, t_{k+1})$ and

$$
\begin{cases}
a_{0,k}^{\pm} = \sum_{j=0}^{1} \frac{(-1)^j}{2-\alpha-j} \binom{1}{j} (t_{k\pm 1}^{2-\alpha-j} - t_k^{2-\alpha-j}) t_k^j \\
\qquad \mp \frac{1}{2h} \sum_{j=0}^{2} \frac{(-1)^j}{3-\alpha-j} \binom{2}{j} (t_{k\pm 1}^{3-\alpha-j} - t_k^{3-\alpha-j}) t_k^j & (8a) \\
a_{1,k}^{\pm} = \pm \frac{1}{2h} \sum_{j=0}^{2} \frac{(-1)^j}{3-\alpha-j} \binom{2}{j} (t_{k\pm 1}^{3-\alpha-j} - t_k^{3-\alpha-j}) t_k^j, & (8b) \\
a_{2,k}^{\pm} = \frac{1}{6} \sum_{j=0}^{3} \frac{(-1)^j}{4-\alpha-j} \binom{3}{j} (t_{k\pm 1}^{4-\alpha-j} - t_k^{4-\alpha-j}) t_k^j \\
\qquad \mp \frac{h}{4} \sum_{j=0}^{2} \frac{(-1)^j}{3-\alpha-j} \binom{2}{j} (t_{k\pm 1}^{3-\alpha-j} - t_k^{3-\alpha-j}) t_k^j, & (8c) \\
a_{3,k}^{\pm} = \pm \frac{h}{4} \sum_{j=0}^{2} \frac{(-1)^j}{3-\alpha-j} \binom{2}{j} (t_{k\pm 1}^{3-\alpha-j} - t_k^{3-\alpha-j}) t_k^j. & (8d)
\end{cases}
$$

Assume that the functions $f^{(4)}(t)$ and $\frac{\partial^4 v}{\partial t^4}(t,s)$ are continuous. Then there are the values $\varsigma_k, \tilde{\varsigma}_k \in (t_{k-1}, t_{k+1})$ and $\zeta_k, \tilde{\zeta}_k \in (\min(\xi_k^{\pm}), \max(\xi_k^{\pm}))$ such that

$$
\begin{cases}
\frac{1}{T_{k-1}} f_k^- + \frac{1}{T_k} f_k^+ = \varphi_k^0 f(t_k) + \varphi_k^- f(t_{k-1}) + \varphi_k^+ f(t_{k+1}) + e_k(f), & (9a) \\
\frac{1}{T_{k-1}} v_k^-(s) + \frac{1}{T_k} v_k^+(s) = \varphi_k^0 v(t_k, s) + \varphi_k^- v(t_{k-1}, s) + \varphi_k^+ v(t_{k+1}, s) + e_k(v(\cdot, s)), & (9b)
\end{cases}
$$

where $\varphi_k^0 = b_{0,k} - \frac{2}{h^2} b_{2,k}$, $\quad \varphi_k^{\pm} = \frac{1}{T_{k-1}} a_{1,k}^{\pm} + \frac{1}{h^2} b_{2,k}$,

$$
\begin{cases}
e_k(f) = -\frac{1}{12} h^2 b_{2,k} f^{(4)}(\varsigma_k) + b_{3,k} f^{(3)}(\zeta_k), \\
e_k(v(\cdot, s)) = -\frac{1}{12} h^2 b_{2,k} \left. \frac{\partial^4 v}{\partial t^4}(t,s) \right|_{t=\tilde{\varsigma}_k} + b_{3,k} \left. \frac{\partial^3 v}{\partial t^3}(t,s) \right|_{t=\tilde{\zeta}_k},
\end{cases}
$$

and $b_{l,k} = \frac{1}{T_{k-1}} a_{l,k}^- + \frac{1}{T_k} a_{l,k}^+$, $l = 0, 2, 3$. Finally, by applying the relations (9) and utilizing a suitable numerical quadrature method of order q with weights $w = (w_0, w_1, \ldots, w_N)^{\top}$, (7) can be reformulated as

$$-\frac{1}{T_{k-1}}y_{k-1} + \left(\frac{1}{T_{k-1}} + \frac{1}{T_k}\right)y_k - \frac{1}{T_k}y_{k+1} + \hat{F}_k + h\sum_{n=0}^{N} w_n u_n \hat{V}_{k,n} = e_k(f) + \mathcal{O}(h^{q+1}),$$

(10)

where for $k = 1, \ldots, N-1$ and $n = 0, 1, \ldots, N$, we set

$$\begin{cases} u_n := u(y_n), \\ \hat{F}_k := \varphi_k^0 f_k + \varphi_k^- f_{k-1} + \varphi_k^+ f_{k+1}, \\ \hat{V}_{k,n} := \varphi_k^0 v_{k,n} + \varphi_k^- v_{k-1,n} + \varphi_k^+ v_{k+1,n}. \end{cases}$$

Therefore, an approximate method to solve the problem (1) can be formulated as follows:

$$-\frac{1}{T_{k-1}}Y_{k-1} + \left(\frac{1}{T_{k-1}} + \frac{1}{T_k}\right)Y_k - \frac{1}{T_k}Y_{k+1} + \hat{F}_k + h\sum_{n=0}^{N} w_n U_n \hat{V}_{k,n} = 0,$$

(11)

where $U_n = u(Y_n)$. Take note that the Newton method can be used to solve the derived nonlinear equations. Let us set $\mathbf{Y} = (Y_1, \ldots, Y_{N-1})^\top$, $\mathbf{F} = (\hat{F}_1, \ldots, \hat{F}_{N-1})^\top$, $\mathbf{V} = \left[\mathbf{V}^1, \ldots, \mathbf{V}^{N-1}\right]$, and $\mathbf{V}^k = \left(\hat{V}_{1,k}, \hat{V}_{2,k}, \ldots, \hat{V}_{N-1,k}\right)^\top$ for $k = 1, \ldots, N-1$. Then the matrix formulation of the proposed method (11) is also written in the following form:

$$\mathbf{T}\,\mathbf{Y} + h\,\mathbf{L}\,u(\mathbf{Y}) = -\mathbf{F} - w_0 U_0 \mathbf{V}^0 - w_N U_N \mathbf{V}^N + \tau_0 Y_0 \mathbf{I}_1 + \tau_N Y_N \mathbf{I}_{N-1}, \quad (12)$$

where $\mathbf{W} = \mathrm{diag}(w_1, \ldots, w_{N-1})$, $\mathbf{L} = \mathbf{V}\mathbf{W}$ and

$$\mathbf{T} = \mathrm{tridiag}(-\mathbf{T}_{N-2}^1, \mathbf{T}_{N-2}^0 + \mathbf{T}_{N-1}^1, -\mathbf{T}_{N-2}^1), \quad (13)$$

is a tridiagonal matrix with $\mathbf{T}_n^k = [\tau_k, \tau_{k+1}, \ldots, \tau_n]^\top$ and $\tau_k = \frac{1}{T_k}$ for $k = 0, 1, \ldots, N$. The symbol $\mathbf{I}_i$ signifies an $(N-1)$-column vector with entry 1 in the position i and 0 elsewhere, where $i = 1, N-1$.

Remark 1. It is worth noting that according to the relation $t^\alpha y'' = (t^\alpha y')' - \alpha t^{\alpha-1} y'$, the presented technique may be utilized for the singularly perturbed Fredholm integro-differential equations discussed in [1, 6, 8, 7], as well as the singularly perturbed boundary value problems considered in [10, 13].

2.1 Convergence analysis

In this section, the convergence analysis of the presented method (11) to solve the SFIDE (1) is performed. To this end, we set $\mathbf{y} = (y_1, y_2, \ldots, y_{N-1})^\top$. Then a matrix formulation of (10) is also derived as

$$\mathbf{T}\,\mathbf{y} + h\,\mathbf{L}\,u(\mathbf{y}) = \mathbf{E}(f) - \mathbf{F} - w_0 U_0 \mathbf{V}^0 - w_N U_N \mathbf{V}^N + \tau_0 Y_0 \mathbf{I}_1$$
$$+ \tau_N Y_N \mathbf{I}_{N-1} + \mathcal{O}(h^{q+1})\mathbf{1}_{N-1}, \tag{14}$$

where $\mathbf{E}(f) = (e_1(f), e_2(f), \ldots, e_{N-1}(f))^\top$ and $\mathbf{1}_{N-1} = (1, 1, \ldots, 1)^\top \in \mathbb{R}^{N-1}$. By subtracting (12) from (14), the error equation can be derived as

$$\mathbf{T}(\mathbf{y} - \mathbf{Y}) + h\,\mathbf{L}\big(u(\mathbf{y}) - u(\mathbf{Y})\big) = \bar{\mathbf{E}},$$

where $\bar{\mathbf{E}} = \mathbf{E}(f) + \mathcal{O}(h^{q+1})\mathbf{1}_{N-1}$. Thus we have

$$(\mathbf{T} + h\,\mathbf{L}\mathbf{J}_U)(\mathbf{y} - \mathbf{Y}) = \bar{\mathbf{E}}, \tag{15}$$

where $\mathbf{J}_U$ is a diagonal matrix containing the Jacobian of kernel $u(y)$. That is, $\mathbf{J}_U = \mathrm{diag}\big(\big[\frac{\partial}{\partial y}u(y)\big|_{y=y_k}\big]_{k=1}^{N-1}\big)$. To formulate an upper bound for the L_2-error $\|\mathbf{y} - \mathbf{Y}_2\|$ derived in (15), we first prove the following lemmas.

Lemma 1. Let $A \in \mathbb{R}^{n \times n}$ and $\mathbf{x} \in \mathbb{R}^n$. Then

$$\|A\mathbf{x}\|_2 \geq \sigma_{\min}(A)\|\mathbf{x}\|_2,$$

in which $\sigma_{\min}(A)$ is the smallest singular value of matrix A.

Proof. If we consider the singular value decomposition of the form $A = S\Sigma Z^\top$, where S and Z are orthogonal and $\Sigma = diag(\sigma_1, \ldots, \sigma_n)$ is the diagonal matrix with singular values $\sigma_k, k = 1, \ldots, n$. Then from the orthogonality of S and Z, we have

$$\|A\mathbf{x}\|_2 = \big\|S\Sigma Z^\top \mathbf{x}\big\|_2 = \big\|S(\Sigma Z^\top \mathbf{x})\big\|_2 = \big\|\Sigma Z^\top \mathbf{x}\big\|_2$$
$$= \left\| \begin{bmatrix} \sigma_1 \sum_{i=1}^n z_{1,i} x_i \\ \vdots \\ \sigma_n \sum_{i=1}^n z_{n,i} x_i \end{bmatrix} \right\|_2 = \sqrt{\sum_{k=1}^n \sigma_k^2 \left(\sum_{i=1}^n z_{k,i} x_i\right)^2},$$

where $\mathbf{x} = (x_1, \ldots, x_N)^\top$. Let $\sigma_{\min}(A) = \min\{\sigma_k; k = 1, \ldots, n\}$. This concludes that

$$\|A\mathbf{x}\|_2 \geq \sigma_{\min}(A)\sqrt{\sum_{k=1}^n \left(\sum_{i=1}^n z_{k,i} x_i\right)^2} = \sigma_{\min}(A)\|Z^\top \mathbf{x}\|_2 = \sigma_{\min}(A)\|\mathbf{x}\|_2.$$

$\square$

In the following, we may try to construct lower and upper triangular matrices L_1 and U_1 such that the tridiagonal matrix A can be expressed as the product $A = L_1 U_1$ of the form

$$
\underbrace{\begin{bmatrix} d_1 & c_1 & & & \\ a_1 & d_2 & c_2 & & \\ & \ddots & \ddots & \ddots & \\ & & a_{n-2} & d_{n-1} & c_{n-1} \\ & & & a_{n-1} & d_n \end{bmatrix}}_{A} = \underbrace{\begin{bmatrix} 1 & & & \\ l_1 & 1 & & \\ & \ddots & \ddots & \\ & & l_{n-1} & 1 \end{bmatrix}}_{L_1} \underbrace{\begin{bmatrix} p_1 & c_1 & & \\ & p_2 & \ddots & \\ & & \ddots & c_{n-1} \\ & & & p_n \end{bmatrix}}_{U_1}.
$$
(16)

Indeed, multiplying L_1 by U_1 yields the following recursive relations:

$$
p_1 = d_1, \quad l_k = \frac{a_k}{p_k}, \quad p_{k+1} = d_{k+1} - l_k c_k, \quad k = 1, 2, \dots, n-1. \tag{17}
$$

In the following lemma, we will exhibit that the decomposition (16)–(17) for the matrices with the property of strictly diagonally dominant is unique.

Lemma 2. If A is a strictly diagonally dominant matrix, then it has a unique LU-factorization of the form (16)–(17).

Proof. It is sufficient to show that the elements p_k in (17) are nonzero for $k = 1, \dots, n$. It can be done by induction. So, we show that $|p_k| \geq \delta_k + |c_k|$, where $\delta_k = |d_k| - |a_{k-1}| - |c_k| > 0, k = 1, 2, \dots, n$, and $a_0 = c_n = 0$. Since $|p_1| = |d_1| = \delta_1 + |c_1|$, assuming $|p_k| \geq \delta_k + |c_k|$ for some $k = 1, 2, \dots, n-1$, concludes that $\frac{|c_k|}{|p_k|} < 1$. Therefore, according to (17) and the strictly diagonal dominant of the matrix A, we have

$$
|p_{k+1}| = |d_{k+1} - l_k c_k| = \left| d_{k+1} - \frac{a_k c_k}{p_k} \right| \geq |d_{k+1}| - \frac{|a_k||c_k|}{|p_k|}
$$
$$
\geq |d_{k+1}| - |a_k| = \delta_{k+1} + |c_{k+1}|.
$$

$\square$

In the next lemma, we can see that under some conditions there is an LU-factorization in the form (16)–(17) for every weak dominant tridiagonal matrix.

Lemma 3. Assume that A is a tridiagonal matrix with the property of weakly diagonally dominant. If in addition $|d_1| > |c_1|$ and $a_k \neq 0, k = 1, 2, \dots, n-2$, then it has a unique LU-factorization of the form (16)–(17). Moreover, if $d_n \neq 0$, then A is nonsingular.

Proof. The proof of this lemma is similar to Lemma 2. $\square$

In the next step, we investigate some properties of the matrix $\mathbf{T}$ given by (13). Since $\mathbf{T}$ is a weak diagonal dominant symmetric matrix with positive diagonal elements, then it is a positive semidefinite matrix. It is easily seen

that the elements of $\mathbf{T}_n^0$ and $\mathbf{T}_n^1$ are not vanished. Consequently, from Lemma 3, we conclude that $|\mathbf{T}| \neq 0$ and $p_k > 0$ for $k = 1, 2, \ldots, N-1$. This yields that all of the singular values σ_k as well as eigenvalues λ_k of $\mathbf{T}$ are positive. Now we establish that $\lambda_{\min}(\mathbf{T}) > h$. To this end, we first define $\mathbf{M} = \frac{1}{h}\mathbf{T} - \mathbf{I}$, in which $\mathbf{I}$ is the identity matrix of order $N-1$. So, if λ is an eigenvalue of $\mathbf{T}$, then $(\frac{1}{h}\lambda - 1)$ is the eigenvalue of $\mathbf{M}$. Hence, it is sufficient to prove that $\lambda_{\min}(\mathbf{M})$ (the smallest eigenvalue of $\mathbf{M}$) is positive. Since $\mathbf{M}$ is a strictly diagonally dominant matrix, then from Lemma 2, there exists a unique LU-factorization of the form (16)–(17) for this matrix with the following coefficients:

$$\begin{cases} d_k = \frac{1}{h}(\tau_{k-1} + \tau_k) - 1, & k = 1, 2, \ldots, N-1, \\ a_k = c_k = -\frac{1}{h}\tau_k, & k = 1, 2, \ldots, N-2. \end{cases}$$

Remark 2. Since $h = \frac{1}{N}$, the coefficients given by (8)–(9) can be approximated as

$$\tau_k \sim \frac{t_k^\alpha}{h}, \quad a_{1,k}^\pm \sim \frac{h^2}{6} t_k^{-\alpha}, \quad b_{0,k} \sim h, \quad b_{2,k} \sim -\frac{h^3}{12}, \quad b_{3,k} \sim -\frac{h^5}{24}, \quad (18)$$

as $h \to 0$. Therefore, the elements $\varphi_k^\pm, \varphi_k^0$, and V^n of the matrix $\mathbf{V}$ will be of order $\mathcal{O}(h)$. Finally, we conclude that the elements of the matrix $\mathbf{L}$ will be of order $\mathcal{O}(h)$.

Lemma 4. If there is an LU-factorization for the matrix $\mathbf{M}$ in the form (16)–(17), then $|\mathbf{M}| > 0$ and $p_k > 0$ for $k = 1, \ldots, N-1$.

Proof. Putting $\bar{p}_k = hp_k, \bar{l}_k = -l_k$ and utilizing (17) yield $\bar{p}_{k+1} = \tau_{k+1} + \tau_k - h - \tau_k \bar{l}_k$. Therefore, we get,

$$\bar{l}_{k+1} = \frac{\gamma_{k+1}}{\gamma_{k+1} + 1 - \bar{l}_k - h/\tau_k},$$

in which $\gamma_{k+1} = \frac{\tau_{k+1}}{\tau_k}, k = 1, \ldots, N-1$. It should be mentioned that $\bar{l}_1 = \frac{\tau_1}{\tau_0 + \tau_1} < 1$ and $\lim_{k \to \infty} \bar{l}_k = 1$. So, for sufficiently small h, assuming $\bar{l}_k < 1 - \frac{h}{\tau_{k-1}} < 1$ concludes $\bar{l}_{k+1} < 1 - \frac{h}{\tau_k} < 1$. Totally, we have $0 < \bar{l}_k < 1, k = 1, \ldots, N-1$. Therefore, with the help of (18), we get

$$\bar{p}_{k+1} > \tau_{k+1} - h > k.$$

$\square$

From Lemma 4, we can reach that the determinant of all upper-left submatrices of $\mathbf{M}$ is positive. Thanks [24, Theorem 7.2], this implies that $\mathbf{M}$ is a positive definite matrix and all its eigenvalues are positive. It means that eigenvalues of $\mathbf{T}$ must be fulfilled $\lambda > h$. Now, We set $A = (\mathbf{T} + h\mathbf{LJ}_U)$ and $\mathbf{x} = \mathbf{y} - \mathbf{Y}$. Then, using the Lemma 1 for (15), we conclude that

$$\|\mathbf{y} - \mathbf{Y}\|_2 \leq \frac{\|\bar{\mathbf{E}}\|_2}{\sigma_{\min}\left(\mathbf{T} + h\mathbf{L}\mathbf{J}_U\right)}. \tag{19}$$

Since $\mathbf{T}$ is a nonsingular positive semidefinite tridiagonal matrix and using Remark 2, we can easily deduce that $\sigma_{\min}\left(\mathbf{T} + h\mathbf{L}\mathbf{J}_U\right) \sim \sigma_{\min}\left(\mathbf{T}\right)$.

Theorem 1. Let the fourth-order derivatives of functions $f(t), v(t,s), u(y(t))$ are continuous. Then for all $\alpha \in (0,1)$, we have $\|\mathbf{y} - \mathbf{Y}\|_2 = \mathcal{O}(h^{\min\{\frac{7}{2}, q - \frac{1}{2}\}})$.

Proof. From the continuity of third and fourth order derivative of the corresponding functions and according to the relations (9) and (18), we conclude that there exists constant $\bar{c} \in \mathbb{R}$ such that

$$\left\|\bar{\mathbf{E}}\right\|_2 \leq \bar{c}h^{\frac{9}{2}} + ch^{q + \frac{1}{2}},$$

and so, from (19), we get

$$\|\mathbf{y} - \mathbf{Y}\|_2 \leq \frac{\bar{c}h^{\frac{9}{2}} + ch^{q + \frac{1}{2}}}{\lambda_{\min}\left(\mathbf{T}\right)} \leq \frac{\bar{c}h^{\frac{9}{2}} + ch^{q + \frac{1}{2}}}{h} = \bar{c}h^{\frac{7}{2}} + ch^{q - \frac{1}{2}}.$$

$\square$

According to Theorem 1, the maximum order of convergence is achieved when $q \geq 4$. Therefore, we use a Simpson quadrature rule to discretize the integral parts of (6).

Remark 3. As is well known, all norms are equivalent for every $\mathbf{z} \in \mathbb{R}^n$; that is, $\|\mathbf{z}\|_1 \leq \sqrt{n}\|\mathbf{z}\|_2 \leq n\|\mathbf{z}\|_\infty$. As a result, if $q \geq 4$, then $\|\mathbf{y} - \mathbf{Y}\|_1 = \mathcal{O}(h^3)$ and $\|\mathbf{y} - \mathbf{Y}\|_\infty = \mathcal{O}(h^4)$ for the proposed method (11).

3 Numerical examples

The performance of the proposed method to solve the SFIDE (1) is demonstrated in this section. In the following numerical simulations, the step size is selected as $h = 2^{-k}, k = 2, \ldots, 8$, and then the error $\|\mathbf{y} - \mathbf{Y}\|_2$ is computed.

Example 1. As a first example for SFIDE (1), we consider $u(y) = \exp(-y)$ and

$$\begin{aligned} v(t,s) =& v_0 \ t^{1+2\alpha} s^{2+\alpha} \sin\left(\mu\pi t^{3+\alpha}\right) \sin\left(\mu\pi s^{3+\alpha}\right), \\ f(t) =& t^{1+2\alpha} \left(3 + \alpha\right) \left(-\mu^2\pi^2 t^{3+\alpha} \left(3 + \alpha\right) \cos\left(\mu\pi t^{3+\alpha}\right) \right. \\ & \left. -2 \left(\mu\pi + 1\right)\left(1 + \alpha\right) \sin\left(\mu\pi t^{3+\alpha}\right)\right). \end{aligned}$$

Then the exact solution is $y(t) = \cos\left(\mu\pi t^{\alpha+3}\right)$, where $v_0 = \frac{2\mu\pi(\alpha+1)(\alpha+3)^2}{\exp(-\cos(\mu\pi)) - \exp(-1)}$ and $\alpha \in (0,1)$.

Table 1: L_2 error and order of the method (11) for Example 1.

N	$\alpha = \mu = \frac{1}{2}$	Order	$\alpha = \frac{1}{2}, \mu = 1$	Order	$\alpha = \frac{1}{8}, \mu = \frac{5}{2}$	Order
4	8.9150e-02	—	6.1400e-01	—	3.0755e+00	—
8	5.9426e-03	3.90	7.0005e-02	3.13	7.0548e-01	2.12
16	4.6722e-04	3.66	6.2258e-03	3.49	6.8774e-02	3.35
32	3.2741e-05	3.83	4.3233e-04	3.84	8.1641e-03	3.07
64	1.1031e-04	3.60	3.5398e-05	3.61	6.9052e-04	3.56
128	2.6914e-06	3.52	3.0591e-06	3.53	6.1420e-05	3.49
256	2.0531e-08	3.50	2.6884e-07	3.50	5.4073e-06	3.50

It should be mentioned that, in this example, the Jacobian of the kernel $u(y)$ is not positive and that increasing μ causes more oscillations of the solution y. We computed the numerical solution of this singular problem by the proposed method (11). The numerical results of this test problem in the form of the L_2 error and the order of the method are reported in Table 1. From this table, we can see that the desired order of convergence of the presented method is obtained. In Figure 1(a), the exact solution of the problem given by Example 1 is compared with the approximate solution derived by the proposed method (11) when $h = 2^{-8}$, $\alpha = \frac{2}{3}$, and $\mu = \frac{7}{2}$. The absolute error of the present method to solve this example is plotted in Figure 1(b).
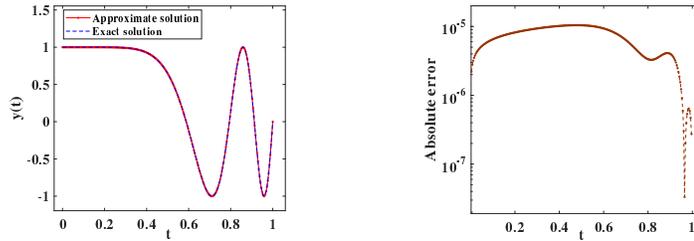
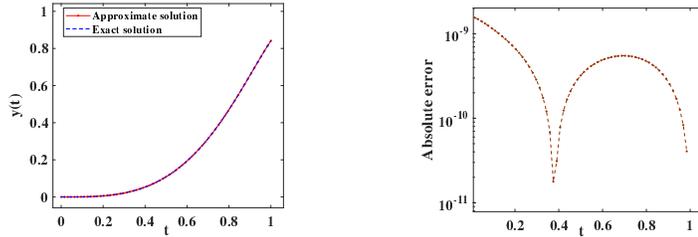(a) Exact and approximate solutions for $h = 2^{-8}$. (b) Absolute error.

Figure 1: Numerical results of the proposed method to solve Example 1 with $\alpha = \frac{2}{3}, \mu = \frac{7}{2}$.

Example 2. As a second example for SFIDE (1), we consider $u(y) = \exp(y)$ and

$$v(t,s) = \mu_0 \ t^{1+2\alpha}s^{2+\alpha} \cos\left(t^{3+\alpha}\right) \cos\left(s^{3+\alpha}\right), \quad f(t) = -(\alpha+3)^2 \ t^{4+3\alpha} \sin\left(t^{3+\alpha}\right).$$

Then the exact solution will be $y(t) = \sin\left(t^{\alpha+3}\right)$, where $\mu_0 = \frac{2(1+\alpha)(3+\alpha)^2}{\exp(\sin(1))-1}$ for $\alpha \in (0,1)$.

This example and Example 1 are similar, but the sign of its Jacobian J_U of the kernel $u(y)$ is unlike that of the ones in Example 1. We report the

(a) Exact and approximate solutions for $h = 2^{-8}$.   (b) Absolute error.

Figure 2: Numerical results of the proposed method to solve Example 2 with $\alpha = \frac{1}{5}$.

Table 2: L_2 error and order of the method (11) for Example 2.

N	$\alpha = \frac{1}{3}$	Order	$\alpha = \frac{1}{2}$	Order	$\alpha = \frac{2}{3}$	Order
4	2.2839e-02	—	3.1452e-02	—	4.3256e-02	—
8	2.6238e-03	3.12	3.8926e-03	3.01	5.7418e-03	2.91
16	2.1676e-04	3.59	3.3580e-04	3.53	5.1593e-04	3.47
32	1.7916e-05	3.59	2.8937e-05	3.53	4.5736e-05	3.49
64	1.4922e-06	3.58	2.5547e-06	3.50	4.1658e-06	3.45
128	1.2372e-07	3.59	2.1784e-07	3.55	3.7527e-07	3.47
256	1.0570e-08	3.54	1.8648e-08	3.54	3.3067e-08	3.50

numerical results of this example with various values of α and step size h. Again, from Table 2, we can observe that the expected order of convergence $\frac{7}{2}$ is achieved. For $\alpha = \frac{1}{5}$ and $h = 2^{-8}$, the exact and approximate solutions are depicted in Figure 2(a). The absolute error plotting in Figure 2(b) shows that the present method is accurate and successful.

Example 3. As a third example for SFIDE (1), we consider $u(y) = -y^5$ and

$$v(t,s) = v_0\, t^{2\alpha} s^{1+\alpha} \left(1 + t^{2+\alpha}\right)^{\beta-1}, \quad f(t) = f_0\, t^{2+3\alpha} \left(1 + t^{2+\alpha}\right)^{-2+\beta}.$$

Then the exact solution will be $y(t) = \left(1 + t^{\alpha+2}\right)^{\beta}$, where $v_0 = \frac{\beta(5\beta+1)(1+2\alpha)(2+\alpha)^2}{1-2^{5\beta+1}}$ and $f_0 = \beta(\beta-1)(2+\alpha)^2$ for $\alpha \in (0,1)$ and $\beta > 0$.

We solved this singular boundary value problem for some selected values of α and β. The L_2 norm of the errors is computed for the presented method (11) and then is exhibited in Table 3. It can be seen that the numerical results verify that the claimed order of the convergence of the method is achieved.

Example 4. Consider the nonlinear SPFIDE from [6] as

$$\begin{cases} -\varepsilon y''(t) + (2 - \exp(-t))y(t) + \frac{1}{2}\int_0^1 \left(\exp(t\cos(\pi s)) - 1\right) y(s)ds = \frac{1}{1+t}, \\ y(0) = 1 \quad y(1) = 0, \end{cases}$$

Table 3: L_2 error and order of the method (11) for Example 3.

N	$\alpha = 0.25, \beta = 3$	Order	$\alpha = 0.5, \beta = 3$	Order	$\alpha = 0.75, \beta = 4$	Order
4	9.9948e-01	—	1.5924e+00	—	1.0037e+01	—
8	4.9934e-01	1.00	2.5881e-01	2.62	2.1306e+00	2.23
16	7.3623e-02	2.76	3.0132e-02	3.10	2.9483e-01	2.85
32	5.6781e-03	3.69	2.9081e-03	3.37	3.1017e-02	3.24
64	5.0210e-04	3.49	2.6321e-04	3.46	2.8995e-03	3.41
128	4.0819e-05	3.62	2.3407e-05	3.49	2.6351e-04	3.45
256	3.4354e-06	3.57	2.0688e-06	3.50	2.3132e-05	3.50

where $\varepsilon \in (0, 1]$ is a perturbation parameter. Since the exact solution to this problem is unknown, we can use the double-mesh principle to estimate the errors and compute numerical solutions [6]. The errors obtained so are denoted by

$$E_\varepsilon^h = \max_k \left| Y_k^{\varepsilon,h} - Y_k^{\varepsilon,h/2} \right|,$$

in which $Y_k^{\varepsilon,h}$ is the corresponding approximate solution with respect to ε and step size h.

In the reported numerical results, we try to compute the estimated convergence rates

$$p_\varepsilon^h = \log_2 \left(\frac{E_\varepsilon^h}{E_\varepsilon^{h/2}} \right), p^h = \log_2 \left(\frac{E^h}{E^{h/2}} \right),$$

where $E^h = \max_\varepsilon E_\varepsilon^h$. The maximum pointwise errors and the rates of convergence $p_\varepsilon^h, p_\varepsilon$ are obtained for the values $\varepsilon = 4^{-j}, j = 0, \ldots, 4$, and $N = 2^l, l = 5, \ldots, 10$, by the proposed method. In this example, the numerical results of the presented method will be compared with those of the numerical reports in [6]. According to the computational results in Table 4, we observe that the presented method is more accurate than that of the method presented in [6]. It can be seen that the numerical results confirm that the methods have achieved the declared order of convergence. Based on Table 4, we can conclude that the order of convergence of the present method is 4 in the L_∞ norm, while the method of [6] is of order 2 with respect to the L_∞ norm.

Example 5. As a final example, we consider the following linear singularly perturbed boundary value problem [21, 16, 13]

$$\begin{cases} \varepsilon y''(t) + y'(t) = 1 + 2t, & 0 \le t \le 1, \\ y(0) = 0, & y(1) = 1. \end{cases}$$

The analytical solution of this problem is

Table 4: L_∞ error and order of the methods for Example 4.

	Present method with $\alpha = 1e - 6$					
ε	$N = 2^5$	$N = 2^6$	$N = 2^7$	$N = 2^8$	$N = 2^9$	$N = 2^{10}$
1	7.024e-07	4.390e-08	2.751e-09	1.719e-10	1.074e-11	6.528e-13
	3.999	3.997	4.000	4.001	4.040	
4^{-1}	2.591e-06	1.594e-07	9.958e-09	6.218e-10	3.888e-11	2.323e-12
	4.023	4.001	4.001	3.999	4.065	
4^{-2}	3.537e-05	2.638e-06	1.727e-07	1.090e-08	6.827e-10	4.264e-11
	3.745	3.933	3.986	3.997	4.001	
4^{-3}	2.305e-04	1.931e-05	1.283e-06	8.137e-08	5.104e-09	3.197e-10
	3.577	3.912	3.979	3.995	3.999	
4^{-4}	4.185e-04	3.902e-05	2.625e-06	1.669e-07	1.048e-08	6.563e-10
	3.423	3.894	3.975	3.993	3.997	
E^h	4.185e-04	3.902e-05	2.625e-06	1.669e-07	1.048e-08	6.563e-10
p^h	3.423	3.894	3.975	3.993	3.997	
	Method of [6]					
ε	$N = 2^5$	$N = 2^6$	$N = 2^7$	$N = 2^8$	$N = 2^9$	$N = 2^{10}$
1	2.882e-02	7.291e-03	1.839e-03	4.624e-04	1.161e-04	2.906e-05
	1.983	1.987	1.992	1.994	1.998	
4^{-1}	2.861e-02	7.251e-03	1.832e-03	4.614e-04	1.507e-04	3.785e-05
	1.98	1.985	1.989	1.991	1.993	
4^{-2}	4.001e-02	1.016e-02	2.575e-03	6.508e-04	1.643e-04	4.139e-05
	1.978	1.98	1.984	1.986	1.989	
4^{-3}	4.331e-02	1.100e-02	2.791e-03	7.070e-04	1.790e-04	4.527e-05
	1.977	1.979	1.981	1.982	1.983	
4^{-4}	4.343e-02	1.105e-02	2.804e-03	7.123e-04	1.809e-04	4.593e-05
	1.975	1.978	1.977	1.977	1.978	
E^h	4.343e-02	1.105e-02	2.804e-03	7.123e-04	1.809e-04	4.593e-05
p^h	1.975	1.978	1.977	1.977	1.978	

$$y(t) = t(t + 1 - 2\varepsilon) + \frac{(2\varepsilon - 1)(1 - \exp(-t/\varepsilon))}{1 - \exp(-1/\varepsilon)}.$$

In this example, we consider the methods [21, 16, 13] to compare the obtained numerical results with the present method. Table 5 contains the computed numerical solution achieved by our method and other methods. From this table, it can be seen that the present method is more accurate than methods [21, 16, 13].

Table 5: Comparison of the approximate solutions of Example 5 for $\varepsilon = h = 10^{-3}$.

t	Method [21]	Method [16]	Method [13]	Present method $(\alpha = 1e-6)$	Analytical solution
0.001	-1.0009970	-0.6311195	-0.6293169	-0.6298615715	-0.6298573177
0.010	-0.9918800	-0.9898546	-0.9878740	-0.9878746043	-0.9878746909
0.020	-0.9815600	-0.9796000	-0.9776400	-0.9776399939	-0.9776399980
0.030	-0.9710400	-0.9691000	-0.9671600	-0.9671599962	-0.9671599999
0.040	-0.9603199	-0.9584000	-0.9564800	-0.9564800000	-0.9564800000
0.050	-0.9493999	-0.9475000	-0.9456000	-0.9456000079	-0.9456000000
0.100	-0.8918000	-0.8900000	-0.8882000	-0.8882000032	-0.8882000000
0.300	-0.6114000	-0.6100000	-0.6086000	-0.6086000008	-0.6086000000
0.500	-0.2510000	-0.2500000	-0.2490000	-0.2490000008	-0.2490000000
0.700	0.1894000	0.1900000	0.1906000	0.1906000004	0.1906000000
0.900	0.7098000	0.7099999	0.7102000	0.7102000001	0.7102000000
1.000	1.0000000	1.0000000	1.0000000	1.0000000	1.0000000

4 Conclusions

In this work, an effective and high-order numerical method for solving a class of nonlinear singular two-point boundary value Fredholm integro-differential equations was presented. After formulation of the method, as well as utilizing an appropriate numerical integration, the original problem was converted to a nonlinear algebraic system. The error analysis was performed to demonstrate the robustness of the method. It was observed that the present methods achieved the order of convergence $O(h^{\min\{\frac{7}{2}, q-\frac{1}{2}\}})$ in the L_2 norm, where q is the order of the quadrature method. Here, some test problems of type SFIDEs and SPFIDEs are solved numerically. Numerical simulations confirmed the theoretical analysis and efficiency of the new method.

Acknowledgments

References

[1] Amiraliyev, G.M., Durmaz, M.E. and Kudu, M. *A numerical method for a second order singularly perturbed Fredholm integro-differential equation*, Miskolc Math. Notes 22(1) (2021) 37–48.

[2] Chambre, P.L. *On the solution of the Poisson-Boltzmann equation with application to the theory of thermal explosions,* J. Chem. Phys. 20 (1952) 1795–1797.

[3] Chandrasekhar, S. *Introduction to the Study of Stellar Structure,* Dover, New York, 1967.

[4] Chen, J., He, M.F. and Huang, Y. *A fast multiscale Galerkin method for solving second order linear fredholm integro-differential equation with Dirichlet boundary conditions,* J. Comput. Appl. Math. 364 (2020) 112352.

[5] Duggan, R.C. and Goodman, A.M. *Pointwise bounds for a nonlinear heat conduction model of the human head,* Bull. Math. Biol. 48 (2) (1986) 229–236.

[6] Durmaz, M.E. and Amiraliyev, G.M. *A robust numerical method for a singularly perturbed Fredholm integro-differential equation,* Mediterr. J. Math. 18(1) (2021), Paper No. 24, 17 pp.

[7] Durmaz, M.E., Amiraliyev, G.M. and Kudu, M. *Numerical solution of a singularly perturbed Fredholm integro differential equation with Robin boundary condition,* Turk. J. Math. 46 (1) (2022) 207–224.

[8] Durmaz, M.E., Cakır, M. and Amirali, G. *Parameter uniform second-order numerical approximation for the integro-differential equations involving boundary layers,* Commun. Fac. Sci. Univ. Ank. Ser. A1 Math. Stat. 71(4) (2022) 954–967.

[9] Flagg, R.C., Luning, C.D. and Perry, W.L. *Implementation of new iterative techniques for solutions of Thomas-Fermi and Emden-Fowler equations,* J. Comput. Phys. 38 (3) (1980) 396–405.

[10] Gebeyehu, M., Garoma, H. and Deressa, A. *Fitted numerical method for singularly perturbed semilinear three-point boundary value problem,* Iranian Journal of Numerical Analysis and Optimization, 12(1) (2022) 145–162.

[11] Gümgüm, S. *Taylor wavelet solution of linear and nonlinear Lane-Emden equations,* Appl. Numer. Math. 158 (2020) 44–53.

[12] Khan, H. and Xu, H. *Series solution to the Thomas-Fermi equation,* Phys. Lett. A 365 (2007) 111–115.

[13] Khan, S. and Khan, A. *A fourth-order method for solving singularly perturbed boundary value problems using nonpolynomial splines,* Iranian Journal of Numerical Analysis and Optimization, 12 (2) (2022) 483–497.

[14] Khuri, S.A. and Sayfy, A. *A novel approach for the solution of a class of singular boundary value problems arising in physiology,* Math. Comput. Model 52 (3) (2010) 626–636.

[15] Kilicman, A., Hashim, I., Tavassoli Kajani, M. and Maleki, M. *On the rational second kind Chebyshev pseudospectral method for the solution of the Thomas-Fermi equation over an infinite interval,* J. Comput. Appl. Math. 257 (7) (2014) 79–85.

[16] Kumar, M., Mishra, H.K. and Singh, P. *A boundary value approach for a class of linear singularly perturbed boundary value problems,* Adv. Eng. Softw. 10 (2009) 298–304.

[17] Lima, P.M. and Morgado, L. *Numerical modeling of oxygen diffusion in cells with Michaelis-Menten uptake kinetics,* J. Math. Chem. 48 (2010) 145–158.

[18] Lin, S.H. *Oxygen diffusion in a spherical cell with nonlinear oxygen uptake kinetics,* J. Theor. Biol. 60 (2) (1976) 449–457.

[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, 12 (2) (2022) 250–276.

[20] Priyadarshana, S., Sahu, S.R. and Mohapatra, J. *Asymptotic and numerical methods for solving singularly perturbed differential difference equations with mixed shifts,* Iranian Journal of Numerical Analysis and Optimization, 12 (1) (2022) 55–72.

[21] Reddy, Y.N. and Chakravarthy, P.P. *An initial-value approach for solving singularly perturbed two-point boundary value problems,* Appl. Math. Comput. 155 (2004) 95–110.

[22] Singh, O.P., Pandey, R.K. and Singh, V.K. *An analytic algorithm of Lane-Emden type equations arising in astrophysics using modified Homotopy analysis method,* Comput. Phys. Commun. 180 (2009) 1116–1124.

[23] Woldaregay, M.M. and Duressa, G.F. *Exponentially fitted tension spline method for singularly perturbed differential difference equations,* Iranian Journal of Numerical Analysis and Optimization, 11 (2) (2021) 261–282.

[24] Zhang, F. *Matrix theory,* Springer, New York, 2011.

How to cite this article

An optimal control approach for solving an inverse heat source problem applying shifted Legendre polynomials

T. Shojaeizadeh*and M. Darehmiraki

Abstract

This study addresses the inverse issue of identifying the space-dependent heat source of the heat equation, which is stated using the optimal control framework. For the numerical solution of this class of problems, an approach based on shifted Legendre polynomials and the associated operational matrix is presented. The approach turns the primary problem into the solution of a system of nonlinear algebraic equations. To do this, the temperature and heat source variables are enlarged in terms of the shifted Legendre polynomials with unknown coefficients employed in the objective function, inverse problem, and initial and Neumann boundary conditions. When paired with their operational matrix, these basis functions provide a quadratic optimization problem with linear constraints, which is then solved using the Lagrange multipliers approach. To assess the method's efficacy and precision, two examples are provided.

―――――――――――――

*Corresponding author

Tahereh Shojaeizadeh
Department of Mathematics, Qom Branch, Islamic Azad University, Qom, Iran.
e-mail: Ta.Shojaeizadeh@iau.ac.ir

Majid Darehmiraki
Department of Mathematics, Behbahan Khatam Alanbia University of Technology, Khouzestan, Iran.
e-mail: darehmiraki@bkatu.ac.ir

1 Introduction

Many inverse problems for the heat equation are applied to many fields of physics and engineering, such as acoustics [19], medical imaging [10], signal processing [35], optic [6], and radar [7]. There are approximately five major classes of inverse heat diffusion equation problems.

(**i**) The problem of reverse time or conducting heat backward from the known last-minute distribution determines the initial temperature distribution.

(**ii**) Inverse heat conduction is the detection of temperature or temperature flux at one inaccessible boundary beyond the data available in the other case that is accessible.

(**iii**) Identify coefficients of over-posed data at the boundaries.

(**iv**) Determining the shape of unknown boundaries or cracks inside the heat conduction body.

(**v**) The identification of the heat source [18, 5].

The heat equation, in this research, treats the heat source as an uncertainty. Applications in the real world where these difficulties are useful include creating the end state of melting, and freezing processes and determining the contaminating source's intensity. Methods such as the generalized finite difference scheme [12], the radial basis function method [28], the sparse regularization approach [25], the meshless generalized finite difference scheme [13], the mollification regularization scheme [36], and the reproducing kernel space scheme [33], have all been applied to the solution of inverse heat source problems. In this paper, we propose a novel numerical method for obtaining the source parameter (or control parameter) in parabolic equations. Iterative methods and a variational approach have recently been proposed to numerically solve this problem [21]. These methods are computationally expensive because they solve a direct problem at each iteration. Tikhonov regularization is proposed in [37] as a stable optimal control solution to the inverse heat source problem. Parameter identification for a nonlinear heat equation in the 2D and 3D space-time domains was solved by Lin and Liu using homogenization functions as the basis [24]. The authors of [29] proposed a perfect method to investigate inverse heat source problems in functionally graded materials using the homogenization function. Due to the given conditions, a homogenization function for the boundary value problem is conceived, and a family of homogenization functions is further derived. Djennadi et al. [11] employed the expansion method and the overdetermination condition to solve the inverse source fractional diffusion problem that contains the Atangana–Baleanu–Caputo fractional derivative. In [20], for the stable reconstruction of the heat source in the parabolic heat equation, an iterative variable conjugate gradient algorithm is proposed based on a sequence of direct problems that are solved using the boundary element method of each iteration step. The gradient descent along with the finite difference method to find the solution nonlinear inverse heat transfer problem in [4]. Ciofalo [9] proposed using finite volume discretization to get a solution for

an inverse heat conduction problem. In which, with the assumption that the thermal boundary conditions in other walls are known, the steady state distribution of the displacement heat transfer coefficient on one slab wall is reconstructed from the temperature distribution in the plate embedded in the slab. With the increasing use of machine learning techniques, including neural networks, the use of these techniques in solving inverse problems has also attracted the attention of many researchers. Li and Hu [23] used a multi-layer neural network to solve the Cauchy inverse problem. Phisyc-informed neural network models are one of the powerful methods in deep learning. Authors in [27] applied it to solve a class of inverse problems related to partial differential equations (PDEs). The authors of [15] proposed a new method for solving large-scale inverse problems based on Bayesian inference, Markov chain Monte Carlo approach, and derivative-free algorithms. Bondarenko [8] presented a finite-difference-based method to investigate the discrete systems of the inverse of the Sturm–Liouville problem. Huntul [16] used the Tikhonov regularization and the nonlinear optimization for the first time in the third-order pseudo-parabolic equation with initial and nonlocal periodic boundary conditions derived from nonlocal integral observation for the inverse space-dependent heat problem. Huntul [17] recovered a source in a high-order pseudo parabolic equation using cubic spline functions. In [14], authors solved the two-dimensional inverse time-fractional diffusion problem with nonlocal boundary conditions using a-polynomials, collocation, and least squares methods. They calculated time using the L_1 method. Wen, Liu, and Wang [34] used the Fourier approach to find the source term and starting data in the time-fractional diffusion equation . Abbaszadeh and Dehghan [1] considered the inverse tempered fractional diffusion equation. They used Crank–Nicolson temporal discretization, a modified element-free Galerkin method, and a meshless method to solve the inverse problem.

In this research, we provide a numerical solution for solving the inverse heat source issue in an optimal control setting by using orthogonal polynomials. This piece is an attempt to provide a fresh strategy for addressing the issue of the mysterious heat source. The optimal control issue is reduced to a set of algebraic equations in the suggested approach [26, 32, 30]. This is achieved by approximating the temperature y and the heat source f in $\mathbf{P_I}$ (see (1)) with the help of shifted Legendre polynomials (SLPs) and their operational matrix with unknown coefficients. By substituting these approximations for the objective function in the inverse problem, we are able to determine not only the unknown coefficients but also the initial and boundary conditions. To conclude, we utilize Lagrange multipliers to connect the algebraic equation produced from the objective function to the algebraic equations derived from the inverse system and the starting and boundary conditions. Then, we can use the constrained extremum method to solve the resulting algebraic system of equations to find the best solution. The authors of [3] investigated the inverse heat equation problem with variable boundary conditions using a weak solution strategy. The Legendre spectral collocation

method was used to solve a fractional inverse heat conduction problem in [2], where both the temperature function and the boundary heat fluxes were unknown. Following the introduction, the article will focus on five primary sections that together address this inverse problem. In Section 2, we present the optimal control issue and the inverse heat source problem. In Section 3, we describe the SLPs and their characteristics. The problem is resolved in Section 4. In Section 5, we provide numerical examples that demonstrate the effectiveness and precision of the suggested approach. The last part explains the results.

2 Problem statement

Suppose the following inverse problem:
Let us suppose $\Theta := (0,1) \times (0,T)$, $T \geq 1$, one is going to find the temperature z and the heat source f that satisfy (1); that is,

$$
\mathbf{P_I} : \begin{cases}
z_t(x,t) - z_{xx}(x,t) = f(x), & (x,t) \in \Theta, \\[2mm]
z(x,0) = \nu(x), & x \in (0,1), \\[2mm]
z_x(0,t) = g_0(t), \;\; z_x(1,t) = g_1(t), & t \in (0,T).
\end{cases} \tag{1}
$$

The second-order parabolic equation $\mathbf{P_I}$ with sufficiently smooth functions $\nu(x)$ (the initial condition), (Neumann boundary conditions) $g_0(t)$ and $g_1(t)$, forms the governing equations.

Assume that the desired function measured data $h_\epsilon(x)$ (desired function) and the actual data $z(x,T) := h(x)$ meet the following relation:

$$
\|h(x) - h_\epsilon(x)\|_{L^2[0,1]} \leqslant \epsilon, \tag{2}
$$

where ϵ is the known noise level and the norm $\|\cdot\|_{L^2[0,1]}$ of a function $z(x)$ is determined by

$$
\|z(x)\|_{L^2[0,1]} = \left(\int_0^1 z^2(x) dx \right)^{\frac{1}{2}}.
$$

In the following part, we convert the problem $\mathbf{P_I}$ into an optimal control problem of $\mathbf{P_{II}}$ and solve it using the suggested approach. The following is a consideration of the optimal control problem:

$$
\mathbf{P_{II}} : \min_{f \in F_{ad}} \quad J(z,f) := \frac{1}{2} \| z(x,T) - h_\varepsilon(x) \|^2_{L^2[0,1]} + \frac{\sigma}{2} \| \nabla f \|^2_{L^2[0,1]}, \tag{3}
$$

where F_{ad} has defined the set of admissible controls of the objective function J as

$$F_{ad} = \{f(x) : 0 \leqslant a \leq f \leq b, \nabla f \in L^2[0,1]\}, \tag{4}$$

with the constant bounds, $a, b \in \mathbb{R}$. Moreover, $z(x,t)$ is the solution of (1) for a given heat source $f(x) \in F_{ad}$, and σ is the regularization parameter. For noisy data $h_\varepsilon(x)$, the purpose of the optimal control problem is to find functions $f(x)$ and $z(x,t)$ that minimize the objective function $\mathbf{P_{II}}$ and satisfy $\mathbf{P_I}$.

3 Shifted Legendre Polynomials (SLPs)

The orthogonal polynomials with regard to the weight function $W(x) = 1$ on $[-1, 1]$ are known as Legendre polynomials of degree m and are denoted by $L_m(x)$ ($m = 0, 1, \ldots$). The following recurrence formula can be used to create these polynomials:

$$L_m(x) = \frac{2m+1}{m+1}xL_m(x) - \frac{m}{m+1}L_{m-1}(x), \qquad m = 1, 2, \ldots, \tag{5}$$

where $L_0(x) = 1$ and $L_1(x) = x$. The well-known SLPs in $[0, 1]$ can be created by changing the variable $x = 2t-1$, which is expressed as $\mathcal{L}_m(t)$ ($m = 0, 1, 2, \ldots,$), by

$$\mathcal{L}_m(t) = \frac{(2m+1)(2t-1)}{m+1}\mathcal{L}_m(t) - \frac{m}{m+1}\mathcal{L}_{m-1}(t), \qquad m = 1, 2, \ldots, \tag{6}$$

where $\mathcal{L}_0(t) = 1$ and $\mathcal{L}_1(t) = 2t - 1$. The explicit formula of the SLPs is as follows [31]:

$$\mathcal{L}_m(t) = \sum_{i=0}^{m} b_{mi}t^i, \tag{7}$$

where $\mathcal{L}_m(0) = (-1)^m$, $\mathcal{L}_m(1) = 1$, and

$$b_{mi} = (-1)^{m+i}\frac{(m+i)!}{(m-i)!(i!)^2}. \tag{8}$$

The orthogonality condition of the SLPs with respect to the weight function $w(t) = 1$ is given by

$$\int_0^1 \mathcal{L}_m(t)\mathcal{L}_n(t)dt = h_m\delta_{mn}, \tag{9}$$

where δ_{mn} is Kronecker's delta function and $h_m = \dfrac{1}{2m+1}$. Any given function $z(t) \in L^2[0,1]$ can be represented in $(n+1)$ terms of the SLPs as

$$z(t) \simeq \sum_{i=0}^{n} z_i \mathcal{L}_i(t) \triangleq Z^T \Phi_n(t), \tag{10}$$

where

$$Z = [z_0 \ z_1 \ \ldots \ z_n]^T,$$

$$\Phi_n(t) \triangleq [\mathcal{L}_0(t) \ \mathcal{L}_1(t) \ \ldots \ \mathcal{L}_n(t)]^T, \tag{11}$$

and

$$z_i = \frac{1}{h_i} \int_0^1 z(t)\mathcal{L}_i(t)dt, \qquad i = 0, 1, \ldots, n.$$

In a similar way, a two-variable function $z(x,t) \in L^2(\Theta)$ can be expanded by the SLPs as

$$z(x,t) \simeq \sum_{i=0}^{m} \sum_{j=0}^{n} z_{ij}\mathcal{L}_i(x)\mathcal{L}_j(t) \triangleq \Phi_m^T(x)Z\Phi_n(t), \tag{12}$$

where $Z = [z_{ij}]$ is the matrix of coefficients with dimensions $(m+1) \times (n+1)$ whose entries are unknown and obtained from the following equation:

$$z_{ij} = \frac{1}{h_i h_j} \int_0^1 \int_0^1 z(x,t)\mathcal{L}_i(x)\mathcal{L}_j(t)dxdt, \tag{13}$$

for $i = 0, 1, \ldots, m, \ j = 0, 1, \ldots, n$. Suppose that $\Phi_n(t)$ is the vector introduced in (11). Then the derivative of this vector is as follows: [31]

$$\frac{d\Phi_n(t)}{dt} = D_t^{(1)}\Phi_n(t), \tag{14}$$

where $D_t^{(1)} = [d_{ij}^{(1)}]$ is called the derivative operational matrix of SLPs of $(n+1)$-order, whose structure is as follows:

$$d_{ij}^{(1)} = \begin{cases} 2(2j+1), & j = i-k, \begin{cases} k = 1, 3, \ldots, n & if \ n \ odd, \\ k = 1, 3, \ldots, n-1 & if \ n \ even, \end{cases} \\ \\ 0 & \text{otherwise.} \end{cases} \tag{15}$$

Remark 1. Generally, the r-derivative operational matrix of SLPs of $\Phi_n(t)$ can be given by [31]

$$\frac{d^r \Phi_n(t)}{dt^r} = D_t^{(r)} \Phi_n(t), \tag{16}$$

in which $D_t^{(r)}$ is obtained by r times multiplying $D_t^{(1)}$ in itself.

4 Convergence analysis

In this section, the convergence analysis of SLPs expansion in two dimensions is investigated.

Theorem 1. Suppose that $z : \Theta \longrightarrow \mathbb{R}$ is $(n + m + 1)$ times continuously differentiable. If $\Phi_m^T(x)Z\Phi_n(t)$ is a unique best approximation of z, then the following inequality holds:

$$\|z(x,t) - \Phi_m^T(x)Z\Phi_n(t)\|_{L^2(\Theta)} \leqslant \frac{\Delta\sqrt{\Gamma}(n+m+2)}{r!(n+m+1-r)!}, \tag{17}$$

where

$$\Delta = \max_{\Theta}\{\left| \frac{\partial^{n+m+1}}{\partial x^{n+m+1-i}\partial t^i}z(x,t) \right| \mid i = 0, 1, \ldots, n+m+1\},$$
$$\Gamma = \max_{T \geqslant 1}\{T^{2n+2m-i+3}, \quad i = 0, 1, \ldots, 2(n+m+1)\}.$$

Proof. Maclaurin's expansion for $z(x,t)$ reads as follows:

$$z(x,t) = p(x,t) + \frac{1}{(n+m+1)!}(x\frac{\partial}{\partial x} + t\frac{\partial}{\partial t})^{n+m+1}z(\xi_0 x, \xi_0 t), \quad \xi_0 \in (0,1),$$

where

$$p(x,t) = \sum_{k=0}^{n+m} \frac{1}{k!}(x\frac{\partial}{\partial x} + t\frac{\partial}{\partial t})^k z(0,0).$$

Thus

$$|z(x,t) - p(x,t)| = |\frac{1}{(n+m+1)!}(x\frac{\partial}{\partial x} + t\frac{\partial}{\partial t})^{n+m+1}z(\xi_0 x, \xi_0 t)|, \quad \xi_0 \in (0,1).$$

On the other hand, since $\Phi_m^T(x)Z\Phi_n(t)$ is the best approximation of $z(x,t)$, we obtain

$$\| z - \Phi_m^T Z\Phi_n \|^2_{L^2(\Theta)} \leq \| z - p \|^2_{L^2(\Theta)}.$$

By the definition of the L^2-norm and expand $(x\frac{\partial}{\partial x} + t\frac{\partial}{\partial t})^{n+m+1}$, we have

$$\| z(x,t) - p(x,t) \|_{L^2(\Theta)}^2$$

$$= \int_0^T \int_0^1 [\frac{1}{(n+m+1)!}(x\frac{\partial}{\partial x} + t\frac{\partial}{\partial t})^{n+m+1} z(\xi_0 x, \xi_0 t)]^2 dxdt$$

$$= \int_0^T \int_0^1 [\frac{1}{(n+m+1)!} \sum_{i=0}^{n+m+1} \binom{n+m+1}{i} x^{n+m+1-i} t^i$$

$$\frac{\partial^{n+m+1}}{\partial x^{n+m+1-i} \partial t^i} z(\xi_0 x, \xi_0 t)]^2 dxdt$$

$$\leq \frac{\Delta^2}{(n+m+1)!^2} \int_0^T \int_0^1 [\sum_{i=0}^{n+m+1} \binom{n+m+1}{r} x^{n+m+1-i} t^i]^2 dxdt,$$

where

$$\binom{n+m+1}{r} = \max \left\{ \binom{n+m+1}{i} ; i = 0, 1, \ldots, n+m+1 \right\}.$$

To find the upper bound for the above expression, we calculate the following terms:

$$\int_0^T \int_0^1 x^{2(n+m+1-i)} t^{2i} dxdt = \frac{T^{1+2i}}{(1+2i)(2n+2m-2i+3)},$$

$$i = 0, 1, \ldots, n+m+1,$$

$$\int_0^T \int_0^1 x^{(2n+2m+1-i)} t^{i+1} dxdt = \frac{T^{2+i}}{(2+i)(2n+2m-i+2)},$$

$$i = 0, 1, \ldots, n+m,$$

$$\vdots$$

$$\int_0^T \int_0^1 x^{(2+i)} t^{2n+2m-i} dxdt = \frac{T^{2n+2m-i+1}}{(3+i)(2n+2m-i+1)}, \qquad i = 0, 1,$$

$$\int_0^T \int_0^1 x^{(1+i)} t^{2n+2m-i+1} dxdt = \frac{T^{2n+2m-i+2}}{(2+i)(2n+2m-i+2)}, \qquad i = 0.$$

Therefore

$$\| z - p \|^2_{L^2(\Theta)} \leq \frac{\Delta^2}{r!^2(n+m+1-r)!^2} \int_0^T \int_0^1 [\sum_{i=0}^{n+m+1} x^{n+m+1-i} t^i]^2 dxdt$$

$$= \frac{\Delta^2}{r!^2(n+m+1-r)!^2} [\sum_{i=0}^{n+m+1} \frac{T^{1+2i}}{(1+2i)(2n+2m-2i+3)}$$

$$+ \sum_{i=0}^{n+m} \frac{T^{2+i}}{(2+i)(2n+2m-i+2)} + \cdots$$

$$+ \sum_{i=0}^{1} \frac{T^{2n+2m-i+1}}{(3+i)(2n+2m+1)} + \frac{T^{2n+2m-i+2}}{2(2n+2m+2)} \Bigg]$$

$$\leq \frac{\Delta^2 \Gamma}{r!^2 (n+m+1-r)!^2}$$

$$\times \big[(n+m+2) + (n+m+1) + \cdots + 2 + 1 \big]$$

$$\leq \frac{\Delta^2 \Gamma}{r!^2 (n+m+1-r)!^2} (n+m+2)^2,$$

which is the desired result. $\qquad \square$

5 Description of the presented method

Now in this section we will use numerical methods to address the problem raised in (1) and (3). We will use numerical methods to address the problem raised in (1) and (3) in this section. SLPs approximate the temperature and heat source for this purpose as follows:

$$z(x,t) \simeq \Phi_m(x)^T Z \Phi_n(t), \tag{18}$$

$$f(x) \simeq F^T \Phi_m(x), \tag{19}$$

where Z and F are the following unknown matrices of coefficients with dimensions $(m+1) \times (n+1)$ and $(m+1) \times 1$, respectively, while $\Phi_m(x)$ and $\Phi_n(t)$ in (11) have been expressed:

$$Z = \begin{pmatrix} z_{00} & z_{01} & \dots & z_{0n} \\ z_{10} & z_{11} & \dots & z_{1n} \\ \vdots & \vdots & & \vdots \\ z_{m0} & z_{m1} & \dots & z_{mn} \end{pmatrix}, \qquad F = \begin{pmatrix} f_0 \\ f_1 \\ \vdots \\ f_m \end{pmatrix}. \tag{20}$$

Set

$$\mathcal{P}(x,t) \triangleq [\mathcal{L}_0(x)\mathcal{L}_0(t), \ \dots, \ \mathcal{L}_m(x)\mathcal{L}_0(t) \ | \ \cdots \ | \ \mathcal{L}_0(x)\mathcal{L}_n(t), \ \dots, \ \mathcal{L}_m(x)\mathcal{L}_n(t)]. \tag{21}$$

According to (21), we can express (18) as

$$z(x,t) \simeq \Phi_m(x)^T Z \Phi_n(t) = \mathcal{P}(x,t) \operatorname{vec}(Z), \tag{22}$$

where

$$\operatorname{vec}(Z) = [z_{00}, \dots, z_{m0} \ | \ \dots \ | \ z_{0n}, \dots, z_{mn}]^T.$$

From (14), (22), and Remark 1, the result will be as follows:

$$z_x(x,t) \simeq \Phi_m(x)^T D_x^{(1)T} Z \Phi_n(t) = \mathcal{P}(x,t)(I_{n+1} \otimes D_x^{(1)T}) \operatorname{vec}(Z), \qquad (23)$$

$$z_{xx}(x,t) \simeq \Phi_m(x)^T D_x^{(2)T} Z \Phi_n(t) = \mathcal{P}(x,t)(I_{n+1} \otimes D_x^{(2)T}) \operatorname{vec}(Z), \qquad (24)$$

$$z_t(x,t) \simeq \Phi_m(x)^T Z D_t^{(1)} \Phi_n(t) = \mathcal{P}(x,t)(D_t^{(1)T} \otimes I_{m+1}) \operatorname{vec}(Z), \qquad (25)$$

so that I_{m+1} and I_{n+1} are identity matrices of orders $m+1$ and $n+1$, respectively. Additionally, $\otimes$ refers to the Kronecker product [22]. Now, (19), (24), and (25) are substituted into the first subequation of (1), and the result is

$$\mathcal{K}(x,t) \operatorname{vec}(Z) - F^T \Phi_m(x) = 0, \qquad (26)$$

in which

$$\mathcal{K}(x,t) \triangleq \mathcal{P}(x,t) \left[(D_t^{(1)T} \otimes I_{m+1}) - (I_{n+1} \otimes D_x^{(2)T}) \right].$$

Thus, as to (22) and (23) and with regards to initial and Neumann boundary conditions in (1), we have

$$\mathcal{P}(x,0) \operatorname{vec}(Z) = \nu(x),$$

$$\mathcal{P}(0,t)(I_{n+1} \otimes D_x^{(1)T}) \operatorname{vec}(Z) = g_0(t), \qquad (27)$$

$$\mathcal{P}(1,t)(I_{n+1} \otimes D_x^{(1)T}) \operatorname{vec}(Z) = g_1(t).$$

We follow the suggested procedure by constructing an $(m+1) \times (n+1)$ algebraic system of equations. For this reason, we derive the following equations from (26) and (27):

$$\begin{cases} \mathcal{K}(\xi_i, \eta_j) \operatorname{vec}(Z) - F^T \Phi_m(\xi_i) = 0, & i = 2,\ldots,m, \;\; j = 2,\ldots,n+1, \\[2mm] \mathcal{P}(\xi_i, 0) \operatorname{vec}(Z) = \nu(\xi_i) & i = 1,\ldots,m+1, \\[2mm] \mathcal{P}(0,\eta_j)(I_{n+1} \otimes D_x^{(1)T}) \operatorname{vec}(Z) = g_0(\eta_j), & j = 2,\ldots,n+1, \\[2mm] \mathcal{P}(1,\eta_j)(I_{n+1} \otimes D_x^{(1)T}) \operatorname{vec}(Z) = g_1(\eta_j), & j = 2,\ldots,n+1, \end{cases}$$

$$(28)$$

where a collocation scheme is defined by evaluating the outcome at the points (ξ_i, η_j) in (28). We employ the shifted Legendre–Gauss–Lobatto nodes ξ_i ($1 \leq i \leq m+1$) and the shifted Legendre roots η_j ($1 \leq j \leq n+1$) of $\mathcal{L}_n(t)$ to find suitable collocation points. It is possible to write (28) as follows:

$$\mathcal{M} \operatorname{vec}(Z) - \mathcal{N} \hat{\mathcal{F}} = Q, \tag{29}$$

in which

$$\mathcal{M} = \begin{pmatrix} \mathcal{T}(2:n+1,:) \otimes \mathcal{X}(2:m,:) \\ \mathcal{T}(1,:) \otimes \mathcal{X}(1:m+1,:) \\ \mathcal{T}(2:n+1,:) \otimes \mathcal{X}(1,:) \\ \mathcal{T}(2:n+1,:) \otimes \mathcal{X}(m+1,:) \end{pmatrix}, \qquad \mathcal{N} = \begin{pmatrix} \mathcal{S} \otimes \mathcal{X}(2:m,:) \\ 0 \\ 0 \\ 0 \end{pmatrix},$$

$$\hat{\mathcal{F}} = \left[f_0, \ldots, f_m | \underbrace{0, \ldots, 0}_{m+1} | \cdots | \underbrace{0, \ldots, 0}_{m+1} \right]^T,$$

$$Q = \left[\overbrace{\underbrace{0, \ldots, 0}_{m-1} | \cdots | \underbrace{0, \ldots, 0}_{m-1}}^{n} | \nu(\xi_1), \ldots, \nu(\xi_{m+1}) \right.$$

$$\left. | g_0(\eta_2), \ldots, g_0(\eta_{n+1}) | g_1(\eta_2), \ldots, g_1(\eta_{n+1}) \right]^T,$$

where $\hat{F}$ and Q are $(m+1)(n+1)$-order vectors and

$$\mathcal{S} \triangleq \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 \end{pmatrix}_{n \times (n+1)},$$

$$\mathcal{T} \triangleq \begin{pmatrix} \mathcal{L}_0(\eta_1) & \mathcal{L}_1(\eta_1) & \ldots & \mathcal{L}_n(\eta_1) \\ \mathcal{L}_0(\eta_2) & \mathcal{L}_1(\eta_2) & \ldots & \mathcal{L}_n(\eta_2) \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{L}_0(\eta_{n+1}) & \mathcal{L}_1(\eta_{n+1}) & \ldots & \mathcal{L}_n(\eta_{n+1}) \end{pmatrix}_{(n+1) \times (n+1)},$$

$$\mathcal{X} \triangleq \begin{pmatrix} \mathcal{L}_0(\xi_1) & \mathcal{L}_1(\xi_1) & \ldots & \mathcal{L}_m(\xi_1) \\ \mathcal{L}_0(\xi_2) & \mathcal{L}_1(\xi_2) & \ldots & \mathcal{L}_m(\xi_2) \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{L}_0(\xi_{m+1}) & \mathcal{L}_1(\xi_{m+1}) & \ldots & \mathcal{L}_m(\xi_{m+1}) \end{pmatrix}_{(m+1) \times (m+1)}.$$

Next, we approximate $\mathbf{P_{II}}$ by SLPs. First, we approximate the desired function $h_\varepsilon(x)$ as

$$h_\varepsilon(x) \simeq H^T \Phi_m(x). \tag{30}$$

From (14) and (19), we have

$$\nabla f(x) \simeq F^T D_x^{(1)} \Phi_m(x). \tag{31}$$

Inserting (18), (30), and (31) into (3) yields that

$$
\begin{aligned}
J(z, f) &\simeq J_{m,n}(Z, F) \\
&= \frac{1}{2} \int_0^1 (\Phi_m(x)^T Z \Phi_n(T) - \Phi_m(x)^T H)(\Phi_m(x)^T Z \Phi_n(T) - \Phi_m(x)^T H)^T dx \\
&\quad + \frac{\sigma}{2} \int_0^1 (F^T D_x^{(1)} \Phi_m(x))(F^T D_x^{(1)} \Phi_m(x))^T dx.
\end{aligned}
$$

The value $\int_0^1 (\phi_m(x)^T H)^2 dx$ is positive, meaning it has no influence on minimization and according to (9), the equation can be expressed as follows:

$$
\begin{aligned}
J_{m,n}(Z, F) &= \frac{1}{2} \mathrm{vec}(Z)^T (\Phi_n(T) \Phi_n(T)^T \otimes \Upsilon_m) \, \mathrm{vec}(Z) \\
&\quad - H^T (\Phi_n(T)^T \otimes \Upsilon_m) \, \mathrm{vec}(Z) \\
&\quad + \frac{\sigma}{2} F(I_{n+1} \otimes (D_x^{(1)} \Upsilon_m D_x^{(1)T})) F, \tag{32}
\end{aligned}
$$

where

$$\Upsilon_m = diag(h_0, \dots, h_m).$$

The problem of optimal control in the discussion has now become a finite dimension optimization. We use the Lagrangian multipliers method to solve the ensuing optimization problem. Let us clarify

$$J^*(z, f) \simeq J^*(Z, F, \Omega) = J_{m,n}(Z, F) + \Lambda^T (\mathcal{M} \, \mathrm{vec}(Z) - \mathcal{N} \hat{\mathcal{F}} - Q), \tag{33}$$

where

$$\Lambda = \begin{bmatrix} \lambda_1 \ \lambda_2 \ \dots \ \lambda_{(m+1) \times (n+1)} \end{bmatrix}^T,$$

which shows the vector of Lagrange multipliers as Λ. The following equations lead to the following optimality conditions:

$$
\begin{cases}
\dfrac{\partial J^*(z, f)}{\partial \, \mathrm{vec}(Z)} = 0, \\[2mm]
\dfrac{\partial J^*(z, f)}{\partial F} = 0, \\[2mm]
\dfrac{\partial J^*(z, f)}{\partial \Lambda} = 0.
\end{cases}
$$

The Newton iterative approach or MATLAB software tools can be used to solve the aforementioned algebraic equation system. We can get the approximate solutions $z(x,t)$ and $f(x)$ from (18) and (19), respectively, by figuring out Z and F.

6 Numerical examples

This section gives two examples along with figures to illustrate how the recommended technique may be implemented successfully and to show its potential. The results of the existing plan are analyzed and compared to the solution that was found analytically and method of [37]. The rand function is used by the MATLAB software to generate noisy data, and the value of h_ε for $0 \le \delta \le 1$ in the collocation points $\{\xi_j\}_{j=1}^{m+1}$ is calculated as follows:

$$h_\varepsilon = h + \delta.rand(size(h)), \tag{34}$$

$$\varepsilon = \|h_\varepsilon - h\|_{l^2} = \left(\frac{1}{m+1} \sum_{j=1}^{m+1} |h_\varepsilon - h|^2\right)^{\frac{1}{2}}. \tag{35}$$

For noisy data $h_\varepsilon(x)$, the goal of the optimal control problem is to find functions $f(x)$ and $z(x,t)$ that minimize the following objective function and satisfy (1):

$$\min J(z,f) = \frac{1}{2} \parallel z(x,1) - h_\varepsilon(x) \parallel^2_{L^2[0,1]} + \frac{\sigma}{2} \parallel f'(x) \parallel^2_{L^2[0,1]}$$
$$= \frac{1}{2} \int_0^1 |z(x,1) - h_\varepsilon(x)|^2 dx + \frac{\sigma}{2} \int_0^1 |f'(x)|^2 dx. \tag{36}$$

We take the regularization parameter $\sigma = \varepsilon^2$, and, in order to observe the convergence of the method described in numerical experiments, we calculate the approximate error resulting from the following equation:

$$e(f) = \parallel \widetilde{f} - f \parallel_{L_\infty}, \tag{37}$$

where $\widetilde{f}$ is the numerical approximation of the exact solution f in the collocation points $\{\xi_i\}_{i=1}^{m+1}$.

Example 1. Consider the inverse problem with $\Theta = (0,1) \times (0,1)$ [37]

$$\begin{cases} z_t(x,t) - z_{xx}(x,t) = f(x), & (x,t) \in \Theta, \\ \\ z(x,0) = 0, & x \in (0,1), \\ \\ z_x(0,t) = z_x(1,t) = 0, & t \in (0,1). \end{cases} \tag{38}$$

We attempt to approximate the heat source defined by

$$f(x) = \pi^2 \cos(\pi x).\tag{39}$$

Then with f given by (39), the forward problem presented by (38) has an analytical solution as follows:

$$z(x, t) = \sum_{n=1}^{\infty} \frac{1 - e^{-(n\pi)^2 t}}{(n\pi)^2} f_n \cos(n\pi x),\tag{40}$$

where f_n is the Fourier coefficient as follows:

$$f_n = 2 \int_0^1 f(x) \cos(n\pi x) dx.\tag{41}$$

From (40), we have

$$h(x) = z(x, 1) = \sum_{n=1}^{\infty} \frac{1 - e^{-(n\pi)^2}}{(n\pi)^2} f_n \, cos(n\pi x).\tag{42}$$

Table 1: Comparison of errors estimate obtained for functions f in Example 1 over a range of σ values between the proposed method and [37]

	$\sigma = 10^{-4}$	$\sigma = 10^{-5}$	$\sigma = 10^{-6}$
Proposed method	$6.1387e - 01$	$6.8907e - 02$	$1.2413e - 02$
Method of [37]	$2.511e - 01$	$4.81e - 02$	$3.26e - 02$

Table 1 analyzes the error behavior of the proposed method in here and the presented method in [37] by varying the value of σ.

Table 2: Errors estimate for the functions f and z in Example 1 over a range of σ values

	$\sigma = 10^{-4}$	$\sigma = 10^{-5}$	$\sigma = 10^{-6}$	$\sigma = 10^{-7}$
Eroor(f)	$6.1387e - 01$	$6.8907e - 02$	$1.2413e - 02$	$8.3175e - 03$
Eroor(z)	$5.5543e - 02$	$6.0304e - 03$	$6.0868e - 04$	$6.3734e - 05$

The approximate solutions for the functions f and z are shown in Figure 1. The approximation inaccuracy are shown in Table 2. Figure 2 depicts the convergence of the suggested approach.
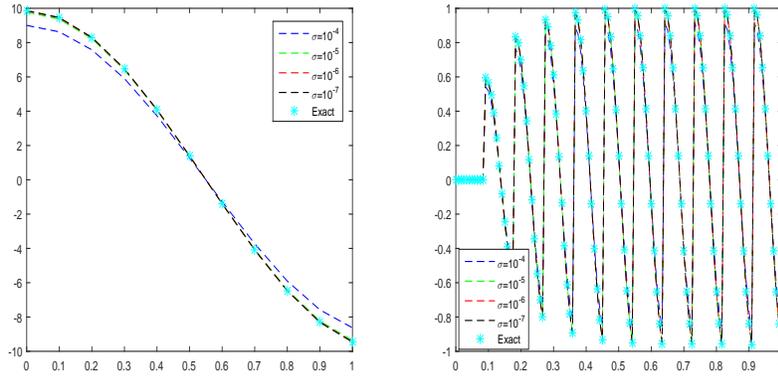
Figure 1: Results of Example 1's numerical solutions for functions f (left) and z (right) for a range of σ values
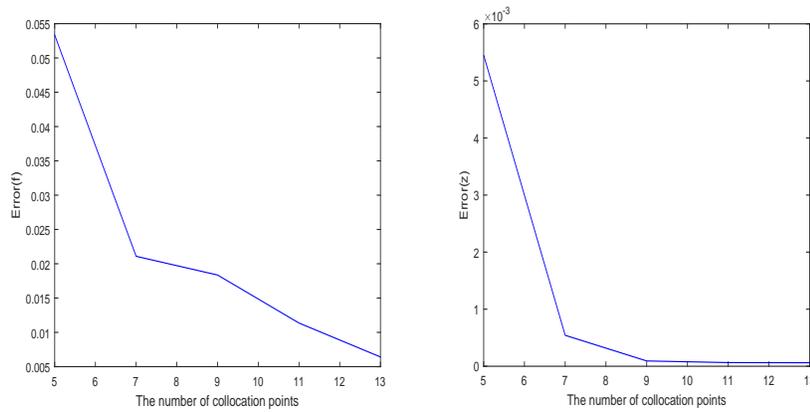
Figure 2: Convergence of the numerical solutions of Example 1 for functions f (left) and z (right) for a range of collocation point values

Example 2. Consider the inverse problem with $\Theta = (0,1) \times (0,1)$ [28]

$$
\begin{cases}
z_t(x,t) - z_{xx}(x,t) = f(x), & (x,t) \in \Theta, \\[2mm]
z(x,0) = \sin(\pi x), & x \in (0,1), \\[2mm]
z_x(1,t) = -z_x(0,t) = \pi(e^{-\pi^2 t} - 2), & t \in (0,1).
\end{cases}
\tag{43}
$$

We attempt to approximate the heat source defined by

$$f(x) = 2\pi^2 \sin(\pi x). \tag{44}$$

Then with f given by (44), the forward problem presented by (43) has an analytical solution as follows:

$$z(x,t) = -(e^{-\pi^2 t} - 2)\sin(\pi x). \tag{45}$$

From (45), we have

$$h(x) = z(x,1) = -(e^{-\pi^2} - 2)\sin(\pi x). \tag{46}$$

Table 3: Values of errors for the functions f and z with different values of σ in Example 2

	$\sigma = 10^{-5}$	$\sigma = 10^{-6}$	$\sigma = 10^{-7}$	$\sigma = 10^{-8}$
Eroor(f)	$4.3462e - 00$	$1.2586e - 00$	$7.4287e - 01$	$6.8760e - 01$
Eroor(z)	$9.7406e - 02$	$1.6259e - 02$	$1.7443e - 03$	$1.7581e - 04$

Figure 3 shows the approximate solutions for the functions f and z. The approximation error is presented in Table 3. The convergence of the proposed method can be seen in Figure 4.
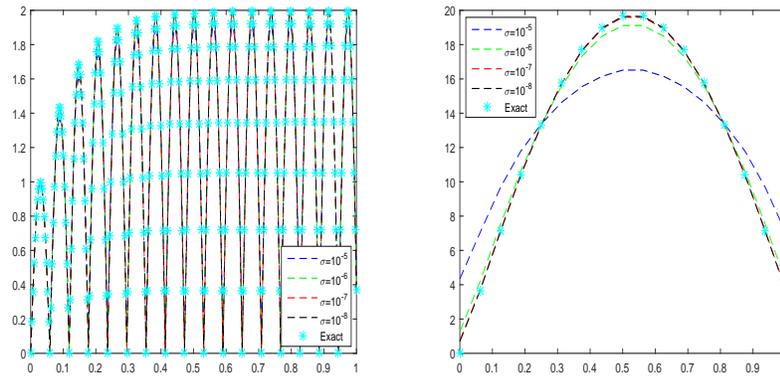
Figure 3: Behavior of the numerical solutions for the functions f (right) and z (left) at some different values of σ in Example 2
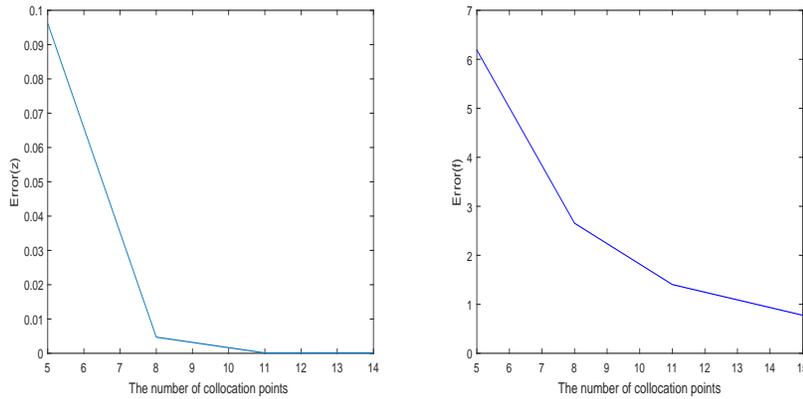
Figure 4: Convergence of the numerical solutions for the functions f (right) and z (left) at some different values of collocation points in Example 2

7 Conclusion

When it comes to finding a regular and stable solution, inverse problems that are related to PDEs provide a significant computing challenge since it is very difficult to do so. The scope of this work is an investigation into an inverse space-dependent source issue for a heat equation. A shifted Legendre polynomial and an optimum control strategy were used in the process of creating a heat source. One of the most popular and efficient tools for resolving computing problems is the Legendre polynomial. The shifted Legendre polynomials operational matrix was utilized to resolve this optimal control problem. By applying the suggested collocation method and using an operational matrix, the issue was converted into a set of equations that can be solved using algebra. When utilizing this method to solve an inverse problem, as demonstrated by the examples provided in the paper, a high level of precision was achieved in the solution. The method presented here was based on the optimal control problem and shifted Legendre polynomials. In future work, we will try to use machine learning techniques, including deep neural networks, to solve this problem.

Acknowledgements

References

[1] Abbaszadeh, M. and Dehghan, M. *Numerical and analytical investigations for solving the inverse tempered fractional diffusion equation via interpolating element-free Galerkin (IEFG) method*, J. Therm. Anal. Calorim. 143(3) (2021), 1917–1933.

[2] Abdelkawy, M.A., Babatin, M.M., Alnahdi, A.S. and Taha, T.M. *Legendre spectral collocation technique for fractional inverse heat conduction problem*, Int. J. Mod. Phys. C. 33(05) (2022), 2250065.

[3] Ait Ben Hassi, E.M., Chorfi, S.E. and Maniar, L. *Identification of source terms in heat equation with dynamic boundary conditions*, Math. Methods Appl. Sci. 45(4) (2022), 2364–2379.

[4] Alpar, S. and Rysbaiuly, B. *Determination of thermophysical characteristics in a nonlinear inverse heat transfer problem*, Appl. Math. Comput. 440 (2023), 127656.

[5] Badia, A.E. and Duong, T.H. *On an inverse source problem for the heat equation: application to a detection problem*, J. Inverse Ill-Posed Probl. 10 (2002), 585–589.

[6] Bal, G. and Schotland, J.C. *Inverse scattering and acousto-optic imaging*, Phys. Rev. Lett. 104(4) (2010), 043902.

[7] Baraniuk, R. and Steeghs, P. *Compressive radar imaging*, Radar conference, Boston (MA): IEEE (2007), 128–133.

[8] Bondarenko, N.P. *Finite-difference approximation of the inverse Sturm–Liouville problem with frozen argument*, Appl. Math. Comput. 413 (2022), 126653.

[9] Ciofalo, M. *Solution of an inverse heat conduction problem with third-type boundary conditions*, Int. J. Therm. Sci. 175 (2022), 107466.

[10] Crossen E., Gockenbach M.S., Jadamba, B., Khan, A.A. and Winkler, B. *An equation error approach for the elasticity imaging inverse problem for predicting tumor location*, Comput. Math. Appl. 67(1) (2014), 122–135.

[11] Djennadi, S., Shawagfeh, N., Osman, M.S., Gomez-Aguilar, J.F. and Arqub, O.A. *The Tikhonov regularization method for the inverse source problem of time fractional heat equation in the view of ABC-fractional technique*, Phys. Scr. 96(9) (2021), 094006.

[12] Gu, Y., Lei, J., Fan, C.M. and He, X.Q. *The generalized finite difference method for an inverse time-dependent source problem associated with three-dimensional heat equation*, Eng. Anal. Bound. Elem. 91 (2018), 73–81.

[13] Gu, Y., Wang, L., Chen, W., Zhang, C. and He, X. *Application of the meshless generalized finite difference method to inverse heat source problems*, Int. J. Heat Mass Transf. 108 (2017), 721–729.

[14] Hajishafieiha, J. and Abbasbandy, S. *Numerical solution of two-dimensional inverse time-fractional diffusion problem with non-local boundary condition using a-polynomials*, J. Appl. Math. Comput. 69(2) (2023), 1945–1965.

[15] Huang, D.Z., Huang, J., Reich, S. and Stuart, A.M. *Efficient derivative-free Bayesian inference for large-scale inverse problems*, Inverse Probl. 38(12) (2022), 125006.

[16] Huntul, M.J. *Space-dependent heat source determination problem with nonlocal periodic boundary conditions*, Results Appl. Math. 12 (2021), 100223.

[17] Huntul, M.J. *Recovering a source term in the higher-order pseudo-parabolic equation via cubic spline functions*, Phys. Scr. 97(3) (2022), 035004.

[18] Ikehata, M. *An inverse source problem for the heat equation and the enclosure method*, Inverse Probl. 23 (2007), 183–202.

[19] Isakov, V. and Wu, S.F. *On theory and application of the Helmholtz equation least squares method in inverse acoustics*, Inverse Probl. 18(4) (2002), 1147.

[20] Johansson, B.T. and Lesnic, D. *A variational method for identifying a spacewise-dependent heat source*, IMA J. Appl. Math. 72(6) (2007), 748–760.

[21] Johansson, T. and Lesnic, D. *Determination of a spacewise dependent heat source*, J. Comput. Appl. Math. 209(1) (2007), 66–80.

[22] Kilicman, A. and Al Zhour, Z.A.A. *Kronecker operational matrices for fractional calculus and some applications*, Appl. Math. Comput. 187(1) (2007), 250–265.

[23] Li, Y. and Hu, X. *Artificial neural network approximations of Cauchy inverse problem for linear PDEs*, Appl. Math. Comput. 414 (2022), 126678.

[24] Lin, J. and Liu, C.S. *Recovering temperature-dependent heat conductivity in 2D and 3D domains with homogenization functions as the bases*, Eng. Comput. 38(3) (2022), 2349–2363.

[25] Lu, Z.R., Pan, T., and Wang, L. *A sparse regularization approach to inverse heat source identification*, Int. J. Heat Mass Transf. 142 (2019), 118430.

[26] Mahmoudi, M., Shojaeizadeh, T. and Darehmiraki, M. *Optimal control of time-fractional convection-diffusion-reaction problem employing compact integrated RBF method*, Math. Sci. (Springer) 17(1) (2023), 1–14.

[27] Mishra, S. and Molinaro, R. *Estimates on the generalization error of physics-informed neural networks for approximating a class of inverse problems for PDEs*, IMA J. Numer. Anal. 42(2) (2022), 981–1022.

[28] Nawaz Khan, M., Ahmad, I. and Ahmad, H. *A radial basis function collocation method for space-dependent inverse heat problems*, J. Appl. Comput. Mech. (2020).

[29] Qiu, L., Lin, J., Wang, F., Qin, Q.H. and Liu, C.S. *A homogenization function method for inverse heat source problems in 3D functionally graded materials*, Appl. Math. Model. 91 (2021), 923–933.

[30] Rezazadeh, A., Mahmoudi, M. and Darehmiraki, M. *A solution for fractional PDE constrained optimization problems using reduced basis method*, Comput. Appl. Math. 39(2) (2020), 1–17.

[31] Saadatmandi, A. and Dehghan, M. *A new operational matrix for solving fractional-order differential equations*, Comput. Math. Appl. 59(3) (2010), 1326–1336.

[32] Shojaeizadeh, T., Mahmoudi, M. and Darehmiraki, M. *Optimal control problem of advection-diffusion-reaction equation of kind fractal-fractional applying shifted Jacobi polynomials*, Chaos Solit. Fractals. 143 (2021), 110568.

[33] Wang, W., Han, B. and Yamamoto, M. *Inverse heat problem of determining time-dependent source parameter in reproducing kernel space*, Nonlinear Anal. Real World Appl. 14(1) (2013), 875–887.

[34] Wen, J., Liu, Z.X. and Wang, S.S. *Conjugate gradient method for simultaneous identification of the source term and initial data in a time-fractional diffusion equation*, Appl. Math. Sci. Eng. 30(1) (2022), 324–338.

[35] Widrow, B. and Walach, E. *Adaptive inverse control, reissue edition: a signal processing approach*, John Wiley and Sons, 2008.

[36] Yang, F. and Fu, C.L. *A mollification regularization method for the inverse spatial-dependent heat source problem*, J. Comput. Appl. Math. 255 (2014), 555–567.

[37] Yang, S. and Xiong, X. *A Tikhonov regularization method for solving an inverse heat source problem*, Bull. Malays. Math. Sci. Soc. 43(1) (2020), 441–452.

How to cite this article
Shojaeizadeh, T. and Darehmiraki, M., An optimal control approach for solving an inverse heat source problem applying shifted Legendre polynomials. *Iran. j. numer. anal. optim.*, 2023; 13(3): 460-480. https://doi.org/10.22067/ijnao.2023.81202.1221

How to cite this article
Research Article

Analysis and optimal control of a fractional MSD model

M. Bagherpoorfard and F. Akhavan Ghassabzade*,

Abstract

In this research, we aim to analyze a mathematical model of Maize streak virus disease as a problem of fractional optimal control. For dynamical analysis, the boundedness and uniqueness of solutions have been investigated and proven. Also, the basic reproduction number is obtained, and local stability conditions are given for the equilibrium points of the model. Then, an optimal control strategy is proposed for the purpose of examining the best strategy to fight the maize streak disease. We solve the fractional optimal control problem by a forward-backward sweep iterative algorithm. In this algorithm, the state variable is obtained in a forward and co-state variable by a backward method where an explicit Runge-Kutta method is used to solve differential equations arising from fractional optimal control problems. Some comparative results are presented in order to verify the model and show the efficacy of the fractional optimal control treatments.

1 Introduction

Maize is an important annual cereal crop of the world belonging to the family Poaceae. It is considered a staple food in many parts of the world. It

---

*Corresponding author

Mina Bagherpoorfard
Department of Mathematics, Fasa Branch, Islamic Azad university, Fasa, Iran.
e-mail: bagherpoorfard@yahoo.com

Fahimeh Akhavan Ghassabzade
Department of Mathematics, Faculty of sciences, University of Gonabad, Gonabad, Iran.
e-mail: akhavan_gh@yahoo.com

is the third leading crop in the world after rice and wheat [20]. Due to its highest yield potential among cereals, it is known globally as the queen of cereals. Maize streak disease (MSD) is the most serious viral crop disease in Sub-Saharan Africa. This disease is caused by the Maize streak virus (MSV), which was first described by the South African entomologist Claude Fuller in 1901 [12]. MSV is mainly transmitted by as many as six leafhopper species in the Genus Cicadulina, but some other leafhopper species are also able to transmit the virus. In addition to maize, this virus can infect over 80 other species in the Family Poaceae. Severe MSD manifests as pronounced, continuous parallel chlorotic streaks on leaves, with severe stunting of the affected plant and, usually, a failure to produce complete cobs or seeds. Erratic epidemics have been occurring every 3-10 years, and the main damage caused is to plants younger than six weeks old [24].

In recent years, mathematical modeling has become a valuable tool to study the mechanisms of plant disease spread, predict the future course of an outbreak, and appraise strategies to control. In most cases, differential equations of the integer order have been used to construct such models; see, for example, [25, 8, 9, 14] and the references therein. The integer-order derivatives and integrals have local properties; that is, the next state is not influenced by the current and previous state. So, the integer-order mathematical models can not describe natural phenomena precisely.

Fractional calculus is an extension of classical calculus that introduces derivatives and integrals of fractional order. Fractional derivatives have nonlocal properties, that is the next state depends on the current state and all previous states. This is the main excellence of fractional derivatives over classical derivatives. Due to this advantage, many applications of fractional calculus can be found in various fields of research, such as biology, economy, physics, control theory, and so on [15, 13, 21, 22, 26, 2]. In [23], a fractional model of tuberculosis disease has presented, and the values of parameters have been evaluated according to the actual clinical cases. In 2020, the dynamics of the fractional HIV infection model were studied by Evirgen Evirgen, Uçar, and Özdemir [11]. Bozkurt et al., in their work [6], have analyzed a fractional model of COVID-19 by considering the fear effects of the media and social networks. In [5], the authors presented a fractional model for the simulation of the Cholera outbreak in Yemen. The authors in [4] proposed a fractional model to study the dynamics of the MSV in the maize plant population by considering the interaction of MSV pathogen with the past invasion.

In light of this significant advantage, we were motivated to develop the model investigated in [3] into a new fractional model involving the Caputo derivatives. The Caputo derivative is of use for modeling phenomena that take account of interactions within the past and also problems with nonlocal properties. In this sense, one can think of the equation as having "memory." After that, we discuss some properties of the fractional version of the model under consideration. Next, fractional optimal control (FOC) is applied as a generalization of the classical optimal control system [3]. The FOC model is

developed with three time-dependent control strategies proposed by Alemneh, Kassa, and Godana [3].

The paper is organized as follows. In section 2, we give a brief review of the Caputo operator and discuss its basic characteristics. In Section 3, the fractional-order model formulation is presented, and the main properties of the fractional model are then given in section 4. Section 5 focuses on the dynamic analysis of the model. The FOC of the model and numerical simulations of the fractional model are presented in section 6. Section 7 also contains concluding remarks.

2 Basic definitions and facts

In this section, we give a brief review of the Caputo operator and discuss its basic characteristics [16].

Definition 1. For a function $f : [0, t_f] \to \mathbb{R}, \nu \in (n-1, n)$, and $n \in \mathbb{N}$, the left- and the right-sided Caputo fractional derivatives of order ν of a function f are defined in the following forms:

$$
{}^{C}_{0}\mathfrak{D}^{\nu}_t f(t) = \frac{1}{\Gamma(n-\nu)} \int_0^t (t-u)^{(n-\nu-1)} f^{(n)}(u) du, \quad t > 0, \tag{1}
$$

and

$$
{}^{C}_{t_f}\mathfrak{D}^{\nu}_t f(t) = \frac{(-1)^n}{\Gamma(n-\nu)} \int_t^{t_f} (t-u)^{(n-\nu-1)} f^{(n)}(u) du, \quad t < t_f. \tag{2}
$$

Here, $\Gamma(\cdot)$ denotes the Gamma function.

Definition 2. The integral operators related to (1) and (2), are specified by

$$
{}^{C}_{0}\mathfrak{I}^{\nu}_t f(t) = \frac{1}{\Gamma(\nu)} \int_0^t (t-\eta)^{\nu-1} f(\eta) d\eta, \tag{3}
$$

$$
{}^{C}_{t}\mathfrak{I}^{\nu}_{t_f} f(t) = \frac{1}{\Gamma(\nu)} \int_t^{t_f} (\eta-t)^{\nu-1} f(\eta) d\eta. \tag{4}
$$

Additionally, if $f \in C^n[a, b]$, then

$$
{}^{C}_{0}\mathfrak{I}^{\nu}_t [{}^{C}_{0}\mathfrak{D}^{\nu}_t f(t)] = f(t) - \sum_{k=0}^{n-1} \frac{f^{(k)}(0)}{k!} t^k, \tag{5}
$$

$$
{}^{C}_{t}\mathfrak{I}^{\nu}_{t_f} [{}^{C}_{t}\mathfrak{D}^{\nu}_{t_f} f(t)] = f(t) - \sum_{k=0}^{n-1} (-1)^k \frac{f^{(k)}(t_f)}{k!} (t_f - t)^k. \tag{6}
$$

For any $\alpha_1, \alpha_2 \in \mathbb{R}$ and $f_1, f_2 \in \mathbb{H}^1(0, t_f)$, we have

$$
{}^{C}_{0}\mathfrak{D}^{\nu}_t (\alpha_1 f_1(t) + \alpha_2 f_2(t)) = \alpha_1 \, {}^{C}_{0}\mathfrak{D}^{\nu}_t f_1(t) + \alpha_2 \, {}^{C}_{0}\mathfrak{D}^{\nu}_t f_2(t), \tag{7}
$$

$$\begin{aligned}
{}_{t}^{C}\mathfrak{D}_{t_f}^{\nu}(\alpha_1 f_1(t) + \alpha_2 f_2(t)) &= \alpha_1 \, {}_{t}^{C}\mathfrak{D}_{t_f}^{\nu} f_1(t) + \alpha_2 \, {}_{t}^{C}\mathfrak{D}_{t_f}^{\nu} f_2(t), \quad (8)
\end{aligned}$$

$$\begin{aligned}
{}_{0}^{C}\mathfrak{I}_{t}^{\nu}(\alpha_1 f_1(t) + \alpha_2 f_2(t)) &= \alpha_1 \, {}_{0}^{C}\mathfrak{I}_{t}^{\nu} f_1(t) + \alpha_2 \, {}_{0}^{C}\mathfrak{I}_{t}^{\nu} f_2(t), \quad (9)
\end{aligned}$$

$$\begin{aligned}
{}_{t}^{C}\mathfrak{I}_{t_f}^{\nu}(\alpha_1 f_1(t) + \alpha_2 f_2(t)) &= \alpha_1 \, {}_{t}^{C}\mathfrak{I}_{t_f}^{\nu} f_1(t) + \alpha_2 \, {}_{t}^{C}\mathfrak{I}_{t_f}^{\nu} f_2(t). \quad (10)
\end{aligned}$$

Let $f(t)$ be a constant function. Then

$$
{}_{0}^{C}\mathfrak{D}_{t}^{\nu} f(t) = {}_{t}^{C}\mathfrak{D}_{t_f}^{\nu} f(t) = 0. \quad (11)
$$

The Caputo derivatives satisfy the Lipschitz condition.

3 New fractional model of MSV disease in maize plant

In this section, we develop a deterministic eco-epidemiological fractional model for the dynamics of MSV disease in maize plants. The original version of this model is a system of ordinary differential equations that have been before presented in [3]. The effect of previous states in the current states of the disease spread has not been considered in this model. One way to overcome this drawback is to replace the integer-order derivatives in the model with noninteger-order derivatives [19]. Hence, we replace the ordinary derivative with the following Caputo fractional derivative operator

$$
\frac{d}{dt} \longrightarrow \frac{1}{\varrho^{1-\nu}} {}_{0}^{C}\mathfrak{D}_{t}^{\nu}, \quad (12)
$$

where the auxiliary parameter $\varrho > 0$ represents the fractional time components in the system. Thus, the new model is described by the system

$$
\begin{cases}
\varrho^{\nu-1} {}_{0}^{C}\mathfrak{D}_{t}^{\nu} S(t) = rS(1 - \dfrac{S+I}{K}) - \dfrac{\beta_1 SY}{A+S}, \\[2mm]
\varrho^{\nu-1} {}_{0}^{C}\mathfrak{D}_{t}^{\nu} I(t) = \dfrac{\beta_1 SY}{A+S} - \mu_1 I, \\[2mm]
\varrho^{\nu-1} {}_{0}^{C}\mathfrak{D}_{t}^{\nu} H(t) = q - \dfrac{\beta_2 IH}{C+I} - \mu_2 H, \\[2mm]
\varrho^{\nu-1} {}_{0}^{C}\mathfrak{D}_{t}^{\nu} Y(t) = \dfrac{b\beta_2 IH}{C+I} - \mu_3 Y,
\end{cases} \quad (13)
$$

$$
S(0) = S_0, \ I(0) = I_0, \ H(0) = H_0, \ Y(0) = Y_0, \quad (14)
$$

where $0 < \alpha \le 1, N_1(t) = S(t) + I(t), N_2(t) = H(t) + Y(t)$, and $(S, I, H, Y) \in \mathbb{R}_+^4$. In this model, $S(t)$ denotes the density of the susceptible maize, and $I(t)$ denotes the density of the infected maize. The susceptible and infected leafhopper vector densities are denoted by $H(t)$ and $Y(t)$, respectively. All parameters in the model are nonnegative. Description of the parameters are found in Table 1.

Table 1: Explanation of MSV model parameters

Parameter	Explanation
β_1	Predation and infection rate of infected leafhopper on susceptible maize plant
β_2	Predation and infection rate of susceptible leafhopper on infected maize plant
A	The half-saturation rate of susceptible maize with infected plant
C	The half-saturation rate of susceptible leafhopper with infected maize plant
K	Carrying capacity
q	Recruitment rate of susceptible leafhopper
b	Infected leafhopper conversion rate
r	Maize population intrinsic growth rate
μ_1	Death rate of infected maize
μ_2	Death rate of susceptible leafhopper
μ_3	Infected leafhopper death rate

As can be observed, model (13) involves a system of nonlinear fractional differential equations. The exact solution of this model may not be available in general. However, a mathematical analysis of the existence and uniqueness of the solution ensures that a unique solution exists under some conditions.

4 Properties of the model

In the following, the model's main properties are provided. Our model can be formulated as

$$
\begin{aligned}
{}_0^C\mathfrak{D}_t^\nu V(t) &= \Phi(t, V(t)), \\
V(0) &= V_0,
\end{aligned}
\tag{15}
$$

where $V(t) = (S(t), I(t), H(t), Y(t))$.

Lemma 1. [17] Let $w(t)$ be a continuous function on $[t_0, \infty)$ and satisfying

$$
\begin{cases}
{}_0^C\mathfrak{D}_t^\nu w(t) \leq -\lambda w(t) + \mu, \\
w(t_0) = w_0,
\end{cases}
\tag{16}
$$

where $0 < \nu < 1, (\lambda, \mu) \in \mathbb{R}^2$, $\lambda \neq 0$, and $t_0 \geq 0$ is the initial time. Then

$$
w(t) \leq (w_0 - \frac{\mu}{\lambda})E_\nu[-\lambda(t - t_0)^\nu] + \frac{\mu}{\lambda},
$$

where E_ν represents Mittag–Leffler function.

Lemma 2. [7] Let $0 < \nu < 1$ and $\lambda < 0$. Then $E_{\nu,\nu}(\lambda t^\nu)$ tends monotonically to zero as $t \to \infty$.

Lemma 3. [10] Let $\Phi : [t_0, \infty) \times \mathbb{R}^n \to \mathbb{R}^n$ be a continuous function and Lipschitz-continuous respecting to the second variable. In addition to, let $\nu \in (0, 1]$ and $V_0 \in \mathbb{R}^n$. Then, the problem

$$\begin{aligned}
{}_{0}^{C}\mathfrak{D}_{t}^{\nu}V(t) &= \Phi(t, V(t)), \quad t > t_0, \\
V(t_0) &= V_0,
\end{aligned} \quad (17)$$

has a unique solution in $C([0, \infty); \mathbb{R}^n)$.

Theorem 1. All solutions of system (13) that initiate in $\mathbb{R}_+^4$ are bounded within the region Ω defined by

$$\Omega = \{(S, I, H, Y) \in \mathbb{R}_+^4 \,|\, S(t) + I(t) + H(t) + \tfrac{1}{b}Y(t) \leq \frac{L}{\rho} + \varepsilon, \text{ for all } \varepsilon > 0\}.$$

Proof. Define a time-dependent function $w(t) = S(t) + I(t) + H(t) + \tfrac{1}{b}Y(t)$. So, for any positive number ρ, we have

$$\begin{aligned}
{}_{0}^{C}&\mathfrak{D}_{t}^{\nu}w(t) + \rho w(t) \\
&= rS(1 - \frac{S + I}{K}) - \mu_1 I - \mu_2 H - \frac{\mu_3}{b}Y + \rho S + \rho I + \rho H + \frac{\rho}{b}Y + q \\
&\leq rS(1 - \frac{S}{K}) + (\rho - \mu_1)I + (\rho - \mu_2)H + (\rho - \mu_3)\frac{1}{b}Y + q \\
&= (r + \rho)S - \frac{r}{K}S^2 + (\rho - \mu_1)I + (\rho - \mu_2)H + (\rho - \mu_3)\frac{1}{b}Y + q \\
&\leq \frac{K}{4r}(r + \rho)^2 + (\rho - \mu_1)I + (\rho - \mu_2)H + (\rho - \mu_3)\frac{1}{b}Y + q.
\end{aligned}$$

Taking $\rho < \min(\mu_1, \mu_2, \mu_3)$, so

$$ {}_{0}^{C}\mathfrak{D}_{t}^{\nu}w(t) + \rho w(t) \leq L, $$

where $L = \frac{K}{4r}(r + \rho)^2 + q$. Now, we apply Lemma 1 and obtain

$$ w(t) \leq (w(0) - \frac{L}{\rho})E_\nu[-\rho t^\nu] + \frac{L}{\rho}. $$

Thus, $w(t) \to \dfrac{L}{\rho}$ as $t \to \infty$ and $0 < w(t) \leq \dfrac{L}{\rho}$. Hence all solutions of system (13) that starts from $\mathbb{R}_+^4$ are confined in the region $\Omega = \{(S, I, H, Y) \in \mathbb{R}_+^4 \,|\, w(t) \leq \dfrac{L}{\rho} + \varepsilon, \text{ for all } \varepsilon > 0\}$. $\square$

Now, we study the existence and uniqueness of system (13) in the region $\Lambda \times [0, T]$, where

$$ \Lambda = \{(S, I, H, Y)\epsilon\mathbb{R}^4 : \max(|S|, |I|, |H|, |Y|) \leq M\}, $$

$T < \infty$ and M is sufficiently large.

Theorem 2. For any nonnegative initial conditions, system (13) has a unique solution.

Proof. Let $X = (S, I, H, Y)$. Consider a mapping

$$Q(X) = (Q_1(X), Q_2(X), Q_3(X), Q_4(X)),$$

where

$$Q_1(X) = rS(1 - \frac{S+I}{K}) - \frac{\beta_1 SY}{A+S},$$

$$Q_2(X) = \frac{\beta_1 SY}{A+S} - \mu_1 I,$$

$$Q_3(X) = q - \frac{\beta_2 IH}{C+I} - \mu_2 H,$$

$$Q_4(X) = \frac{b\beta_2 IH}{C+I} - \mu_3 Y.$$

For any $X, \overline{X} \in \Lambda$, we have

$$\|Q(X) - Q(\overline{X})\|$$
$$= |Q_1(X) - Q_1(\overline{X})| + |Q_2(X) - Q_2(\overline{X})| + |Q_3(X) - Q_3(\overline{X})| + |Q_4(X) - Q_4(\overline{X})|$$
$$= |rS(1 - \frac{S+I}{K}) - \frac{\beta_1 SY}{A+S} - r\overline{S}(1 - \frac{\overline{S}+\overline{I}}{K}) + \frac{\beta_1 \overline{S}.\overline{Y}}{A+\overline{S}}|$$
$$\quad + |\frac{\beta_1 SY}{A+S} - \mu_1 I - \frac{\beta_1 \overline{S}.\overline{Y}}{A+\overline{S}} + \mu_1 \overline{I}| + |q - \frac{\beta_2 IH}{C+I} - \mu_2 H - q + \frac{\beta_2 \overline{I}.\overline{H}}{C+\overline{I}} + \mu_2 \overline{H}|$$
$$\quad + |\frac{b\beta_2 IH}{C+I} - \mu_3 Y - \frac{b\beta_2 \overline{I}.\overline{H}}{C+\overline{I}} + \mu_3 \overline{Y}|$$
$$= |r(S - \overline{S}) - \frac{r}{K}(S^2 - \overline{S}^2) - \frac{r}{K}(SI - \overline{S}.\overline{I}) - \beta_1(\frac{SY}{A+S} - \frac{\overline{S}.\overline{Y}}{A+\overline{S}})|$$
$$\quad + |\beta_1(\frac{SY}{A+S} - \frac{\overline{S}.\overline{Y}}{A+\overline{S}}) - \mu_1(I - \overline{I})| + |-\beta_2(\frac{IH}{C+I} - \frac{\overline{I}.\overline{H}}{C+\overline{I}}) - \mu_2(H - \overline{H})|$$
$$\quad + |b\beta_2(\frac{IH}{C+I} - \frac{\overline{I}.\overline{H}}{C+\overline{I}}) - \mu_3(Y - \overline{Y})|$$
$$\leq (r + \frac{3rM}{K})|S - \overline{S}| + 2\beta_1|\frac{SY}{A+S} - \frac{\overline{S}.\overline{Y}}{A+\overline{S}}| + \beta_2(1+b)|\frac{IH}{C+I} - \frac{\overline{I}.\overline{H}}{C+\overline{I}}|$$
$$\quad + \mu_1|I - \overline{I}| + \mu_2|H - \overline{H}| + \mu_3|Y - \overline{Y}|$$
$$\leq (r + \frac{3rM}{K} + \frac{2\beta_1 M}{A})|S - \overline{S}| + (\frac{2\beta_1(A+M)M}{A^2} + \mu_3)|Y - \overline{Y}|$$
$$\quad + (\frac{\beta_2 M}{C}(1+b) + \mu_1)|I - \overline{I}| + (\frac{\beta_2 M(C+M)}{C^2}(1+b) + \mu_2)|H - \overline{H}|$$
$$\leq H\|X - \overline{X}\|,$$

where

$$H = \max\{r + \frac{3rM}{K} + \frac{2\beta_1 M}{A}, \frac{2\beta_1(A+M)M}{A^2} + \mu_3,$$
$$\frac{\beta_2 M(1+b)}{C} + \mu_1, \frac{\beta_2 M(C+M)(1+b)}{C^2} + \mu_2\}.$$

Thus, $Q(X)$ satisfies the Lipschitz condition with respect to X. Hence there exists a unique solution of the system (13) with conditions (14) on $\Lambda \times [0, T]$. $\square$

5 Dynamical behaviors

One of the key concepts in epidemiology is the basic reproduction number (BRN). The aim of this section is to obtain the BRN for model (13) and study the local stability behavior of the model at its disease-free equilibriums.

5.1 Basic reproduction number

Consider the following fractional differential system:

$$
\begin{cases}
\varrho^{\nu-1}{}_0^C\mathfrak{D}_t^\nu X(t) = F(X, Y), \\
\varrho^{\nu-1}{}_0^C\mathfrak{D}_t^\nu Y(t) = G(X, Y), \\
G(X, 0) = 0,
\end{cases}
\tag{18}
$$

with nonnegative initial conditions $X(0) = X_0 \in \mathbb{R}^2$ and $Y(0) = Y_0 \in \mathbb{R}^2$, where the components of vector $X = (S, H)$ represent the number of susceptible maize and leafhopper, and the components of vector $Y = (I, Y)$ indicate the number of infected maize and leafhopper. Furthermore, we presume that the function G is of class C^1, F is continuous, and the system (18) with the initial conditions $X(0) = X_0$ and $Y(0) = Y_0$ admits a unique solution. Also, suppose that $E = (X^*, 0) \in \mathbb{R}^4$ denotes the disease-free equilibrium point of the system (18). Let $A = \frac{\partial G}{\partial Y}(X^*, 0) = M - D$, where M, D are two square matrices that $D > 0$ is a diagonal matrix and $M \geq 0$. Then the BRN $\mathcal{R}_0$ is obtained as the spectral radius of MD^{-1}.
For system (13), we have

$$
A = \begin{bmatrix} -\mu_1 & \frac{\beta_1 S}{A+S} \\ \frac{b\beta_2 HC}{(C+I)^2} & -\mu_3 \end{bmatrix} = \begin{bmatrix} 0 & \frac{\beta_1 S}{A+S} \\ \frac{b\beta_2 HC}{(C+I)^2} & 0 \end{bmatrix} - \begin{bmatrix} \mu_1 & 0 \\ 0 & \mu_3 \end{bmatrix}
\tag{19}
$$

So,

$$
\mathcal{R}_0 := \rho\left(\begin{bmatrix} 0 & \frac{\beta_1 S}{\mu_3(A+S)} \\ \frac{b\beta_2 HC}{\mu_1(C+I)^2} & 0 \end{bmatrix} \right) = \sqrt{\frac{b\beta_1\beta_2 HCS}{\mu_1\mu_3(A+S)(C+I)^2}}.
\tag{20}
$$

5.2 Local stability analysis

Theorem 3. The disease free-equilibrium point $E_0 = (0, 0, \frac{q}{\mu_2}, 0)$ of system (13) is always unstable while the disease-free equilibrium point $E_1 = (K, 0, \frac{q}{\mu_2}, 0)$, is locally asymptotically stable if $\mathcal{R}_0 < 1$.

Proof. Conforming to Mittag-Leffler function [18], the disease free equilibrium E of system (13) is locally asymptotically stable if all eigenvalues $\lambda_i, i = 1, 2, 3, 4$ of J_E satisfy $|arg(\lambda_i)| > \frac{\nu\pi}{2}, i = 1, 2, 3, 4$. The Jacobian matrix associated to E_0 is given by

$$J_{E_0} = \begin{bmatrix} r & 0 & 0 & 0 \\ 0 & -\mu_1 & 0 & 0 \\ 0 & \frac{-\beta_2 q}{C\mu_2} & -\mu_2 & 0 \\ 0 & \frac{b\beta_2 q}{C\mu_2} & 0 & -\mu_3 \end{bmatrix}.$$

The eigenvalues of the matrix J_{E_0} are $\lambda_1 = r > 0$, $\lambda_2 = -\mu_1 < 0$ $\lambda_3 = -\mu_2 < 0$ $\lambda_4 = -\mu_3 < 0$. We observed that $|arg(\lambda_1)| = 0 < \frac{\nu\pi}{2}$. So, the equilibrium point E_0 is unstable.

The Jacobian matrix associated to E_1 is the following one:

$$J_{E_1} = \begin{bmatrix} -r & -r & 0 & \frac{-\beta K}{A+K} \\ 0 & -\mu_1 & 0 & \frac{\beta K}{A+K} \\ 0 & \frac{-\beta_2 q}{C\mu_2} & -\mu_2 & 0 \\ 0 & \frac{b\beta_2 q}{C\mu_2} & 0 & -\mu_3 \end{bmatrix}.$$

The following characteristic equation is obtained from J_{E_1}:

$$\phi(\lambda) = (r + \lambda)(\mu_2 + \lambda)\left(\lambda^2 + (\mu_1 + \mu_3)\lambda + (\mu_1\mu_3 - \frac{Kbq\beta_1\beta_2}{(A+K)\mu_2 C})\right).$$

We observe that two roots of the characteristic equation $\phi(\lambda)$ are

$$\lambda_1 = -r < 0, \quad \lambda_2 = -\mu_2 < 0.$$

It is obvious that $|arg(\lambda_1)| > \frac{\nu\pi}{2}$ and $|arg(\lambda_2)| > \frac{\nu\pi}{2}$. The remaining eigenvalues are given by

$$\lambda^2 + (\mu_1 + \mu_3)\lambda + (\mu_1\mu_3 - \frac{Kbq\beta_1\beta_2}{(A+K)\mu_2 C}) = 0. \tag{21}$$

By the Routh–Hurwitz criteria, all the roots of the polynomial (21) are negative or have negative real part if and only if

$$\mu_1\mu_3 - \frac{Kbq\beta_1\beta_2}{(A+K)\mu_2 C} > 0,$$

or

$$\mathcal{R}_0 = \sqrt{\frac{Kbq\beta_1\beta_2}{(A+K)\mu_1\mu_2\mu_3 C}} < 1. \tag{22}$$

Hence, E_1 is locally asymptotically stable if $\mathcal{R}_0 < 1$. □

6 Optimal control problem

In this section, to attain the minimized number of infected maize and infected leafhoppers, we reconsider the model (13) and formulate an optimal control problem with three control variables $u_1(t), u_2(t)$, and $u_3(t)$. Let

$$U = \{(u_1, u_2, u_3)|u_1, u_2, \text{ and } u_3 \text{ are Lebesgue measurable on } [0, t_f],$$
$$0 \le u_1, u_2, u_3 \le 1, \text{ for all } t \in [0, t_f]\},$$

be the admissible control set. With the existence of control u_1, it is expected to diminish the number of infected maize as compared to those without control cases. The control variable u_2 is used to control the number of infected leafhoppers. Furthermore, u_3 is chemical control that is used as an intervention strategy to optimize the objective functional $\mathcal{F}$. After incorporating the control variables $u_1(t), u_2(t)$, and $u_3(t)$ in the model (13), the optimal control model is as follows:

$$\begin{cases} \varrho^{\nu-1}{}_0^C\mathfrak{D}_t^\nu S(t) = rS(1 - \dfrac{S+I}{K}) - (1-u_1)\dfrac{\beta_1 SY}{A+S}, \\ \varrho^{\nu-1}{}_0^C\mathfrak{D}_t^\nu I(t) = (1-u_1)\dfrac{\beta_1 SY}{A+S} - (\mu_1 + u_2)I, \\ \varrho^{\nu-1}{}_0^C\mathfrak{D}_t^\nu H(t) = q - (1-u_2)\dfrac{\beta_2 IH}{C+I} - (u_3 + \mu_2)H, \\ \varrho^{\nu-1}{}_0^C\mathfrak{D}_t^\nu Y(t) = (1-u_2)\dfrac{b\beta_2 IH}{C+I} - (u_3 + \mu_3)Y, \\ S(0), I(0), H(0), Y(0) \ge 0. \end{cases} \tag{23}$$

Consider the following objective functional:

$$\mathcal{F} = \int_0^{t_f} \left(d_1 I + d_2 Y + \frac{1}{2}(w_1 u_1^2 + w_2 u_2^2 + w_3 u_3^2)\right) dt, \tag{24}$$

where d_1, d_2 are the weights on the state variables and w_1, w_2, and w_3 are relative weights of the treatment related to the control functions u_1, u_2, and u_3.

Our aim is to minimize the cost value $\mathcal{F}$ by the state and control variables I^*, Y^*, u_1^*, u_2^*, and u_3^* satisfying the constraints (23). For this purpose, we use a kind of Pontryagin maximum principle in the fractional order state [1]. We define the Hamiltonian function as below:

$$\mathfrak{H}(S, I, H, Y) = d_1 I + d_2 Y + \frac{1}{2}\left(w_1 u_1^2 + w_2 u_2^2 + w_3 u_3^2\right)$$

$$+ \varpi_1 \left(rS\left(1 - \frac{S+I}{K}\right) - (1-u_1)\frac{\beta_1 SY}{A+S}\right)$$

$$+ \varpi_2 \left((1-u_1)\frac{\beta_1 SY}{A+S} - (\mu_1 + u_2)I\right)$$

$$+ \varpi_3 \left(q - (1-u_2)\frac{\beta_2 IH}{C+I} - (u_3 + \mu_2)H\right)$$

$$+ \varpi_4 \left((1-u_2)\frac{b\beta_2 IH}{C+I} - (u_3 + \mu_3)Y\right),$$

where $\varpi_i, i = 1, 2, 3, 4$ are the co-state variables or adjoint variables. The optimality conditions are obtained from

$$\frac{\partial \mathfrak{H}}{\partial u_1} = \frac{\partial \mathfrak{H}}{\partial u_2} = \frac{\partial \mathfrak{H}}{\partial u_3} = 0.$$

Hence, we have

$$u_1 = \frac{\beta_2(\varpi_2 - \varpi_1)IH}{w_1(A+S)},$$

$$u_2 = \frac{\varpi_2 I}{w_2} + \frac{\beta_2(b\varpi_4 - \varpi_3)IH}{w_2(C+I)},$$

$$u_3 = \frac{\varpi_4 Y - \varpi_3 H}{w_3}, \tag{25}$$

where the adjoint variables satisfy

$$_t^C\mathfrak{D}_{t_f}^\nu \varpi_1(t) = \frac{\partial \mathfrak{H}}{\partial S} = \left\{r\left(\frac{K - 2S - I}{K}\right) - (1-u_1)\frac{\beta_1 YA}{(A+S)^2}\right\}\varpi_1$$

$$+ \left\{\frac{\beta_1(1-u_1)SY}{(A+S)^2}\right\}\varpi_2,$$

$$_t^C\mathfrak{D}_{t_f}^\nu \varpi_2(t) = \frac{\partial \mathfrak{H}}{\partial I} = d_1 - \frac{rS}{K}\varpi_1 - (\mu_1 + u_2)\varpi_2 - (1-u_2)\frac{\beta_2 CH}{(C+I)^2}\varpi_3$$

$$+ (1-u_2)\frac{\beta_2 bCH}{(C+I)^2}\varpi_4,$$

$$_t^C\mathfrak{D}_{t_f}^\nu \varpi_3(t) = \frac{\partial \mathfrak{H}}{\partial H} = \left\{(u_2 - 1)\frac{\beta_2 I}{C+I} - \mu_2 - u_3\right\}\varpi_3 + \frac{b(1-u_2)\beta_2 I}{C+I}\varpi_4,$$

$$_t^C\mathfrak{D}_{t_f}^\nu \varpi_4(t) = \frac{\partial \mathfrak{H}}{\partial Y} = d_2 + \frac{\beta_1(u_1 - 1)S}{A+S}(\varpi_1 - \varpi_2) - (u_3 + \mu_3)\varpi_4,$$

$$\varpi_1(t_f) = \varpi_2(t_f) = \varpi_3(t_f) = \varpi_4(t_f) = 0. \tag{26}$$

Then, we have the following boundary value problem for optimal treatment:

$$
\begin{cases}
\varrho^{\nu-1}{}_0^C\mathfrak{D}_t^\nu S(t) = rS(1 - \dfrac{S+I}{K}) - (1-u_1)\dfrac{\beta_1 SY}{A+S}, \\[2mm]
\varrho^{\nu-1}{}_0^C\mathfrak{D}_t^\nu I(t) = (1-u_1)\dfrac{\beta_1 SY}{A+S} - (\mu_1+u_2)I, \\[2mm]
\varrho^{\nu-1}{}_0^C\mathfrak{D}_t^\nu H(t) = q - (1-u_2)\frac{\beta_2 IH}{C+I} - (u_3+\mu_2)H, \\[2mm]
\varrho^{\nu-1}{}_0^C\mathfrak{D}_t^\nu Y(t) = (1-u_2)\dfrac{b\beta_2 IH}{C+I} - (u_3+\mu_3)Y, \\[2mm]
{}_t^C\mathfrak{D}_{t_f}^\nu \varpi_1(t) = \{r(\dfrac{K-2S-I}{K}) - (1-u_1)\dfrac{\beta_1 YA}{(A+S)^2}\}\varpi_1 \\[2mm]
\qquad\qquad + \{\dfrac{\beta_1(1-u_1)SY}{(A+S)^2}\}\varpi_2, \\[2mm]
{}_t^C\mathfrak{D}_{t_f}^\nu \varpi_2(t) = d_1 - \frac{rS}{K}\varpi_1 - (\mu_1+u_2)\varpi_2 - (1-u_2)\frac{\beta_2 CH}{(C+I)^2}\varpi_3 \\[2mm]
\qquad\qquad + (1-u_2)\frac{\beta_2 bCH}{(C+I)^2}\varpi_4, \\[2mm]
{}_t^C\mathfrak{D}_{t_f}^\nu \varpi_3(t) = \{(u_2-1)\dfrac{\beta_2 I}{C+I} - \mu_2 - u_3\}\varpi_3 + \dfrac{b(1-u_2)\beta_2 I}{C+I}\varpi_4, \\[2mm]
{}_t^C\mathfrak{D}_{t_f}^\nu \varpi_4(t) = d_2 + \dfrac{\beta_1(u_1-1)S}{A+S}(\varpi_1-\varpi_2) - (u_3+\mu_3)\varpi_4, \\[2mm]
\varpi_1(t_f) = \varpi_2(t_f) = \varpi_3(t_f) = \varpi_4(t_f) = 0, \\[2mm]
S(0) = S_0, I(0) = I_0, H(0) = H_0, Y(0) = Y_0,
\end{cases}
\tag{27}
$$

where $u_1(t), u_2(t)$, and $u_3(t)$ are given by (25). In turn, the optimality conditions of Pontryagin's Minimum Principle establish that the optimal controls $u_1^*(t), u_2^*(t)$, and $u_3^*(t)$ are defined by

$$
u_1^* = \min\{\max\{0, \frac{\beta_2(\varpi_2 - \varpi_1)IH}{w_1(A+S)}\}, 1\},
$$

$$
u_2^* = \min\{\max\{0, \frac{\varpi_2 I}{w_2} + \frac{\beta_2(b\varpi_4 - \varpi_3)IH}{w_2(C+I)}\}, 1\},
$$

$$
u_3^* = \min\{\max\{0, \frac{\varpi_4 Y - \varpi_3 H}{w_3}\}, 1\}.
$$

Simulation and discussion

In this part, the effects of fractional operators on the behavior of controlled system for the dynamics of MSV disease are investigated. We develop the fractional version of fourth-order Runge-Kutta (RK4) algorithm for the coupled system (27) and apply the iterative process as follows:

We use $S(0) = 1000, I(0) = 20, H(0) = 100$, and $Y(0) = 0$ as initial values. In addition, the parameter values can be seen in [3].

The dynamical behaviors of all variables in the new fractional model without applying any control for different values of the fractional orders and the classic integer-order are plotted in Figure 1. As seen in this figure, infectious

Algorithm 1

Step 1 Set the initial values for the control functions $u_1(t), u_2(t)$, and $u_3(t)$.

Step 2 Use the current values of control functions and apply the forward fractional RK4 method for the control system and obtain the original variables.

Step 3 Apply the backward fractional RK4 method to compute the adjoint variables using the current values of the original variables and control functions.

Step 4 Update the value of control functions.

Step 5 If the updated values of the original variables, adjoint variables, and control functions are not close enough to their previous values, then go to **Step 2**.

---

maize and leafhopper densities increase with the fractional orders decrease and tend uniformly to the integer-order trajectory. Furthermore, when the fractional orders decrease, the densities of susceptible maize and leafhopper are reduced and go to the $\nu = 1$ state.

To indicate the efficiency of the new optimal control model, the same impact rate has been considered for all three controls, and the numerical results of the new model are compared with the classical integer model, in Figure 2. As can be seen in this figure, the participation of controls leads to a further reduction of infected maize in the new model than in the classical model. Therefore, the effect of controls on the fractional system is more successful than applying controls on the integer system, and the difference between them is significant. Of course, it should be noted that if no control is applied, the fractional model still leads to a significantly lower infection density than the integer model (Figure 2).

In the following, we numerically examine the effect of several optimal control scenarios, where each scenario includes more than one interventionist:

Scenario 1 Applying quarantine (u_2) and chemical control (u_3) along with elimination of prevention (u_1).

Scenario 2 Applying prevention (u_1) and quarantine (u_2) along with elimination of chemical control (u_3).

Scenario 3 Applying prevention (u_1) and chemical control (u_3) along with elimination of quarantine (u_2).

Scenario 4 Applying all three controls u_1, u_2, and u_3.

In the first scenario, the prevention ($u_1(t)$) effect is removed, and two control functions $u_2(t)$ and $u_3(t)$ are used. Figure 3 shows that in this control scenario, the number of infected maize decreases, while if no control is applied, the number of them increases over time. In addition, for fractional derivatives with lower orders, the rate of reduction of infectious cases is more
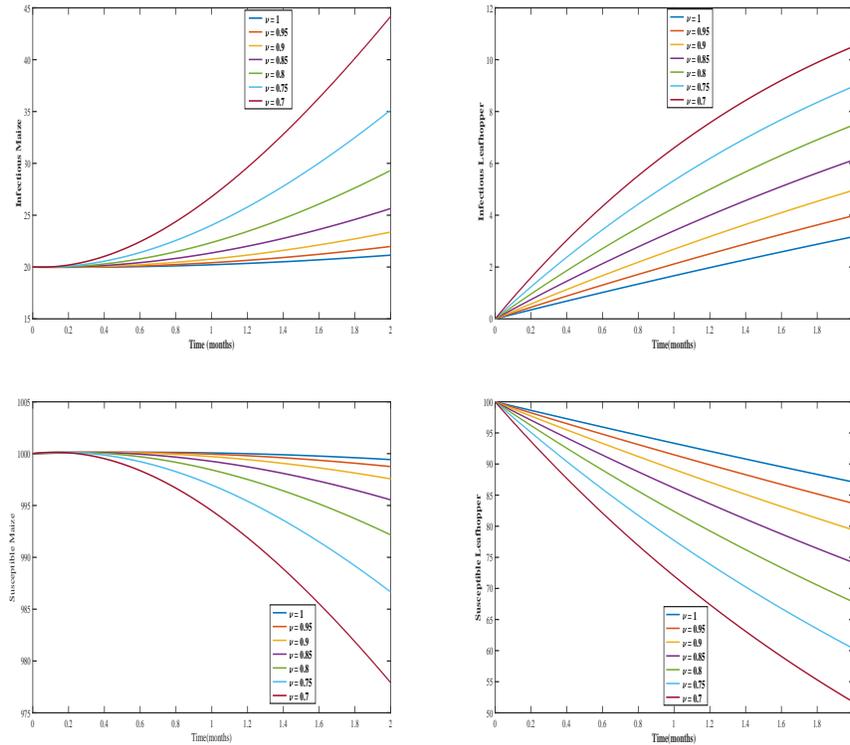
Figure 1: Numerical solutions for classical and fractional order models without controls

significant. Hence, this scenario is effective in the decline of infection in the maize community, especially in the fractional model.

The results of scenario 2 are presented in Figure 4. In this case, the quarantine control $(u_2(t))$ is maintained as in the previous scenario, but the chemical control $(u_3(t))$ is replaced by the prevention $(u_1(t))$. This scenario prevents the spread of infected maize and reduces their number. Therefore, this strategy is also successful in eliminating the disease in the maize community.

Figure 5 indicates the results of using scenario 3. In this scenario, the quarantine control is removed, unlike the previous two scenarios, and the rest of the controls are applied. Based on this figure, the number of infected is reduced compared to the case where there is no control, and this scenario also limits the growth of infectious cases.

Finally, all control interventions are considered together. As you can see in Figure 6, the use of these controls is a successful plan and causes the infected maize to be destroyed as passing the time.

Based on Figures 3–6, the quarantine control is more effective than other controls. Without quarantine control, despite preventive and chemical controls, the number of infected maize will spread, but this increasing process is much slower than the case where there is no control, and also, the difference between their increasing manner is very significant.
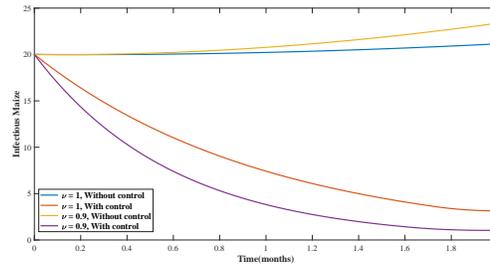
Figure 2: Numerical solutions of $I(t)$, with uncontrolled and controlled conditions for classical and fractional order models
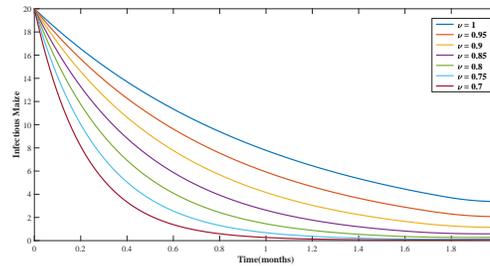
Figure 3: Numerical solutions of $I(t)$ in classic and fractional model, with quarantine and chemical controls ($u_2 \neq 0, u_3 \neq 0, u_1 = 0$)
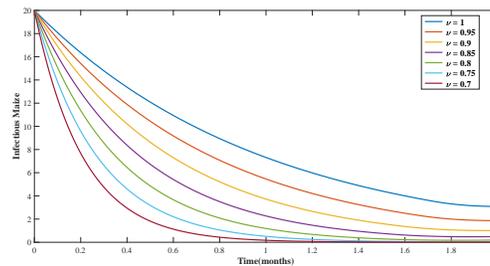
Figure 4: Numerical solutions of $I(t)$ in classic and fractional model, with prevention and quarantine controls ($u_1 \neq 0, u_2 \neq 0, u_3 = 0$)
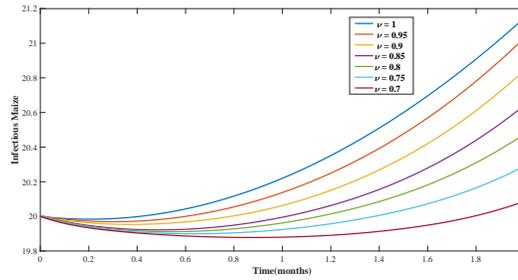
Figure 5: Numerical solutions of $I(t)$ in classic and fractional model, with prevention and chemical controls($u_1 \neq 0, u_3 \neq 0, u_2 = 0$)
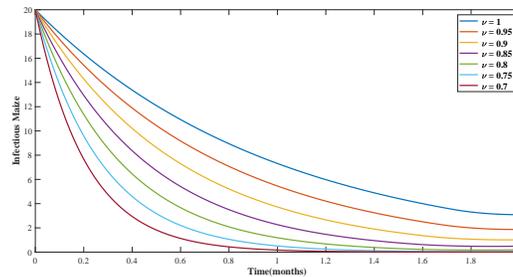
Figure 6: Numerical solutions of $I(t)$ in classic and fractional model, with prevention, quarantine and chemical controls($u_1 \neq 0, u_2 \neq 0, u_3 \neq 0$)

7 Conclusion

In the present study, we developed a new mathematical model involving the Caputo fractional derivative for MSV disease in maize plants. First, we proved that the solution of this model system exists uniquely and that all solutions remain positive and bounded whenever they start with positive initial values, thus justifying the well-posedness of a biological model. We also determined the BRN for the model. Then, we studied the local stability of the disease-free equilibrium points of the model. The study demonstrated that one of the equilibrium points is always unstable, and the other equilibrium point is locally asymptotically stable if the model's BRN is less than unity. Next, an optimization problem is formulated. Our main focus in this work is to investigate the influence of fractional-order derivatives on the optimal control problem. The optimality system was solved numerically by use of a forward and backward RK4 scheme. The effectiveness of preventive, quarantine, and chemical controls on the fractional model is investigated in the figures. Different scenarios for the participation of these controls were evaluated for various fractional-order values. We observed that in all scenarios,

the efficiency of the controls increases by moving away from the integer-order and reducing the fractional orders. Moreover, it was observed that the quarantine control is more effective than other controls. Without quarantine control, despite preventive and chemical controls, the infectious density of maize progresses with an increasing trend. Of course, it should be noted that the increasing trend is much slower than the case where there is no control, and also, the difference between their increasing manner is very significant.

Acknowledgements

References

[1] Agrawal, O.P., Defterli, O. and Baleanu, D. *Fractional optimal control problems with several state and control variables*, J. Vib. Control 16 (2010), 1967–1976.

[2] Akhavan Ghassabzadeh, F., Tohidi, E., Singh, H. and Shateyi, S. *RBF collocation approach to calculate numerically the solution of the nonlinear system of qFDEs*, J. King Saud. Univ. Sci. 33(2) (2021), 101288.

[3] Alemneh, H.T., Kassa, A.S. and Godana, A.A. *An optimal control model with cost effectiveness analysis of Maize streak virus disease in maize plant*, Infect. Dis. Model. 6 (2020), 169–182.

[4] Ameen, I.G., Baleanu, D. and Mohamed Ali, H. *Different strategies to confront maize streak disease based on fractional optimal control formulation*, Chaos, Solitons & Fractals, 164 (2022), 112699.

[5] Baleanu, D., Akhavan Ghassabzade, F., Nieto, J.J. and Jajarmi, A. *On a new and generalized fractional model for a real cholera outbreak*, Alexandria Eng. J. 61(11) (2022), 9175–9186.

[6] Bozkurt, F., Yousef, A., Abdeljawad, T., Kalinli, A. and Al Mdallal, Q. *A fractional-order model of COVID-19 considering the fear effect of the media and social networks on the community*, Chaos Solitons Fract. 152 (2021), 111403.

[7] Choi, S.K., Kang, B. and Koo, N. *Stability for Caputo fractional differential systems*, Abstr. Appl. Anal. (2014), Art. ID 631419, 6 pp.

[8] Collins, O.C. and Duffy, K.J. *Optimal control of maize foliar diseases using the plants population dynamics*, Acta Agriculturae Scandinavica, Section B– Soil and Plant Science 66(1) (2016), 20–26.

[9] Cunniffe, N.J. and Gilligan, C.A. *A theoretical framework for biological control of soil-borne plant pathogens: identifying effective strategies*, J. Theor. Biol. 278 (2011), 32–43.

[10] Diethelm, K., Siegmund, S. and Tuan, H.T. *Asymptotic behavior of solutions of linear multi-order fractional differential systems*, Fract. Calc. Appl. Anal. 20(5) (2017), 1165–1195.

[11] Evirgen, F., Uçar, S. and Özdemir, N. *System analysis of HIV infection model with CD4+T under non-singular kernel derivative*, Appl. Math. Nonlinear Sci. 5(1) (2020), 13–146.

[12] Fuller, C. *Mealie variegation In: 1st Report of the Government Entomologist, Natal, 1899-1900*, Pietermaritzburg, Natal, South Africa: P. Davis & Sons, Government Printers, 1901.

[13] Hristov, J. *Transient heat diffusion with a non-singular fading memory: from the Cattaneo constitutive equation with Jeffrey's kernel to the Caputo-Fabrizio time-fractional derivative*, Thermal Sci. 20(2) (2016), 757–762.

[14] Hugo, A., Lusekelo, E.M. and Kitengeso, R. *Optimal control and cost effectiveness analysis of tomato yellow leaf curl virus disease epidemic model*, Applied Mathematics, 9(3) (2019), 82–88.

[15] Ionescu, C., Lopes, A., Copot, D., Machado, J.A.T. and Bates, J.H.T. *The role of fractional calculus in modeling biological phenomena: A review*, Commun. Nonlinear Sci. Numer. Simul. 51 (2017), 141–159.

[16] Kilbas, A.A., Srivastava, H.H. and Trujillo, J.J. *Theory and applications of fractional differential equations*, North-Holland Mathematics Studies, 204. Elsevier Science B.V., Amsterdam, 2006.

[17] Li, H.L., Zhang, L., Hu, C., Jiang, Y.-L. and Teng, Z. *Dynamical analysis of a fractional-order predator-prey model incorporating a prey refuge*, J. Appl. Math. Comput. 54 (2017), 435–449.

[18] Matignon, D. *Stability results on fractional differential equations to control processing*, In Proceedings of the Computational Engineering in Systems and Application Multiconference; IMACS, IEEE-SMC: Lille, France, 2 (1996), 963–968.

[19] Saeedian, M., Khalighi, M., Azimi-Tafreshi, N., Jafari, G.R. and Ausloos, M. *Memory effects on epidemic evolution: the susceptible-infected-recovered epidemic model*, Phys. Rev. 95(2) (2017), 0224091–0224099.

[20] Sandhu, K.S., Singh, N. and Malhi, N.S. *Some properties of corn grains and their flours I: Physicochemical, functional and chapati-making properties of flours*, Food Chem. 101(3) (2007), 938–946.

[21] Sene, N. *Integral balance methods for Stokes' first, equation described by the left generalized fractional derivative*, Physics, 1(1) (2019), 154–166.

[22] Sene, N. *Second-grade fluid model with Caputo-Liouville generalized fractional derivative* , Chaos, Solit. Fractals, 133 (2020), 109631.

[23] Shatanawi, W., Abdo, M.S., Abdulwasaa, M.A., Shah, K., Panchal, S.K., Kawale, S.V. and Ghadle K.P. *A fractional dynamics of tuberculosis (TB) model in the frame of generalized Atangana- Baleanu derivative*, Res. Phys. 29 (2021), 104739.

[24] Shepherd, D.N., Martin, D.P., Van Der Walt, E., Dent, K., Varsani, A. and Rybicki, E.P. *Maize streak virus: an old and complex 'emerging' pathogen*, Mol. Plant Pathol. 11(1) (2010), 1–12.

[25] Shi, R., Zhao, H. and Tang, S. *Global dynamic analysis of a vector-borne plant disease model*, Adv. Difference Equ. (59) (2014), 16.

[26] Traore, A. and Sene, N. *Model of economic growth in the context of fractional derivative*, Alex. Eng. J. 59(6) (2020), 4843–4850.

How to cite this article

Cubic hat-functions approximation for linear and nonlinear fractional integral-differential equations with weakly singular kernels

H. Ebrahimi and J. Biazar*,

Abstract

In the current study, a new numerical algorithm is presented to solve a class of nonlinear fractional integral-differential equations with weakly singular kernels. Cubic hat functions (CHFs) and their properties are introduced for the first time. A new fractional-order operational matrix of integration via CHFs is presented. Utilizing the operational matrices of CHFs, the main problem is transformed into a number of trivariate polynomial equations. Error analysis and the convergence of the proposed method are evaluated, and the convergence rate is addressed. Ultimately, three examples are provided to illustrate the precision and capabilities of this algorithm. The numerical results are presented in some tables and figures.

---

*Corresponding author

Received 25 December 2022; revised 31 March 2023; accepted 12 April 2023

Hamed Ebrahimi
Department of Applied Mathematics, Faculty of Mathematical Sciences, University of Guilan, Rasht, Iran.
e-mail: hamed_ebrahimi@phd.guilan.ac.ir

Jafar Biazar
Department of Applied Mathematics, Faculty of Mathematical Sciences, University of Guilan, Rasht, Iran.
e-mail: biazar@guilan.ac.ir

1 Introduction

The mathematical modeling of many phenomena in various branches of science leads to nonlinear integral-differential equations. Fractional calculus is applied extensively by many scientists in the mathematical modeling and control of numerous dynamic systems [30, 31]. This class of equations arises in the field of signal processing [21], waves and brain modeling [18, 26], radiative equilibrium [17], and so on. Commonly, it is impractical to obtain an analytical solution to integral-differential equations. As a result, the improvement of some numerical methods and the introduction of new high-accuracy numerical algorithms is very important to obtain approximate solutions. So various numerical methods have been developed to solve these types of equations by many researchers. Some of the prominent methods are modified differential transform [15, 20], Adomian decomposition, Homotopy analysis [9, 12], Galerkin [10, 28], collocation [16, 25], product integration [1], Euler wavelets [7], haar wavelets [4], Legendre wavelets [29], Chebyshev wavelets [13], Hermite cubic splines [27], Hat functions [11], Taylor series [8], and so forth. The nonlinear fractional integral-differential equation with a weakly singular kernel appears in the following form:

$$
\begin{aligned}
{}_{0}^{C}D_t^\alpha u(t) &= g(t) + p(t)u(t) + \lambda \int_0^t (t-s)^{-\beta} u^m(s) ds, \quad \alpha > 0, \quad 0 < \beta < 1, \\
u^{(i)}(0) &= u_0^{(i)}, \quad i = 0, 1, \ldots, \lceil \alpha \rceil - 1, \quad m \in \mathbb{N}, \quad t \in I(t),
\end{aligned}
\tag{1}
$$

where $u(t)$ is an unknown function to be determined, λ is an appropriate parameter, $g(t)$ and $p(t)$ are known continuous functions on $I(T) := [0, T]$, and ${}_{0}^{C}D_t^\alpha$ is the Caputo fractional differential operator of order α. Some numerical methods convert such an integral-differential equation into a system of algebraic equations that can be easily solved.

Wang and Zhu [32] applied the second kind of Chebyshev wavelets method to give approximate solutions for the fractional integral-differential equations with a weakly singular kernel. Nemati and Lima [22] applied a numerical method based on modified hat functions (MHFs) for solving the problem (1). Xie et al. [33] used the Haar wavelets to solve a coupled system of fractional-order integral-differential equations. Riahi Beni [29] proposed a novel technique for nonlinear fractional Volterra–Fredholm integro-differential equations. Also, a numerical solution for a fractional integro-differential equation via a method based on the Gegenbauer wavelets was suggested by Özaltun, Konuralp, and Gümgüm [23]. In this paper, we introduce a high-precision numerical algorithm for the problem (1) in terms of Cubic hat-functions (CHFs).

The present work discusses some of the properties of Riemann–Liouville integral operators to solve the nonlinear fractional integral-differential equa-

tions. Applying the operational matrix method, the principal problem will be reduced to solving several nonlinear trivariate polynomial equations. In Section 2, some basic definitions and characteristics of fractional calculus are presented. Section 3 is devoted to introducing the operational matrix of CHFs basis. The fourth section studies the absolute error of approximation of a function by a truncated series of CHFs. The fifth section presents a numerical method for Problem (1). The convergence analysis of the proposed scheme is discussed in Section 6. To show the validity and accuracy of the utilized approach, three numerical examples are provided in Section 7, and the paper ends in Section 8, with a conclusion and discussion.

2 Basic concepts and definitions

In this section, some definitions and properties, which have been used in this manuscript, are explained. In this research, the Riemann–Liouville integral operator of the αth order (I_t^α) and the Caputo fractional differential operator of order α $({}_0^c D_t^\alpha)$ will be used. They are well addressed in [24].

Definition 1. Suppose that $\alpha \in R$, $n - 1 < \alpha \le n$, $n \in \mathbb{N}$, and let $u(t)$ be a continuous function defined on $[0, 1]$. The Caputo fractional derivative of order $\alpha > 0$ is defined as follows:

$$
{}_0^C D_t^\alpha u(t) = \begin{cases} \frac{1}{\Gamma(n-\alpha)} \int_0^t (t-\tau)^{(n-\alpha-1)} \frac{d^n}{d\tau} u(\tau) d\tau, & n - 1 < \alpha < n, \\ u^{(n)}(t), & \alpha = n, \end{cases} \tag{2}
$$

wherein

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

Definition 2. Assume that $\alpha > 0$ and that $u(t)$ is a continuous function defined on the closed interval $[0, 1]$. The Riemann–Liouville integral operator of order α is defined as follows:

$$
I_t^\alpha u(t) = \frac{1}{\Gamma(\alpha)} \int_0^t (t-\tau)^{\alpha-1} u(\tau) d\tau. \tag{3}
$$

The Riemann–Liouville integral operator and the Caputo fractional derivative operator satisfy the following properties [24]:

$$
I_t^\alpha (I_t^\beta u(t)) = I_t^\beta (I_t^\alpha u(t)) = I_t^{\alpha+\beta} u(t), \quad \alpha, \beta > 0,
$$

$$
I_t^\alpha ({}_0^C D_t^\alpha u(t)) = u(t) - \sum_{i=0}^{n-1} u^{(i)}(0) \frac{t^i}{i!}, \quad n - 1 < \alpha \le n, \quad t > 0. \tag{4}
$$

2.1 Definition of CHFs

First, let us state a history of Hat functions and then some definitions and properties of CHFs. In 2011, Babolian and Mordad [6] described generalized Hat functions to solve systems of Fredholm and Volterra integral equations. In 2016, Mirzaee and Hadadiyan [19] introduced MHFs to solve Volterra–Fredholm integral equations. In this paper, we improve the hat functions method and use the method for solving linear and nonlinear fractional integral-differential equations with weakly singular kernels. CHFs are defined on the closed interval $[0, T]$ and have a hat-like shape. The interval is divided into n subintervals, with equal lengths h, where $h = \frac{T}{n}$, and $n = 3K$, $K \in \mathbb{N}$.

CHFs are defined as follows:

$$\phi_0(t) = \begin{cases} -\frac{1}{6h^3}(t-h)(t-2h)(t-3h), & 0 \le t \le 3h, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

For $i = 3\nu - 2$, $\nu = 1, 2, \ldots, n/3$,

$$\phi_i(t) = \begin{cases} \frac{1}{2h^3}(t-(i-1)h)(t-(i+1)h)(t-(i+2)h), & (i-1)h \le t \le (i+2)h, \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

For $i = 3\nu - 1$, $\nu = 1, 2, \ldots, n/3$,

$$\phi_i(t) = \begin{cases} -\frac{1}{2h^3}(t-(i-2)h)(t-(i-1)h)(t-(i+1)h), & (i-2)h \le t \le (i+1)h, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

When $i = 3\nu$, $\nu = 1, 2, \ldots, (n-3)/3$,

$$\phi_i(t) = \begin{cases} \frac{1}{6h^3}(t-(i-3)h)(t-(i-2)h)(t-(i-1)h), & (i-3)h \le t \le ih, \\ -\frac{1}{6h^3}(t-(i+1)h)(t-(i+2)h)(t-(i+3)h), & ih \le t \le (i+3)h, \\ 0, & \text{otherwise,} \end{cases} \tag{8}$$

and

$$\phi_n(t) = \begin{cases} \frac{1}{6h^3}(t-(n-3)h)(t-(n-2)h)(t-(n-1)h), & (n-3)h \le t \le nh, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

A function $u(t)$ can be expressed in terms of CHFs as follows:

$$u(t) \approx u_n(t) = \sum_{i=0}^{n} a_i \phi_i(t) = A^{\mathrm{T}} \Phi(t) = \Phi(t)^T A, \tag{10}$$
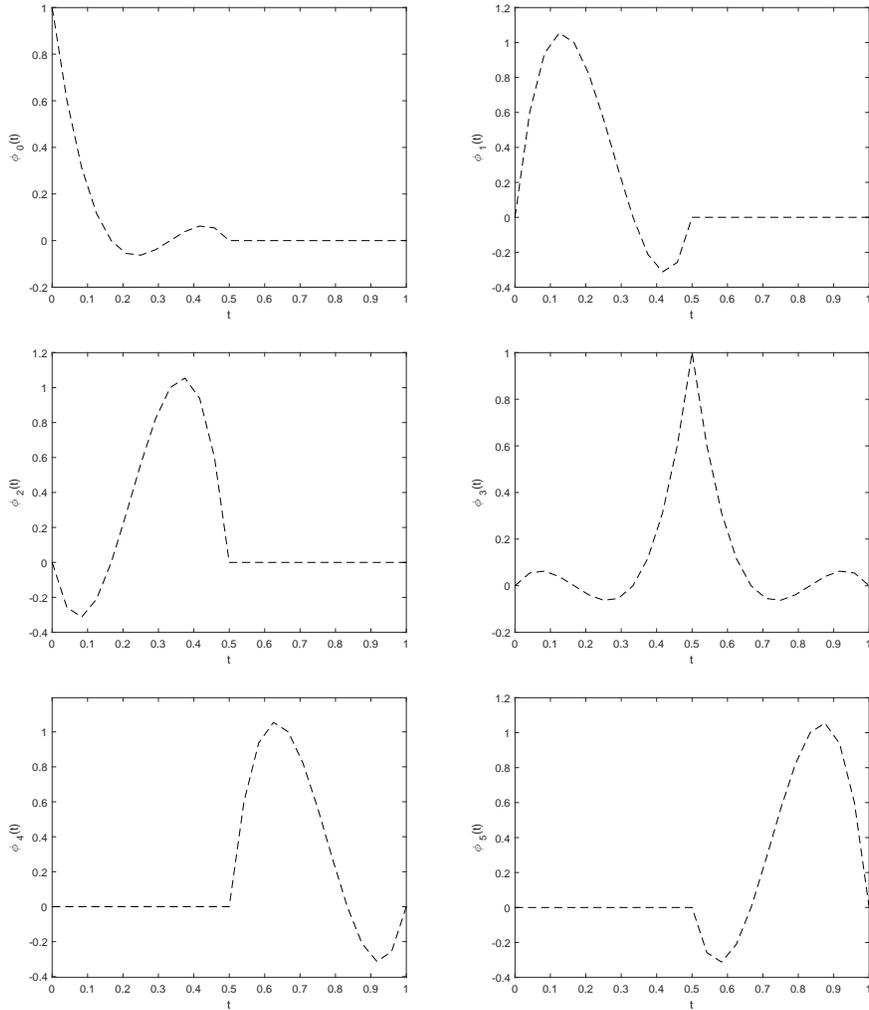
so that

$$\Phi(t) = [\phi_0(t), \phi_1(t), \ldots, \phi_n(t)]^T, \tag{11}$$

and

$$A = [a_0, a_1, \ldots, a_n]^T, \tag{12}$$

wherein $a_i = u(ih)$, $i = 0, \ldots, n$, are unknown coefficients of the CHFs. Figure 1 shows the CHFs plotted on the interval $[0, 1]$ for $n = 6$ using MATLAB package.

2.1.1 Properties of CHFs

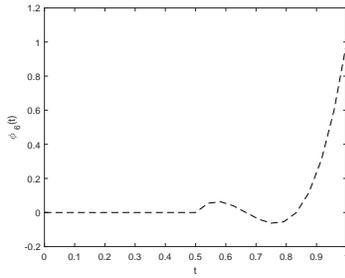Using the CHFs definition, the following properties can be obtained:

Figure 1: Plots of the CHFs, up to $n = 6, T = 1$

$$\sum_{i=0}^{n} \phi_i(t) = 1, \quad \phi_i(jh) = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases} \tag{13}$$

Multiplying both sides of this summation to $\phi_j(t)$ attains

$$\left(\sum_{i=0}^{n} \phi_j(t)\phi_i(t) \right) = \phi_j(t). \tag{14}$$

Thus, for $t = jh$, we have

$$\sum_{i=0}^{n} \phi_j(jh)\phi_i(jh) = \phi_j(jh),$$
$$[(\phi_j(jh)\phi_0(jh)) + \cdots + (\phi_j(jh)\phi_j(jh)) + \cdots + (\phi_j(jh)\phi_n(jh))] = \psi_j(jh),$$
$$[(\phi_j(jh) \times 0) + \cdots + (\phi_j(jh) \times \phi_j(jh)) + \cdots + (\phi_j(jh) \times 0)] = \phi_j(jh).$$
$$\tag{15}$$

As a result,

$$\phi_j(jh)\phi_j(jh) = \phi_j(jh). \tag{16}$$

Taking these properties, one has

$$\phi_i(jh)\phi_j(jh) \approx \begin{cases} \phi_i(jh), & j = i, \\ 0, & j \neq i. \end{cases} \tag{17}$$

Then, from the relations (17) and (11), it can be concluded that

$$\Phi(t)\Phi^T(t) \simeq \text{diag}\,[\phi_0(t), \phi_1(t), \ldots, \phi_{n-1}(t), \phi_n(t)]^T = \text{diag}\,(\Phi(t)). \tag{18}$$

2.1.2 Nonlinear approximation of CHFs

Using (18) and (10), $u^m(t)$, $m = 1, 2, \ldots$, can be calculated as follows:

$$u^2(t) \simeq A^T \Phi(t)\Phi^T(t)A = A^T \text{diag}(\Phi(t))A = A^T \text{diag}(A)\Phi(t)$$

$$= A_2^T \Phi(t), A_2 = [a_0^2, a_1^2, \ldots, a_n^2]^T,$$

$$u^3(t) \simeq u^2(t)u(t) = A_2^T \Phi(t)\Phi^T(t)A = A_2^T \operatorname{diag}(\Phi(t))A = A_2^T \operatorname{diag}(A)\Phi(t)$$

$$= A_3^T \Phi(t), A_3 = [a_0^3, a_1^3, \ldots, a_n^3]^T,$$

$$\vdots$$

$$u^m(t) \simeq \sum_{i=0}^{n} a_i{}^m \phi_i(t) = A_m^T \Phi(t), \quad A_m = [a_0^m, a_1^m, \ldots, a_n^m]^T. \tag{19}$$

3 Operational matrices of CHFs

In this part of the study, we achieve the fractional-order integral operational matrix using CHFs.

3.1 Fractional order operational matrix of integration

Let us state the following theorem.

Theorem 1. Let $\Phi(t)$ be given by (11) and let $\alpha > 0$. Then

$$I_t^\alpha \Phi(t) \simeq Q^\alpha \Phi(t), \tag{20}$$

where Q^α is called the $(n+1) \times (n+1)$ operational matrix of fractional integration of order α and is defined as follows:

$$Q^{(\alpha)} = \frac{h^\alpha}{6\Gamma(\alpha+4)} \begin{pmatrix} 0 & \rho_1 & \rho_2 & \rho_3 & \rho_4 & \rho_5 & \rho_6 & \cdots & \rho_{n-2} & \rho_{n-1} & \rho_n \\ 0 & \sigma_1 & \sigma_2 & \sigma_3 & \sigma_4 & \sigma_5 & \sigma_6 & \cdots & \sigma_{n-2} & \sigma_{n-1} & \sigma_n \\ 0 & \kappa_1 & \kappa_2 & \kappa_3 & \kappa_4 & \kappa_5 & \kappa_6 & \cdots & \kappa_{n-2} & \kappa_{n-1} & \kappa_n \\ 0 & \mu_1 & \mu_2 & \mu_3 & \mu_4 & \mu_5 & \mu_6 & \cdots & \mu_{n-2} & \mu_{n-1} & \mu_n \\ 0 & 0 & 0 & 0 & \sigma_1 & \sigma_2 & \sigma_3 & \cdots & \sigma_{n-5} & \sigma_{n-4} & \sigma_{n-3} \\ 0 & 0 & 0 & 0 & \kappa_1 & \kappa_2 & \kappa_3 & \cdots & \kappa_{n-5} & \kappa_{n-4} & \kappa_{n-3} \\ 0 & 0 & 0 & 0 & \mu_1 & \mu_2 & \mu_3 & \cdots & \mu_{n-5} & \mu_{n-4} & \mu_{n-3} \\ 0 & \vdots & \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & \sigma_1 & \sigma_2 & \sigma_3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & \kappa_1 & \kappa_2 & \kappa_3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & \mu_1 & \mu_2 & \mu_3 \end{pmatrix}, \tag{21}$$

wherein

$$\rho_k = 6k^\alpha(\alpha+3)(\alpha+2)(\alpha+1) - 11k^{\alpha+1}(\alpha+3)(\alpha+2)$$
$$+ 12k^{\alpha+2}(\alpha+3) - 6k^{\alpha+3}, \quad k = 1, 2, 3,$$

$$\rho_k = 6k^\alpha(\alpha+3)(\alpha+2)(\alpha+1) - \left(11k^{\alpha+1} - 2(k-3)^{\alpha+1}\right)(\alpha+3)(\alpha+2)$$

$$+ 6\left(2k^{\alpha+2} + (k-3)^{\alpha+2}\right)(\alpha+3)$$

$$- 6\left(k^{\alpha+3} - (k-3)^{\alpha+3}\right), \qquad k = 4, \ldots, n,$$

$$\sigma_k = 3\left(6(k)^{\alpha+1}(\alpha+3)(\alpha+2) - 10(k)^{\alpha+2}(\alpha+3) + 6(k)^{\alpha+3}\right), \quad k = 1, 2, 3,$$

$$\sigma_k = 9\left(2(k)^{\alpha+1} - (k-3)^{\alpha+1}\right)(\alpha+3)(\alpha+2)$$

$$- 6\left(5(k)^{\alpha+2} + 4(k-3)^{\alpha+2}\right)(\alpha+3)$$

$$+ 18\left((k)^{\alpha+3} - (k-3)^{\alpha+3}\right), \quad k = 4, \ldots, n,$$

$$\kappa_k = -3\left(3(k)^{\alpha+1}(\alpha+3)(\alpha+2) - 8(k)^{\alpha+2}(\alpha+3) + 6(k)^{\alpha+3}\right), \quad k = 1, 2, 3,$$

$$\kappa_k = -9\left((k)^{\alpha+1} - 2(k-3)^{\alpha+1}\right)(\alpha+3)(\alpha+2)$$

$$+ 6\left(4(k)^{\alpha+2} + 5(k-3)^{\alpha+2}\right)(\alpha+3)$$

$$- 18\left((k)^{\alpha+3} - (k-3)^{\alpha+3}\right), \qquad k = 4, \ldots, n,$$

$$\mu_k = 2(k)^{\alpha+1}(\alpha+3)(\alpha+2) - 6(k)^{\alpha+2}(\alpha+3) + 6(k)^{\alpha+3}, \qquad k = 1, 2, 3,$$

$$\mu_k = 2\left((k)^{\alpha+1} - 11(k-3)^{\alpha+1}\right)(\alpha+3)(\alpha+2) - 6(k)^{\alpha+2}(\alpha+3)$$

$$+ 6\left((k)^{\alpha+3} - 2(k-3)^{\alpha+3}\right), \qquad k = 4, 5, 6,$$

$$\mu_k = 2\left((k)^{\alpha+1} - 11(k-3)^{\alpha+1} + (k-6)^{\alpha+1}\right)(\alpha+3)(\alpha+2)$$

$$- 6\left((k)^{\alpha+2} - (k-6)^{\alpha+2}\right)(\alpha+3)$$

$$+ 6\left((k)^{\alpha+3} - 2(k-3)^{\alpha+3} + (k-6)^{\alpha+3}\right), \quad k = 7, \ldots, n. \tag{22}$$

Proof. First, for $\phi_i(t)$, $i = 0, \ldots, n$, we have the definition of the Riemann–Liouville integral operator as follows:

$$I_t^\alpha \phi_i(t) = \frac{1}{\Gamma(\alpha)} \int_0^t (t - \tau)^{\alpha-1} \phi_i(\tau) d\tau. \tag{23}$$

We expand $I_t^\alpha \phi_i(t)$, in terms of the cubic hat basis functions as follows:

$$I_t^\alpha \phi_i(t) \simeq \sum_{j=0}^n \gamma_{ij} \phi_j(t), \quad i = 0, \ldots, n, \tag{24}$$

where the values of $I_t^\alpha \phi_i(t)$ at jth node point, (jh), represent the coefficients γ_{ij}. Thus, we have

$$\gamma_{ij} = \frac{1}{\Gamma(\alpha)} \int_0^{jh} (jh - \tau)^{\alpha-1} \phi_i(\tau) d\tau, \quad i, j = 0, 1, \ldots, n. \qquad (25)$$

Using equations (5)–(9), we calculate the integral (25). For $i = 0$, by substituting (5) in (25), we introduce the coefficient as follows:

$$\gamma_{0j} = -\frac{h^\alpha}{6\Gamma(\alpha+4)} \begin{cases} \begin{pmatrix} 6j^{\alpha+3} - 12j^{\alpha+2}(\alpha+3) \\ +11j^{\alpha+1}(\alpha+3)(\alpha+2) \\ -6j^\alpha(\alpha+3)(\alpha+2)(\alpha+1), \end{pmatrix}, & j \le 2, \\[2em] \begin{pmatrix} 6\left(j^{\alpha+3} - (j-3)^{\alpha+3}\right) \\ -6\left(2j^{\alpha+2} + (j-3)^{\alpha+2}\right)(\alpha+3) \\ +\left(11j^{\alpha+1} - 2(j-3)^{\alpha+1}\right)(\alpha+3)(\alpha+2) \\ -6j^\alpha(\alpha+3)(\alpha+2)(\alpha+1), \end{pmatrix}, & j \ge 3. \end{cases}$$

$$(26)$$

For $i = 3\nu - 2$, $\nu = 1, 2, \ldots, n/3$, we obtain

$$\gamma_{ij} = \frac{h^\alpha}{2\Gamma(\alpha+4)} \begin{cases} 0, & j \le i-1, \\[1em] \begin{pmatrix} 6(j-i+1)^{\alpha+3} \\ -10(j-i+1)^{\alpha+2}(\alpha+3) \\ +6(j-i+1)^{\alpha+1}(\alpha+3)(\alpha+2) \end{pmatrix}, & i \le j \le i+2, \\[2em] \begin{pmatrix} 6\left((j-i+1)^{\alpha+3} - (j-i-2)^{\alpha+3}\right) \\ -2\left(5(j-i+1)^{\alpha+2} \\ +4(j-i-2)^{\alpha+2}\right)(\alpha+3) \\ +3\left(2(j-i+1)^{\alpha+1} \\ -(j-i-1)^{\alpha+1}\right)(\alpha+3)(\alpha+2) \end{pmatrix}, & j \ge i+3. \end{cases}$$

$$(27)$$

For $i = 3\nu - 1$, $\nu = 1, 2, \ldots, n/3$, replacing (7) into Eq. (25) yields

$$\gamma_{ij} = -\frac{h^\alpha}{2\Gamma(\alpha+4)} \begin{cases} 0, & j \leq i-2, \\ \begin{pmatrix} 6(j-i+2)^{\alpha+3} \\ -8(j-i+2)^{\alpha+2}(\alpha+3) \\ +3(j-i+2)^{\alpha+1}(\alpha+3)(\alpha+2) \end{pmatrix}, & i-1 \leq j \leq i+1, \\ \begin{pmatrix} 6\left((j-i+2)^{\alpha+3} - (j-i-1)^{\alpha+3}\right) \\ -2\left(4(j-i+2)^{\alpha+2} \\ +5(j-i-1)^{\alpha+2}\right)(\alpha+3) \\ +3\left((j-i+1)^{\alpha+1} \\ -2(j-i-1)^{\alpha+1}\right)(\alpha+3)(\alpha+2) \end{pmatrix}, & j \geq i+2. \end{cases}$$

(28)

Now, we attain (25) for $i = 3\nu$, $\nu = 1, 2, \ldots, n/3$,

$$\gamma_{ij} = \frac{h^\alpha}{6\Gamma(\alpha+4)} \begin{cases} 0, & j \leq i-3, \\ \begin{pmatrix} 6(j-i+3)^{\alpha+3} \\ -6(j-i+3)^{\alpha+2}(\alpha+3) \\ +2(j-i+3)^{\alpha+1}(\alpha+3)(\alpha+2) \end{pmatrix}, & i-2 \leq j \leq i, \\ \begin{pmatrix} 6\left((j-i+3)^{\alpha+3} - 2(j-i)^{\alpha+3}\right) \\ -6(j-i+3)^{\alpha+2}(\alpha+3) \\ +2\left((j-i+3)^{\alpha+1} \\ -11(j-i)^{\alpha+1}\right)(\alpha+3)(\alpha+2) \end{pmatrix}, & i+1 \leq j \leq i+3, \\ \begin{pmatrix} 6\left((j-i+3)^{\alpha+3} \\ -2(j-i)^{\alpha+3} + (j-i-3)^{\alpha+3}\right) \\ -6\left((j-i+3)^{\alpha+2} \\ -(j-i-3)^{\alpha+2}\right)(\alpha+3) \\ +2\begin{pmatrix} (j-i+3)^{\alpha+1} \\ -11(j-i)^{\alpha+1} \\ +(j-i-3)^{\alpha+1} \end{pmatrix}(\alpha+3)(\alpha+2) \end{pmatrix}, & j \geq i+3, \end{cases}$$

(29)

Consider $3\nu - 2 = i$ in (27), $3\nu - 1 = i$ in (28), and $3\nu = i$ in (29). Then apply $3\nu + k = j$ to all (26)–(29), $\nu = 1, \ldots, n/3$ and $k = 1, \ldots, n$. Some simple manipulations completes the proof. $\square$

As a result of using (10) and (20), we can approximate the integral of a nonlinear function as follows:

$$I_t^\alpha u^m(t) \simeq I_t^\alpha \left(\sum_{i=0}^n a_i^m \phi_i(t) \right) \simeq I_t^\alpha \left(A_m^T \Phi(t) \right) \simeq A_m^T Q^\alpha \Phi(t), \quad , m = 1, 2, \ldots.$$

$$(30)$$

For instance, when $\alpha = 1$ and $n = 3$, using the operational matrix (21), we get

$$
\begin{cases}
\text{Examples} & \text{Composite trapezoidal rule} \quad \text{CHFs} \\
& \text{solutions with } n = 6 \quad \text{solutions} \\
\text{Example 1} : \int_0^1 \sin s \, \cos^3 s \, ds & 0.2251 \quad 0.2327 \\
\text{Example 2} : \int_0^{\frac{2}{3}} 2^x x^3 ds & 0.0742 \quad 0.0730 \\
\text{Example 3} : \int_0^1 x^{\frac{1}{2}} \ln(x+1) \, ds & 0.3055 \quad 0.3053
\end{cases}
\quad (31)
$$

4 Error analysis

In this section, our analysis shows that when using CHFs to approximate a function, the order of accuracy is $O(h^4)$. Let us approximate a function $u(t)$, as (10), where

$$u_n(t) = \sum_{i=0}^n u(ih)\phi_i(t), \quad n = 3K, \quad K \in \mathbb{N}. \quad (32)$$

In the first step, for $t \in (jh, (j+1)h)$, $j = 0, 3, 6, \ldots, n-3$, using (5)–(9) and doing some computation, we obtain

$$
\begin{aligned}
u_n(t) &= \sum_{i=0}^n u(jh)\phi_i(t) \\
&= \phi_j(t)u(jh) + \phi_{j+1}(t)u((j+1)h) \\
&\quad + \phi_{j+2}(t)u((j+2)h) + \phi_{j+3}(t)u((j+3)h) \\
&= u(jh) \left(\frac{(t-(j+1)h)(t-(j+2)h)(t-(j+3)h)}{-6h^3} \right) \\
&\quad + u(jh+h) \left(\frac{(t-jh)(t-(j+2)h)(t-(j+3)h)}{2h^3} \right) \\
&\quad + u(jh+2h) \left(\frac{(t-jh)(t-(j+1)h)(t-(j+3)h)}{-2h^3} \right) \\
&\quad + u(jh+3h) \left(\frac{(t-jh)(t-(j+1)h)(t-(j+2)h)}{6h^3} \right).
\end{aligned}
$$

Then by simplifying the current relationship, we have

$$u_n(t) = u(jh) \left(\frac{(t - jh)^3 - 6h(t - jh)^2 + 11h^2 (t - jh) - 6h^3}{-6h^3} \right)$$

$$+ u(jh + h) \left(\frac{(t - jh)^3 - 5h(t - jh)^2 + 6h^2 (t - jh)}{2h^3} \right)$$

$$+ u(jh + 2h) \left(\frac{(t - jh)^3 - 4h(t - jh)^2 + 3h^2 (t - jh)}{-2h^3} \right)$$

$$+ u(jh + 3h) \left(\frac{(t - jh)^3 - 3h(t - jh)^2 + 2h^2 (t - jh)}{6h^3} \right).$$

Therefore

$$u_n(t) = u(jh)$$

$$+ (t - jh) \left(\frac{-11u(jh) + 18u(jh + h) - 9u(jh + 2h) + 2u(jh + 3h)}{6h} \right)$$

$$+ \frac{(t - jh)^2}{2} \left(\frac{2u(jh) - 5u(jh + h) + 4u(jh + 2h) - u(jh + 3h)}{h^2} \right)$$

$$+ \frac{(t - jh)^3}{6} \left(\frac{-u(jh) + 3u(jh + h) - 3u(jh + 2h) + u(jh + 3h)}{h^3} \right).$$

It is known that the kth, $k = 1, 2, 3$, order derivative of $u(t)$ about the point (jh) is as follows:

$$u'(jh) = \frac{-11u(jh) + 18u(jh + h) - 9u(jh + 2h) + 2u(jh + 3h)}{6h} + O(h^4),$$

$$u''(jh) = \frac{2u(jh) - 5u(jh + h) + 4u(jh + 2h) - u(jh + 3h)}{h^2} + O(h^4),$$

$$u'''(jh) = \frac{-u(jh) + 3u(jh + h) - 3u(jh + 2h) + u(jh + 3h)}{h^3} + O(h^4).$$

So, assuming $h \to 0$ results in

$$u_n(t) = u(jh) + (t - jh)u'(jh) + \frac{(t - jh)^2}{2}u''(jh) + \frac{(t - jh)^3}{6}u'''(jh) + O(h^4). \tag{33}$$

Expanding $u(t)$ in the Taylor's series, about the point $t = jh$, we have

$$u(t) = \sum_{k=0}^{3} \frac{(t - jh)^k}{k!} u^{(k)}(jh) + O(t - jh)^4. \tag{34}$$

According to (33) and (34), we can obtain the error between the exact and approximate values of $u(t)$ as follows:

$$u(t) - u_n(t) = O(t - jh)^4. \tag{35}$$

Thus, for $t \to jh$, $j = 0, 3, 6, \ldots, n - 3$ and $h \to 0$, since $(t - jh) < h$, from (35), we attain

$$|u(t) - u_n(t)| = O(h^4). \tag{36}$$

In the second step, for $j = 1, 4, 7, \ldots, n - 2$ and $jh < t < (j + 1)h$, we get

$$
\begin{aligned}
u_n(t) &= \sum_{i=0}^{n} u(jh)\phi_i(t) \\
&= \phi_{j-1}(t)u\left((j-1)h\right) + \phi_j(t)u\left(jh\right) \\
&\quad + \phi_{j+1}(t)u\left((j+1)h\right) + \phi_{j+2}(t)u\left((j+2)h\right) \\
&= u(jh - h)\left(\frac{(t - jh)\left(t - (j+1)h\right)\left(t - (j+2)h\right)}{-6h^3}\right) \\
&\quad + u(jh)\left(\frac{\left(t - (j-1)h\right)\left(t - (j+1)h\right)\left(t - (j+2)h\right)}{2h^3}\right) \\
&\quad + u(jh + h)\left(\frac{\left(t - (j-1)h\right)(t - jh)\left(t - (j+2)h\right)}{-2h^3}\right) \\
&\quad + u(jh + 2h)\left(\frac{\left(t - (j-1)h\right)(t - jh)\left(t - (j+1)h\right)}{6h^3}\right).
\end{aligned}
$$

Hence

$$
\begin{aligned}
u_n(t) &= u(jh) \\
&\quad + (t - jh)\left(\frac{-2u(jh - h) - 3u(jh) + 6u(jh + h) - u(jh + 2h)}{6h}\right) \\
&\quad + \frac{(t - jh)^2}{2}\left(\frac{u(jh - h) - 2u(jh) + u(jh + h)}{h^2}\right) \\
&\quad + \frac{(t - jh)^3}{6}\left(\frac{-u(jh - h) + 3u(jh) - 3u(jh + h) + u(jh + 2h)}{h^3}\right).
\end{aligned}
\tag{37}
$$

As a reminder, the derivatives of $u(t)$ about the point jh are as follows:

$$
\begin{aligned}
u'(jh) &= \frac{-2u(jh - h) - 3u(jh) + 6u(jh + h) - u(jh + 2h)}{6h} + O(h^4), \\
u''(jh) &= \frac{u(jh - h) - 2u(jh) + u(jh + h)}{h^2} + O(h^4), \\
u'''(jh) &= \frac{-u(jh - h) + 3u(jh) - 3u(jh + h) + u(jh + 2h)}{h^3} + O(h^4). \tag{38}
\end{aligned}
$$

As $h \to 0$, from (37)–(38), we get

$$u_n(t) = u(jh) + (t-jh)u'(jh) + \frac{(t-jh)^2}{2}u''(jh) + \frac{(t-jh)^3}{6}u'''(jh) + O(h^4).$$
$$(39)$$

Considering (39), (34), and $h \to 0$ for $j = 1, 4, 7, \ldots, n-2$, one has

$$|u(t) - u_n(t)| = O(h^4).$$
$$(40)$$

In the final step, for $j = 2, 5, 8, \ldots, n-1$ and $t \in (jh, (j+1)h)$, we attain

$$
\begin{aligned}
u_n(t) &= \sum_{i=0}^{n} u(jh)\phi_i(t) \\
&= \phi_{j-2}(t)u\left((j-2)h\right) + \phi_{j-1}(t)u\left((j-1)h\right) \\
&\quad + \phi_j(t)u\left(jh\right) + \phi_{j+1}(t)u\left((j+1)h\right) \\
&= u(jh - 2h)\left(\frac{(t-(j-1)h)\,(t-jh)\,(t-(j+1)h)}{-6h^3}\right) \\
&\quad + u(jh - h)\left(\frac{(t-(j-2)h)\,(t-jh)\,(t-(j+1)h)}{2h^3}\right) \\
&\quad + u(jh)\left(\frac{(t-(j-2)h)\,(t-(j-1)h)\,(t-(j+1)h)}{-2h^3}\right) \\
&\quad + u(jh + h)\left(\frac{(t-(j-2)h)\,(t-(j-1)h)\,(t-jh)}{6h^3}\right).
\end{aligned}
$$

As a result,

$$
\begin{aligned}
u_n(t) &= u(jh) + (t-jh)\left(\frac{u(jh-2h) - 6u(jh-h) + 3u(jh) + 2u(jh+h)}{6h}\right) \\
&\quad + \frac{(t-jh)^2}{2}\left(\frac{u(jh-h) - 2u(jh) + u(jh+h)}{h^2}\right) \\
&\quad + \frac{(t-jh)^3}{6}\left(\frac{-u(jh-2h) + 3u(jh-h) - 3u(jh) + u(jh+h)}{h^3}\right).
\end{aligned}
$$
$$(41)$$

On the other hand, according to the derivatives of $u(t)$ about the point jh, (41) can be written as follows:

$$u_n(t) = u(jh) + (t-jh)u'(jh) + \frac{(t-jh)^2}{2}u''(jh) + \frac{(t-jh)^3}{6}u'''(jh) + O(h^4).$$
$$(42)$$

Thus, assuming (42), (34), and $h \to 0$, for $j = 2, 5, 8, \ldots, n-1$, one has

$$|u(t) - u_n(t)| = O(h^4).$$
$$(43)$$

Finally, for $t \in (jh, (j+1)h)$, $j = 0, 1, 2, \ldots, n$, and $h \to 0$, using (36), (40), and (43), we get

$$|u(t) - u_n(t)| = O(h^4). \tag{44}$$

5 Numerical algorithm

In this section, a numerical algorithm is offered to solve problem (1). Consider the following nonlinear fractional integral-differential equation with weakly singular kernel :

$$\,_0^C D_t^\alpha u(t) = g(t) + p(t)u(t) + \lambda \int_0^t (t-s)^{-\beta} u^m(s)ds, \tag{45}$$

$$\alpha > 0, \quad 0 < \beta < 1, \quad m \in \mathbb{N}, \quad t \in I(t).$$

First, putting $-\beta = \omega - 1$, $0 < \omega < 1$ in the third term on the right of this equation, we get

$$\int_0^t (t-s)^{-\beta} u^m(s)ds = \Gamma(\omega) \left(\frac{1}{\Gamma(\omega)} \int_0^t (t-s)^{\omega-1} u^m(s)ds \right), \quad 0 < \omega < 1. \tag{46}$$

By the definition of Riemann–Liouville fractional integral operator, [24], the current relationship can be rewritten as follows:

$$\int_0^t (t-s)^{-\beta} u^m(s)ds = \Gamma(\omega) I_t^\omega \left(u^m(t) \right). \tag{47}$$

Now, by applying (47), the Riemann–Liouville integral operator of order α on the both sides of (45), one gets

$$I_t^\alpha \left(\,_0^C D_t^\alpha u(t) \right) = I_t^\alpha \left(g(t) \right) + I_t^\alpha \left(p(t)u(t) \right) + \lambda \Gamma(\omega) I_t^\alpha I_t^\omega \left(u^m(t) \right),$$
$$u(t) = z(t) + I_t^\alpha \left(g(t) \right) + I_t^\alpha \left(p(t)u(t) \right) + \lambda \Gamma(\omega) I_t^{\alpha+\omega} \left(u^m(t) \right), \tag{48}$$

where

$$z(t) = \sum_{i=0}^{\lceil \alpha \rceil - 1} u^{(i)}(0) \frac{t^i}{i!}, \quad \alpha > 0.$$

Now, by approximating the functions in (48) by CHFs (10) and (19), we attain

$$u(t) \simeq \sum_{i=0}^n a_i \phi_i(t) = A^T \Phi(t), \quad A_m = [a_0, a_1, \ldots, a_n]^T. \tag{49}$$

$$u^m(t) \simeq \sum_{i=0}^n a_i{}^m \phi_i(t) = A_m^T \Phi(t), \quad A_m = [a_0^m, a_1^m, \ldots, a_n^m]^T. \tag{50}$$

$$z(t) \simeq \sum_{i=0}^{n} z(ih)\phi_i(t) = z^T \Psi(t), \quad Z = [z(0), z(h), \ldots, z(nh)]^T, \qquad (51)$$

$$g(t) \simeq \sum_{i=0}^{n} g(ih)\phi_i(t) = G^T \Psi(t), \quad G = [g(0), g(h), \ldots, g(nh)]^T, \qquad (52)$$

and

$$p(t) \simeq \sum_{i=0}^{n} p(ih)\phi_i(t) = p^T \Psi(t), \quad P = [p(0), p(h), \ldots, p(nh)]^T, \qquad (53)$$

wherein n is an integer multiple of 3. Utilizing (20)–(22), and (18) and the substitution of (49)–(53) in (48) become

$$A^T \Phi(t) = Z^T \Phi(t) + I_t^\alpha (G^T \Phi(t)) + I_t^\alpha (P^T \Phi(t)\Phi(t)^T A) + \lambda\Gamma(\omega)I_t^{\alpha+\omega}(A_m^T \Phi(t)),$$

$$A^T \Phi(t) - Z^T \Phi(t) - I_t^\alpha (G^T \Phi(t)) - I_t^\alpha (P^T diag\,(\Phi(t))\,A) - \lambda\Gamma(\omega)I_t^{\alpha+\omega}(A_m^T \Phi(t)) = 0,$$

$$A^T \Phi(t) - Z^T \Phi(t) - I_t^\alpha (G^T \Phi(t)) - I_t^\alpha (P^T diag\,(A)\,\Phi(t)) - \lambda\Gamma(\omega)I_t^{\alpha+\omega}(A_m^T \Phi(t)) = 0,$$

$$A^T \Phi(t) - Z^T \Phi(t) - G^T Q^{(\alpha)}\Phi(t)\ - P^T diag\,(A)\,Q^{(\alpha)}\Phi(t) - \lambda\Gamma(\omega)A_m^T Q^{(\alpha+\omega)}\Phi(t) = 0.$$

Thus

$$A^T - Z^T - G^T Q^{(\alpha)}\ - P^T \operatorname{diag}(A)\,Q^{(\alpha)} - \lambda\Gamma(\omega)A_m^T Q^{(\alpha+\omega)} = 0,$$

$$\alpha > 0, \quad 0 < \omega < 1, \quad \omega = 1 - \beta. \qquad (54)$$

This system has the dimension $(n+1) \times (n+1)$.
Suppose that

$$Q^{(\alpha)} = [\gamma_{ij}], \quad Q^\omega = [\theta_{ij}], \quad i, j = 0, 1, 2, \ldots, n. \qquad (55)$$

Then, from the operational matrix (23), one gets

$$\gamma_{i0} = \theta_{i0} = 0, \quad i = 0, 1, 2, \ldots, n,$$

$$\gamma_{ij} = \theta_{ij} = 0, \quad j = 1, 3, \ldots, n-1, \quad i = j + 3, j + 4, \ldots, n,$$

$$\gamma_{ij} = \theta_{ij} = 0, \quad j = 2, 4, \ldots, n, \quad i = j + 2, j + 3, \ldots, n.$$

Using (54), the unknown coefficients can be determined. We start to find the first unknown coefficient as follows:

$$a_0 = z(0). \qquad (56)$$

In the next step, we get

$$\textbf{system 1:} \begin{cases} eq_1 : [a_1] - [z(1h)] - \left[\sum\limits_{i=0}^{3} g(ih)\gamma_{i1}\right] \\ \quad - \left[\sum\limits_{i=0}^{3} p(ih)\gamma_{i1}a_i\right] - \left[\lambda\Gamma(\omega)\sum\limits_{i=0}^{3} \theta_{i1}a_i{}^m\right] = 0, \\ eq_2 : [a_2] - [z(2h)] - \left[\sum\limits_{i=0}^{3} g(ih)\gamma_{i2}\right] \\ \quad - \left[\sum\limits_{i=0}^{3} p(ih)\gamma_{i2}a_i\right] - \left[\lambda\Gamma(\omega)\sum\limits_{i=0}^{3} \theta_{i2}a_i{}^m\right] = 0, \\ eq_3 : [a_3] - [z(3h)] - \left[\sum\limits_{i=0}^{3} g(ih)\gamma_{i3}\right] \\ \quad - \left[\sum\limits_{i=0}^{3} p(ih)\gamma_{i3}a_i\right] - \left[\lambda\Gamma(\omega)\sum\limits_{i=0}^{3} \theta_{i3}a_i{}^m\right] = 0. \end{cases}$$

Solving the first system allows us to calculate a_1, a_2, and a_3, then we solve the following system:

$$\textbf{system 2:} \begin{cases} eq_4 : [a_4] - [z(4h)] - \left[\sum\limits_{i=0}^{6} g(ih)\gamma_{i4}\right] \\ \quad - \left[\sum\limits_{i=0}^{6} p(ih)\gamma_{i4}a_i\right] - \left[\lambda\Gamma(\omega)\sum\limits_{i=0}^{6} \theta_{i4}a_i{}^m\right] = 0, \\ eq_5 : [a_5] - [z(5h)] - \left[\sum\limits_{i=0}^{6} g(ih)\gamma_{i5}\right] \\ \quad - \left[\sum\limits_{i=0}^{6} p(ih)\gamma_{i5}a_i\right] - \left[\lambda\Gamma(\omega)\sum\limits_{i=0}^{6} \theta_{i5}a_i{}^m\right] = 0, \\ eq_6 : [a_6] - [z(6h)] - \left[\sum\limits_{i=0}^{6} g(ih)\gamma_{i6}\right] \\ \quad - \left[\sum\limits_{i=0}^{6} p(ih)\gamma_{i6}a_i\right] - \left[\lambda\Gamma(\omega)\sum\limits_{i=0}^{6} \theta_{i6}a_i{}^m\right] = 0. \end{cases}$$

By solving **system 2**, the values of the unknown parameters a_4, a_5, and a_6 are calculated. Then we can get the values of a_7, a_8, and a_9 using **system 3**:

$$\textbf{system 3:} \begin{cases} eq_7 : [a_7] - [z(7h)] - \left[\sum\limits_{i=0}^{9} g(ih)\gamma_{i7}\right] \\ \quad \left[\sum\limits_{i=0}^{9} p(ih)\gamma_{i7}a_i\right] - \left[\lambda\Gamma(\omega)\sum\limits_{i=0}^{9} \theta_{i7}a_i{}^m\right] = 0, \\ eq_8 : [a_8] - [z(8h)] - \left[\sum\limits_{i=0}^{9} g(ih)\gamma_{i8}\right] \\ \quad - \left[\sum\limits_{i=0}^{9} p(ih)\gamma_{i8}a_i\right] - \left[\lambda\Gamma(\omega)\sum\limits_{i=0}^{9} \theta_{i8}a_i{}^m\right] = 0, \\ eq_9 : [a_9] - [z(9h)] - \left[\sum\limits_{i=0}^{9} g(ih)\gamma_{i9}\right] \\ \quad - \left[\sum\limits_{i=0}^{9} p(ih)\gamma_{i9}a_i\right] - \left[\lambda\Gamma(\omega)\sum\limits_{i=0}^{9} \theta_{i9}a_i{}^m\right] = 0. \end{cases}$$

The process can be continued up to the following form:

$$
\textbf{system n/3:} \begin{cases}
eq_{n-2} : [a_{n-2}] - [z((n-2)h)] - \left[\sum_{i=0}^{n} g(ih)\gamma_{i(n-2)}\right] \\
\quad - \left[\sum_{i=0}^{n} p(ih)\gamma_{i(n-2)}a_i\right] - \left[\lambda\Gamma(\omega)\sum_{i=0}^{n} \theta_{i(n-2)}a_i{}^m\right] = 0, \\
eq_{n-1} : [a_{n-1}] - [z((n-1)h)] - \left[\sum_{i=0}^{n} g(ih)\gamma_{i(n-1)}\right] \\
\quad - \left[\sum_{i=0}^{n} p(ih)\gamma_{i(n-1)}a_i\right] - \left[\lambda\Gamma(\omega)\sum_{i=0}^{n} \theta_{i(n-1)}a_i{}^m\right] = 0, \\
eq_{n} : [a_{n}] - [z(nh)] - \left[\sum_{i=0}^{n} g(ih)\gamma_{in}\right] \\
\quad - \left[\sum_{i=0}^{n} p(ih)\gamma_{in}a_i\right] - \left[\lambda\Gamma(\omega)\sum_{i=0}^{n} \theta_{in}a_i{}^m\right] = 0.
\end{cases}
$$

As a result, the values of a_{n-2}, a_{n-1}, and a_n are derived using system $n/3$. Therefore, we can obtain an approximate solution via (10). To solve the nonlinear equations, see [34]. The computations were handled by MATLAB package. The following theorem outlines the proposed method.

Theorem 2. Consider the principal problem (1). To obtain a numerical solution to (1) using CHFs, the following iterative algorithm is offered:

Proof. See the scheme proposed in this section, (56)–(57). $\qquad\square$

6 Convergence analysis

In this section, we will verify the convergence of the numerical proposed scheme.

Theorem 3. Let $u_n(t)$ be the numerical solution of (1) obtained by the proposed method in Section (5). Moreover, $u(t)$ is an exact solution and $E_n(t)$ is the residual error for numerical solution. Also, suppose that M and K are positive constants. Then, $E_n(t)$ tends to zero, as $n \to \infty$, where

$$
M = \sup_{t,\tau \in [0,T]} \left| \Gamma^{-1}(\alpha)(t-\tau)^{\alpha-1} p(\tau) \right|,
$$
$$
K = \sup_{t,\tau \in [0,T]} \left| \lambda m L \Gamma^{-1}(\alpha+\omega)(t-\tau)^{\alpha+\omega-1} \right|.
$$

Proof. Applying the Riemann–Liouville integral operator of order α and (48), it is appropriate to rewrite (1) in the integral form

$$
u(t) = z(t) + I_t^{\alpha}\left(g(t)\right) + I_t^{\alpha}\left(p(t)u(t)\right) + \lambda\Gamma(\omega)I_t^{\alpha+\omega}\left(u^m(t)\right), \qquad (57)
$$

where

Algorithm 1 An algorithm for approximation using CHsF

Step 1: Inputs, n (integer multiple of 3), $\alpha, \beta, \lambda, T, g(t), p(t)$,
$\quad\quad u^{(i)}(0), \quad i = 0, 1, \ldots, \lceil \alpha \rceil - 1$.

Step 2: Set $\omega = 1 - \beta$, $h = T/n$, and $t_i = ih$, $i = 0, \ldots, n$.

Step 3: $z(t) = \sum\limits_{i=0}^{n-1} u^{(i)}(0) \frac{t^i}{i!}$.

Step 4: Compute the elements of $Q^{(\alpha)} = [\gamma_{ij}]$ and $Q^{(\alpha+\omega)} = [\theta_{ij}]$,
$\quad\quad i, j = 0, \ldots, n$.

Step 5: Set and solve recursive trivariable system v, $v = 1 : n/3$.

$\quad\quad\quad \Big|\, a_0 = z(0)$,
$\quad\quad\quad \Big|\, \text{for} \quad v = 1 : n/3$
$\quad\quad\quad \Big|\, \text{Solution of the } v\text{th system determines}$
$\quad\quad\quad \Big|\, \text{the unknown parameter.}$

$$\text{system } v : \left\{
\begin{array}{l}
\left[a_{(3v-2)} \right] - \left[z((3v-2)h) \right] - \left[\sum\limits_{k=0}^{3v} g(ih)\gamma_{i(3v-2)} \right] \\[3mm]
- \left[\sum\limits_{k=0}^{3v} p(ih)\gamma_{i(3v-2)} a_i \right] \\[3mm]
- \left[\Gamma(\omega) \sum\limits_{i=0}^{3v} \theta_{i(3v-2)} a_i{}^m \right] = 0, \\[4mm]
\left[a_{(3v-1)} \right] - \left[z((3v-1)h) \right] - \left[\sum\limits_{k=0}^{3v} g(ih)\gamma_{i(3v-1)} \right] \\[3mm]
- \left[\sum\limits_{i=0}^{3v} p(ih)\gamma_{i(3v-1)} a_i \right] \\[3mm]
- \left[\Gamma(\omega) \sum\limits_{i=0}^{3v} \theta_{i(3v-1)} a_i{}^m \right] = 0, \\[4mm]
\left[a_{(3v)} \right] - \left[z((3v)h) \right] - \left[\sum\limits_{i=0}^{3v} g(ih)\gamma_{i(3v)} \right] \\[3mm]
- \left[\sum\limits_{i=0}^{3v} p(ih)\gamma_{i(3v)} a_i \right] \\[3mm]
- \left[\Gamma(\omega) \sum\limits_{i=0}^{3v} \theta_{i(3v)} a_i{}^m \right] = 0,
\end{array}
\right.$$

$\quad\quad\quad\quad \text{end.}$

Step 6: Calculate fully $a_i, \quad i = 0, 1, \ldots, n$.

Step 7: Define CHFs: $(\phi_i(t), \quad i = 0, 1, \ldots, n)$.

Step 8: Determine the approximate solutions: $u_n(t) = \sum\limits_{i=0}^{n} a_i \phi_i(t)$.

$$z(t) = \sum_{i=0}^{\lceil \alpha \rceil - 1} u^{(i)}(0) \frac{t^i}{i!}, \quad \alpha > 0, \quad \omega = 1 - \beta, \quad 0 < \beta < 1, \quad t \in I(t).$$

Thus, $u_n(t)$ satisfies the following equation:

$$u_n(t) = z(t) + I_t^\alpha \left(g(t) \right) + I_t^\alpha \left(p(t)u_n(t) \right) + \lambda \Gamma(\omega) I_t^{\alpha+\omega} \left(u_n^m(t) \right). \qquad (58)$$

If the residual function $E_n(t)$ is not zero, then we can obtain it by using the following relation:

$$E_n(t) = e_n[u](t) - J_n^\alpha[u](t) - V_n^{\alpha+\omega}[u^m](t), \qquad (59)$$

where

$$e_n[u](t) = u(t) - u_n(t), \qquad (60)$$

$$J_n^\alpha[u](t) = \frac{1}{\Gamma(\alpha)} \int_0^t (t-\tau)^{\alpha-1} p(\tau)(u(\tau) - u_n(\tau))d\tau, \qquad (61)$$

and

$$V_n^{\alpha+\omega}[u^m](t) = \frac{\lambda}{\Gamma(\alpha+\omega)} \int_0^t (t-\tau)^{\alpha+\omega-1} (u^m(\tau) - u_n^m(\tau))d\tau. \qquad (62)$$

Then, we get

$$|E_n(t)| \leq |e_n[u](t)| + |J_n^\alpha[u](t)| + \left| V_n^{\alpha+\omega}[u^m](t) \right|. \qquad (63)$$

For $t \in (ih, (i+3)h), i = 0, 3, 6, \ldots, n-3$, according to (44), the approximation of the absolute error using CHFs yields

$$|u(t) - u_n(t)| = O(h^4). \qquad (64)$$

By using (60), we have

$$|e_n[u](t)| = O(h^4), \qquad (65)$$

when $h \to 0$, $|e_n[u](t)| \to 0$. Then, by using (61) and (64), we attain

$$\begin{aligned}
|J_n^\alpha[u](t)| &= \frac{1}{\Gamma(\alpha)} \left| \int_0^t (t-\tau)^{\alpha-1} p(\tau)(u(\tau) - u_n(\tau))d\tau \right| \\
&\leq \frac{|p(\tau)|}{\Gamma(\alpha)} \int_0^t (t-\tau)^{\alpha-1} |u(\tau) - u_n(\tau)| \, d\tau \\
&\leq \frac{|p(\tau)|}{\Gamma(\alpha)} \int_0^t (t-\tau)^{\alpha-1} |u(\tau) - u_n(\tau)| \, d\tau \\
&\leq MO(h^4), \qquad (66)
\end{aligned}$$

wherein

$$M = \sup_{t,\tau \in [0,T]} \left| \Gamma^{-1}(\alpha)(t-\tau)^{\alpha-1} p(\tau) \right| \text{ and when} h \to 0, \quad \left| J_n^{\beta-\alpha}[u](t) \right| \to 0.$$

In addition, the following inequality holds [11]:

$$|u^m(t) - u_n^m(t)| \leq mL \, |u(t) - u_n(t)| \,, \tag{67}$$

where $L = \left| (\max(u(t), u_n(t)))^{m-1} \right|$. As well, from (62), (64), and (67), we have

$$
\begin{aligned}
\left| V_n^{\alpha+\omega}[u^m](t) \right| &= \frac{1}{\Gamma(\alpha+\omega)} \left| \lambda \int_0^t (t-\tau)^{\alpha+\omega-1} (u^m(\tau) - u_n^m(\tau)) d\tau \right| \\
&\leq \frac{|\lambda|}{\Gamma(\alpha+\omega)} \int_0^t (t-\tau)^{\alpha+\omega-1} \left| (u^m(\tau) - u_n^m(\tau)) \right| d\tau \\
&\leq \frac{mL \, |\lambda|}{\Gamma(\alpha+\omega)} \int_0^t (t-\tau)^{\alpha+\omega-1} \left| (u(\tau) - u_n(\tau)) \right| d\tau \\
&\leq K O(h^4), \tag{68}
\end{aligned}
$$

wherein

$$K = \sup_{t,\tau \in [0,T]} \left| \lambda m L \Gamma^{-1}(\alpha+\omega)(t-\tau)^{\alpha+\omega-1} \right| \text{ and as } h \to 0, \quad \left| V_n^{\alpha+\omega}[u^m](t) \right| \to 0.$$

Then, from relations (65), (66), (68), and (63), it is obvious that $|E_n(t)|$ tends to zero, as $h \to 0$, or $n \to \infty$. □

7 Numerical examples

In this section, the theoretical results of the previous sections are used for solving linear and nonlinear fractional integral-differential equations with the weakly singular kernel, that is, the initial condition equation (1). For assessing the accuracy of the scheme, let us define the maximum absolute error (L_∞-norm error) as

$$\|\xi_n\|_\infty = \sup_{[t_i = ih]_{i=0}^n} \{|u(t_i) - u_n(t_i)|\}. \tag{69}$$

Using this definition, the order of convergence, with respect to this norm, is introduced as follows:

$$\text{Rate} = \log_2 \left(\frac{\|\xi_n\|_\infty}{\|\xi_{2n}\|_\infty} \right), \tag{70}$$

For some problems, there are no exact solutions, so the L_2-norm error is calculated by the following formula:

$$\widetilde{\|\xi_n\|_2} = \left(\sum_{i=0}^n (u_n(t_i) - u_{2n}(t_i))^2 \right)^{\frac{1}{2}}, \quad t_i = ih, \quad i = 0, \ldots, n, \tag{71}$$

where $u_n(t)$, $n = T/h$ is the approximate solution defined as (49). In addition, the results of different values of α are compared with each other and with MHFs method, [22].

Example 1. Consider a nonlinear integral-differential equation with weakly singular kernel [22]:

$$_o^C D_t^\alpha u(t) = g(t) + p(t)u(t) + \int_0^t (t-s)^{-\beta} u^2(s)ds, \quad t \in [0,1],$$

$$g(t) = 3t^2 - \left(\frac{\sqrt{\pi}\Gamma(7)}{\Gamma\left(\frac{15}{6}\right)} \right) t^{\frac{13}{2}}, \quad p(t) = 0, \quad u(0) = 0.$$

For $\alpha = 1$ and $\beta = \frac{1}{2}$, the exact solution is $u(t) = t^3$. Approximate numerical results using different values of n are shown in Tables 1–3 and Figures 2–4. Table 1 shows the approximate and exact solutions to the problem at some points. Also, the L_∞-norm errors and convergence orders obtained by the current method are compared with the MHFs method [22] and the methods presented in [14, 5] in Table 2. Table 3 shows the comparison of the result of the l_2-norm error $\widetilde{\|\xi_n\|_2}$ obtained by the proposed method. It is clear from Table 2 that the results of the present method for less or similar n are better than the results obtained in [22]. Figure 2 indicates the behavior of absolute errors for Example 1. Also, Figure 3 shows the logarithm of the L_∞-norm errors. As can be seen from the plot, as n increases, the error decreases. Also, the comparison of the results obtained for different values of alpha with the exact solutions of the equation is plotted in Figure 4.

Table 1: Numerical results of Example 1

Points s	Exact solutions $u(s)$	Approximate solutions $n = 12$	Approximate solutions $n = 24$	Approximate solutions $n = 48$
0	0.00000000000	0.00000000000	0.00000000000	0.00000000000
$1/12$	0.00057870370	0.00057871890	0.00057870463	0.00057870368
$1/6$	0.00462962963	0.00462979683	0.00462962554	0.00462962982
$1/4$	0.01562500000	0.01562510533	0.01562500074	0.01562499998
$1/3$	0.03703703704	0.03703629382	0.03703707090	0.03703703603
$5/12$	0.07233796296	0.07233970584	0.07233791123	0.07233796502
$1/2$	0.12500000000	0.12500014128	0.12499999725	0.12499999984
$7/12$	0.19849537037	0.19849178055	0.19849551759	0.19849536600
$2/3$	0.29629629630	0.29630236028	0.29629611403	0.29629630318
$3/4$	0.42187500000	0.42187499884	0.42187498758	0.42187499953
$5/6$	0.57870370370	0.57869407976	0.57870407228	0.57870369262
$11/12$	0.77025462963	0.77026833822	0.77025420450	0.77025464509
1	1.00000000000	0.99999965644	0.99999997045	0.99999999902

Table 2: Comparison of the L_∞-norm error and convergence order for Example 1

MHFs method [22]			CHFs method		
n	$\|\xi_n\|_\infty$	Rate of convergence	n	$\|\xi_n\|_\infty$	Rate of convergence
4	$4.81704E-03$	3.93	3	$5.64350E-03$	4.22
8	$3.15569E-04$	3.49	6	$3.03330E-04$	4.47
16	$2.81379E-05$	3.88	12	$1.37086E-05$	4.71
32	$1.90987E-06$	3.93	24	$5.24867E-07$	4.85
64	$1.25460E-07$	3.95	48	$1.82605E-08$	4.92
128	$8.09506E-09$	3.97	96	$6.04389E-10$	4.96
256	$5.16785E-10$	$-$	192	$1.950051E-11$	$-$

	Method presented in [14]	Method presented in [5]
n	$\|\xi_n\|_\infty$	$\|\xi_n\|_\infty$
4	$3.52E-09$	$3.5E-04$
8	$1.40E-14$	$1.11022E-16$
16	$1.51E-14$	$1.11022E-16$

Table 3: Numerical results of the L_2-norm error functions $\widetilde{\|\xi_n\|_2}$ for Example 1

n	3	6	12	24	48	96	192
$\widetilde{\|\xi_n\|_2}$	$7.9E-03$	$3.5E-04$	$1.9E-05$	$9.3E-07$	$4.4E-08$	$2.0E-09$	$9.6E-11$
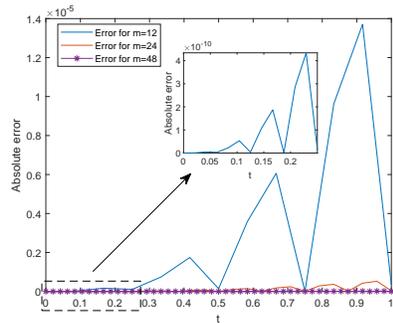
Figure 2: Absolute errors of Example 1, for $n = 12, 24, 48$

Example 2. Consider the following nonlinear fractional integral-differential equation with weakly singular kernel [22] :

$$\,^{C}_{\circ}D^\alpha_t u(t) = g(t) + p(t)u(t) + \int_0^t (t-s)^{-\beta} u^2(s)ds, \quad t \in [0,1],$$

$$g(t) = \frac{3\Gamma\left(\frac{1}{2}\right)}{4\Gamma\left(\frac{11}{6}\right)} t^{\frac{5}{6}} - t^{\frac{5}{2}} - \frac{32}{35} t^{\frac{7}{2}},$$

$$p(t) = t, \quad u(0) = 0.$$
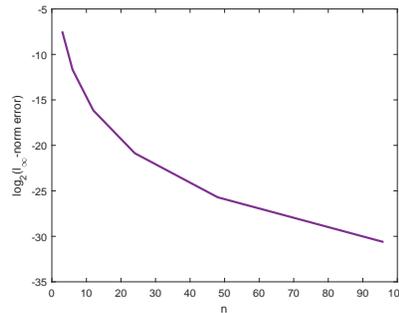
Figure 3: Logarithm of the L_∞-norm error in Example 1
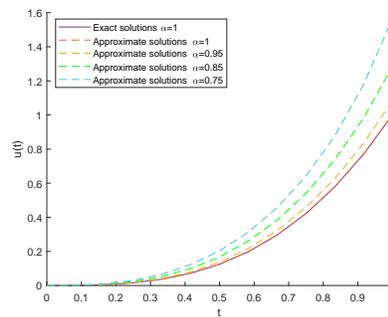
Figure 4: Exact and approximate solutions of Example 1, for $n = 12$

For $\alpha = \frac{2}{3}$ and $\beta = \frac{1}{2}$, the analytic solution is $u(t) = t^{\frac{3}{2}}$. Approximate numerical results using different values of n are shown in Tables 4–6, and Figures 5–7. Table 4 shows the approximate and exact solutions to the problem at some points. Table 5 indicates the L_∞-norm errors and convergence orders for various values of n. Table 6 shows the L_2-norm errors at some values of n. As can be compared in Table 5, this new method provides a higher order of convergence compared to the other method. Figure 5 indicates the absolute errors of Example 2, at $n = 12, 24, 48$. Figure 6 shows that the logarithm of the L_∞-norm error decreases as n increases. A comparison between the changes in the fractional orders of the equation is shown in Figure 7.

Example 3. Consider the following linear fractional integral-differential equation with weakly singular kernel [22]:

$$
{}^{C}_{\circ}D^{\alpha}_{t}u(t) = g(t) + p(t)u(t) + \int_0^t (t-s)^{-\beta}u(s)ds, \quad t \in [0,1],
$$

$$
g(t) = \frac{6t^{\frac{8}{3}}}{\Gamma\left(\frac{11}{3}\right)} + \left(\frac{32}{35} - \frac{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{7}{3}\right)}{\Gamma\left(\frac{17}{6}\right)}\right)t^{\frac{11}{6}} + \Gamma\left(\frac{7}{3}\right)t,
$$

Table 4: Numerical results of Example 2

Points s	Exact solutions $u(s)$	Approximate solutions $n = 12$	Approximate solutions $n = 24$	Approximate solutions $n = 48$
0	0.00000000	0.00000000	0.00000000	0.00000000
$^1/_{12}$	0.02405626	0.02340582	0.02394753	0.02400902
$^1/_6$	0.06804138	0.06771773	0.06790284	0.06800459
$^1/_4$	0.12500000	0.12449685	0.12487612	0.12496621
$^1/_3$	0.19245009	0.19199415	0.19232923	0.19241677
$^5/_{12}$	0.26895718	0.26849729	0.26883160	0.26892254
$^1/_2$	0.35355339	0.35305089	0.35341692	0.35351566
$^7/_{12}$	0.44552819	0.44495789	0.44537319	0.44548528
$^2/_3$	0.54433105	0.54365866	0.54414702	0.54428012
$^3/_4$	0.64951905	0.64868490	0.64929131	0.64945599
$^5/_6$	0.76072577	0.75964752	0.76043179	0.76064433
$^{11}/_{12}$	0.87764152	0.87619255	0.87724539	0.87753181
1	1.00000000	0.99796079	0.99944320	0.99984577

Table 5: Comparison of the L_∞-norm error and convergence order for Example 2

MHFs method [22]			CHFs method		
n	$\|\xi_n\|_\infty$	Rate of convergence	n	$\|\xi_n\|_\infty$	Rate of convergence
4	$1.39991E - 02$	2.08	3	$4.05382E - 02$	2.38
8	$3.30803E - 03$	1.90	6	$7.80949E - 03$	1.94
16	$8.84780E - 04$	1.84	12	$2.03921E - 03$	1.87
32	$2.46509E - 04$	1.83	24	$5.56795E - 04$	1.85
64	$6.92324E - 05$	1.51	48	$1.54232E - 04$	1.84
128	$2.42498E - 05$	1.50	96	$4.30102E - 05$	1.84
256	$8.57216E - 06$	$-$	192	$1.20325E - 05$	$-$

Table 6: Numerical results of the L_2-norm error functions $\widetilde{\|\xi_n\|_2}$ for Example 2

n	3	6	12	24	48	96	192
$\widetilde{\|\xi_n\|_2}$	$3.5E - 02$	$7.2E - 03$	$2.4E - 03$	$8.5E - 04$	$3.3E - 04$	$1.3E - 04$	$5.0E - 05$

$$p(t) = -\frac{32}{35}t^{\frac{1}{2}}, \quad u(0) = 0.$$

For $\alpha = \frac{1}{3}$ and $\beta = \frac{1}{2}$, the analytic solution is $u(t) = t^3 + t^{\frac{4}{3}}$. Tables 7–9 and Figures 8–10 show approximate numerical results using different values of n. Table 7 indicates the approximate and exact solutions to the problem at some grid points. Table 7 shows the advantage of the proposed method compared to the MHF method by presenting the order of convergence and the maximum norm error. Figure 8 shows the behavior of absolute errors for Example 3. Figure 9 shows the logarithm of the L_∞-norm errors. In
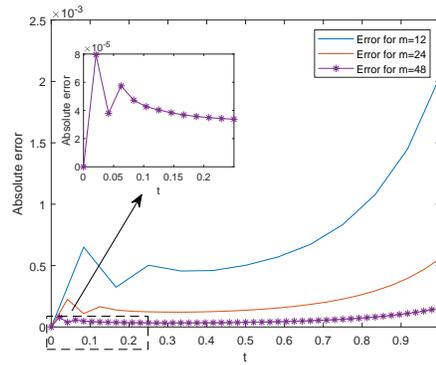
Figure 5: Absolute errors of Example 2, for $n = 12, 24, 48$
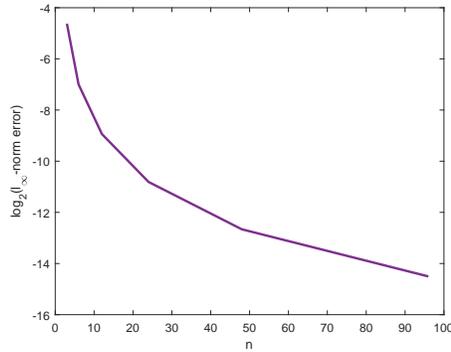

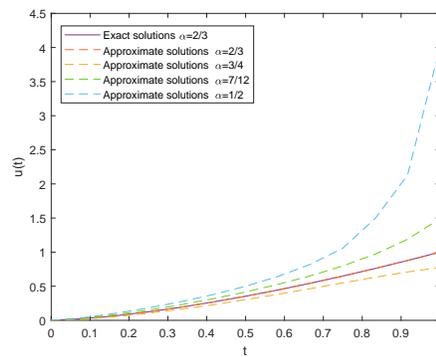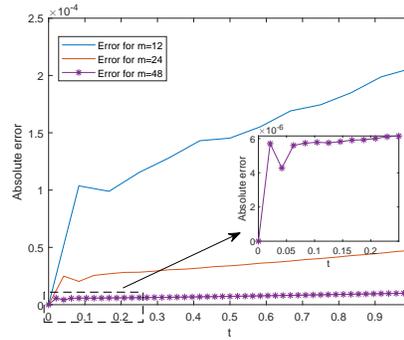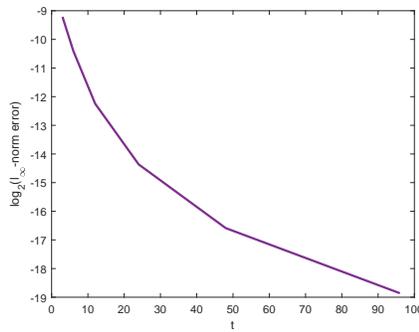Figure 6: Logarithm of the L_∞-norm error in Example 2

Figure 7: Exact and approximate solutions of Example 2, for $n = 12$

addition, a comparison of the results for different values of α with the exact solution of the equation is shown in Figure 10.

Table 7: Numerical results of Example 3

Points s	Exact solutions $u(s)$	Approximate solutions $n = 12$	Approximate solutions $n = 24$	Approximate solutions $n = 48$
0	0.00000000	0.00000000	0.00000000	0.00000000
$1/12$	0.03697789	0.03708156	0.03699820	0.03698362
$1/6$	0.09634983	0.09644877	0.09637681	0.09635574
$1/4$	0.17311513	0.17323059	0.17314347	0.17312128
$1/3$	0.26815746	0.26828582	0.26818799	0.26816396
$5/12$	0.38354663	0.38368969	0.38357860	0.38355352
$1/2$	0.52185026	0.52199562	0.52188416	0.52185755
$7/12$	0.68589934	0.68605454	0.68593554	0.68590705
$2/3$	0.87868327	0.87885234	0.87872136	0.87869145
$3/4$	1.10329522	1.10346969	1.10333552	1.10330386
$5/6$	1.36290039	1.36308529	1.36294318	1.36290952
$11/12$	1.66071634	1.66091514	1.66076132	1.66072598
1	2.00000000	2.00020619	2.00004743	2.00001015

Table 8: Comparison of the L_∞-norm error and convergence order for Example 3

MHFs method [22]			CHFs method		
n	$\|\xi_n\|_\infty$	Rate of convergence	n	$\|\xi_n\|_\infty$	Rate of convergence
4	$1.14967E - 03$	1.86	3	$1.68472E - 03$	1.20
8	$3.15569E - 04$	1.83	6	$7.35848E - 04$	1.84
16	$8.90424E - 05$	2.20	12	$2.06192E - 04$	2.12
32	$1.93665E - 05$	2.30	24	$4.74299E - 05$	2.22
64	$3.94225E - 06$	2.09	48	$1.01541E - 05$	2.27
128	$9.26153E - 07$	2.13	96	$2.10201E - 06$	2.30
256	$2.12088E - 07$	$-$	192	$4.27476E - 07$	$-$

Table 9: Numerical results of the L_2-norm error $\widetilde{\|\xi_n\|_2}$ for Example 3

n	3	6	12	24	48	96	192
$\widetilde{\|\xi_n\|_2}$	$1.2E - 03$	$1.0E - 03$	$4.2E - 04$	$1.4E - 04$	$4.2E - 05$	$1.2E - 05$	$3.6E - 06$

8 Conclusion

In this paper, we proposed a numerical scheme for solving a class of non-linear fractional integral-differential equations with weakly singular kernels

Figure 8: Absolute errors of Example 3, for $n = 12, 24, 48$

Figure 9: Logarithm of the L_∞-norm error in Example 3
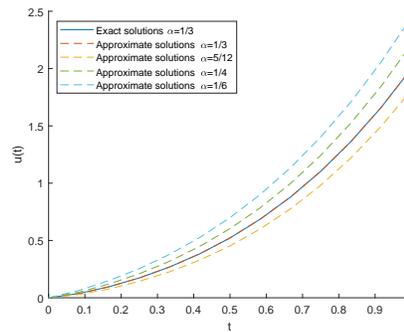
Figure 10: Exact and approximate solutions of Example 3, for $n = 12$

based on CHFs. CHF and the corresponding operational matrix were introduced. The proposed method transforms the original problem into an iterative algorithm, including polynomial equations with three unknown co-

efficients, using the fractional-order operational matrix of integration. An analysis of the method's absolute errors and convergence was conducted. In order to validate the accuracy and effectiveness of this new method, three numerical examples were presented. In Example 1, the absolute error is lower at the nodal points near the beginning of the interval, as shown in Figure 2. Table 2 shows that the new method offers more accurate solutions at the same lengths h than the MHFs approaches. In Examples 2 and 3, the error clearly increases as the time variable approaches one; see Figures 5 and 8, respectively. A study of the results shows that, generally, as n increases, the accuracy of the approximate solution increases, and the absolute error decreases. One of the advantages of this proposed algorithm is that instead of solving a system of $(n+1) \times (n+1)$ equations, it only needs to solve $n/3$ systems of three-variable nonlinear equations. In addition, the order of convergence for the cubic hat functions is $O(h^4)$, while the order of convergence for the generalized hat functions method [6] and the MHFs method [22] are $O(h^2)$ and $O(h^3)$, respectively. Finally, the proposed method (CHFs) can be used for a large number of similar problems, and we will continue to work on developing this method.

Acknowledgements

References

[1] Abdelaty, A.M., Roshdy, M., Said, L.A. and Radwan, A. G. *Numerical simulations and FPGA implementations of fractional-order systems based on product integration rules*, IEEE Access. 8 (2020), 102093–102105.

[2] Ahmed, S.A., Elzaki, T.M. and Hassan, A.A. *Solution of integral differential equations by new double integral transform (Laplace–Sumudu transform)*, Abstr. Appl. Anal. (2020), Art. ID 4725150, 7.

[3] Al-Ahmad, S., Sulaiman, I.M. and Mamat, M. *An efficient modification of differential transform method for solving integral and integrodifferential equations*, Aust. J. Math. Anal. Appl. 17(2) (2020), Art. 5, 15.

[4] Amin, R., Shah, K., Asif, M., Khan, I. and Ullah, F. *An efficient algorithm for numerical solution of fractional integro-differential equations via Haar wavelet*, J. Comput. Appl. Math. 381(113028) (2021), 17.

[5] Arsalan Sajjadi, S., Saberi Najafi, H. and Aminikhah, H. *A numerical study on the non-smooth solutions of the nonlinear weakly singular fractional Volterra integro-differential equations*, Math. Methods Appl. Sci. 46(4) (2023), 4070–4084.

[6] Babolian, E. and Mordad, M. *A numerical method for solving systems of linear and nonlinear integral equations of the second kind by hat basis functions*, Comput. Math. Appl. 62(1) (2011), 187–198.

[7] Behera, S. and Ray, S. Saha. *An efficient numerical method based on Euler wavelets for solving fractional order pantograph Volterra delay-integro-differential equations*, J. Comput. Appl. Math. 406(113825) (2022), 23.

[8] Behera, S. and Ray, S. Saha. *On a wavelet-based numerical method for linear and nonlinear fractional Volterra integro-differential equations with weakly singular kernels*, Comput. Appl. Math. 41(5) (2022), 1–32.

[9] Biazar, J. *Solution of systems of integral–differential equations by Adomian decomposition method*, Appl. Math. Comput. 168(2) (2005), 1232–1238.

[10] Biazar, J. and Ebrahimi, H. *Orthonormal Bernstein polynomials for Volterra integral equations of the second kind*, Int. J. Appl. Math. Res. 9(1) (2019), 9–20.

[11] Biazar, J. and Ebrahimi, H. *A numerical algorithm for a class of nonlinear fractional Volterra integral equations via modified hat functions*, J. Integral Equ. Appl. 34(3) (2022), 295–316.

[12] Biazar, J. and Montazeri, R. *Optimal homotopy asymptotic and multistage optimal homotopy asymptotic methods for solving system of volterra integral equations of the second kind*, J. Appl. Math. 2019 (2019), Art. ID 3037273, 17.

[13] Derakhshan, M. *A numerical scheme based on the Chebyshev functions to find approximate solutions of the coupled nonlinear sine-Gordon equations with fractional variable orders*, Abstr. Appl. Anal. 2021 (2021), Art. ID 8830727, 20.

[14] Du, H., Chen, Z. and Yang, T. *A stable least residue method in reproducing kernel space for solving a nonlinear fractional integro-differential equation with a weakly singular kernel*, Appl. Numer. Math. 157 (2020), 210–222.

[15] Hashemi Mehne, S.H. *Differential transform method: A comprehensive review and analysis*, Iranian Journal of Numerical Analysis and Optimization 12(3) (Special Issue), (2022), 629–657.

[16] Kukreja, V. K. *An improvised collocation algorithm with specific end conditions for solving modified Burgers equation*, Numer. Methods Partial Differ. Equ. 37(1) (2021), 874–896.

[17] Kythe, P.K. and Puri, P. *Computational methods for linear integral equations*, Birkhäuser Boston, Inc., Boston, MA, 2002.

[18] Leitman, M.J. *An integro-differential equation for plane waves propagating into a random fluid: asymptotic behavior*, SIAM J. Math. Anal. 12(4) (1981), 560–571.

[19] Mirzaee, F. and Hadadiyan, E. *Numerical solution of Volterra–Fredholm integral equations via modification of hat functions*, Appl. Math. Comput. 280 (2016), 110–123.

[20] Moosavi Noori, S.R. and Taghizadeh, N. *Modified differential transform method for solving linear and nonlinear pantograph type of differential and Volterra integro-differential equations with proportional delays*, Adv. Differ. Equ. 2020(1) (2020), 1–25.

[21] Ndiaye, A. and Mansal, F. *Existence and uniqueness results of Volterra-Fredholm integro-differential equations via Caputo fractional derivative*, J. Math. (2021), Art. ID 5623388, 8.

[22] Nemati, S. and Lima, P M. *Numerical solution of nonlinear fractional integro-differential equations with weakly singular kernels via a modification of hat functions*, Appl. Math. Comput. 327 (2018), 79–92.

[23] Özaltun, G., Konuralp, A. and Gümgüm, S. *Gegenbauer wavelet solutions of fractional integro-differential equations*, J. Comput. Appl. Math. 420(114830) (2023), 11.

[24] Podlubny, I. *Fractional differential equations*, Math. Sci. Eng. 198 (1999), Academic Press.

[25] Qiao L. and Xu, D. *A fast ADI orthogonal spline collocation method with graded meshes for the two-dimensional fractional integro-differential equation*, Adv. Comput. Math. 47(5) (2021), 1–22.

[26] Quentin, R., King, J.R., Sallard, E., Fishman, N., Thompson, R., Buch E.R. and Cohen, L.G. *Differential brain mechanisms of selection and maintenance of information during working memory* J. Neurosci. 39(19) (2019), 3728–3740.

[27] Rabbath, C.A. and Corriveau, D. *A comparison of piecewise cubic Hermite interpolating polynomials, cubic splines and piecewise linear functions for the approximation of projectile aerodynamics*, Defence Technology 15(5) (2019), 741–757.

[28] Rakshit G. and Rane, A.S. *Asymptotic expansion of iterated Galerkin solution of Fredholm integral equations of the second kind with Green's kernel*, J. Integral Equ. Appl. 32(4) (2020), 495–507.

[29] Riahi Beni, M. *Legendre wavelet method combined with the Gauss quadrature rule for numerical solution of fractional integro-differential equations*, Iranian Journal of Numerical Analysis and Optimization 12(1) (2022), 229–249.

[30] Sabatier, J., Aoun, M., Oustaloup, A., Gregoire, G., Ragot, F. and Roy, P. *Fractional system identification for lead acid battery state of charge estimation*, Signal Process 86(10) (2006), 2645–2657.

[31] Vinagre, B.M. ,Monje, C.A. ,Calderón A.J. and Suárez, J.I. *Fractional PID controllers for industry application. A brief introduction*, J. Vib. Control, 13(9-10) (2007), 1419–1429.

[32] Wang, Y. and Zhu, Li. *SCW method for solving the fractional integro-differential equations with a weakly singular kernel*, Appl. Math. Comput. 275 (2022), 72–80.

[33] Xie, J., Wang, T., Ren, Z., Zhang J. and Quan, L. *Haar wavelet method for approximating the solution of a coupled system of fractional-order integral–differential equations*, Math. Comput. Simul 163 (2019), 80–89.

[34] Yang, Z. *Gröbner Bases for Solving Multivariate Polynomial Equations*, Computing Equilibria and Fixed Points: The Solution of Nonlinear Inequalities (1999) 265–288.

How to cite this article
Ebrahimi, H. and Biazar, J., Cubic hat-functions approximation for linear and nonlinear fractional integral-differential equations with weakly singular kernels. *Iran. j. numer. anal. optim.*, 2023; 13(3): 500-531.
https://doi.org/10.22067/ijnao.2022.73126.1209

Nearest fuzzy number of type L-R to an arbitrary fuzzy number with applications to fuzzy linear system

S.M. Alavi

Abstract

The fuzzy operations on fuzzy numbers of type L-R are much easier than general fuzzy numbers. It would be interesting to approximate a fuzzy number by a fuzzy number of type L-R. In this paper, we state and prove two significant application inequalities in the monotonic functions set. These inequalities show that under a condition, the nearest fuzzy number of type L-R to an arbitrary fuzzy number exists and is unique. After that, the nearest fuzzy number of type L-R can be obtained by solving a linear system. Note that the trapezoidal fuzzy numbers are a particular case of the fuzzy numbers of type L-R. The proposed method can represent the nearest trapezoidal fuzzy number to a given fuzzy number. Finally, to approximate fuzzy solutions of a fuzzy linear system, we apply our idea to construct a framework to find solutions of crisp linear systems instead of the fuzzy linear system. The crisp linear systems give the nearest fuzzy numbers of type L-R to fuzzy solutions of a fuzzy linear system. The proposed method is illustrated with some examples.

1 Introduction

One of the most important topics related to fuzzy mathematics is to study fuzzy numbers. Fuzzy numbers were first introduced by Zadeh [22, 23, 24]

Seyed Majid Alavi

Department of Mathematics, Faculty of Basic Sciences, Arak Branch, Islamic Azad University, Arak, Iran. e-mail: sm.alavi@iau.ac.ir

and then were developed by other researchers [16, 11]. Fuzzy numbers are important essential tools to represent uncertainty models, formalize fuzzy variable functions, defuzzification [3], and linguistic variables [4]. The researchers in [2] proposed an efficient approach to assigning a distance between fuzzy numbers and describing a pseudo-metric on the set of fuzzy numbers and a metric on the set of trapezoidal fuzzy numbers. Trapezoidal fuzzy numbers have a good application in dealing with uncertain information. In fact, to describe the specification of some uncertain events, we have to use fuzzy numbers. Voxman [20] represented two canonical representations of discrete fuzzy numbers. Nevertheless, some fuzzy numbers are too complicated. Hence, the approximation of general fuzzy numbers with regular fuzzy numbers like fuzzy numbers of type L-R, help decision-makers to make better decisions. Trapezoidal fuzzy numbers and triangular fuzzy numbers are particular cases of fuzzy numbers of type L-R. Abbasbandy et al. showed that the nearest trapezoidal fuzzy number to a given fuzzy number exists and is unique. Hajjari [12] used the concept of 0.5-Level and mean Core to approximate fuzzy numbers. In [13], authors introduced a trapezoidal approximation of an arbitrary fuzzy number by Core and support of the fuzzy numbers. Some researchers presented efficient methods to find the nearest trapezoidal fuzzy number or triangular fuzzy number to a given fuzzy number [8, 21, 5, 10]. Lucian Coroianu [9] proved that quadratic programs give the nearest trapezoidal approximation of general fuzzy numbers with respect to weighted metrics with or without additional constraints. Amirfakhrian and Bagherian [6, 7] represented a parametric distance and used it to find the nearest approximation of a given fuzzy Number. Zhou, Yang, and Wang [25] represented fuzzy arithmetic on L-R fuzzy numbers and showed that the proposed model could be transferred to an equivalent crisp programming model by the operational law and then solved with the aid of some well-developed optimization software packages. Ghanbari et al. [14] used an effective approximate multiplication operation on L-R fuzzy numbers and their application.

In this paper, we present two inequalities in monotonic function. These inequalities show that under a condition, the nearest fuzzy number of type L-R to a given fuzzy number exists and is unique. For this purpose, we represent a constrained optimization problem and prove that it has a unique solution. The unique solution is obtained by solving a linear system. Since the trapezoidal fuzzy numbers are a kind of fuzzy number of type L-R, our method can find the nearest trapezoidal fuzzy number to an arbitrary fuzzy number. Here some examples are given to illustrate the main results. Due to the presented method, it is easy to obtain the nearest fuzzy numbers of type L-R to the solutions of a fuzzy linear system, fuzzy linear differential equations, or fuzzy linear integral equations, etc. For instance, finding the nearest fuzzy numbers of type L-R to fuzzy solutions of a linear system is explained. An example is given to approximate a fuzzy solution to a 2×2 fuzzy linear system. In Section 2, we recall some notations and basic definitions of

fuzzy sets and fuzzy numbers. In Section 3, two basic inequalities and theorems are stated and proved, and the proposed method is described. Section 5 represents our method with an example to approximate fuzzy solutions of a fuzzy linear system with L-R fuzzy numbers.

2 Notations and basic definitions

The concept of real numbers is generalized to the concept of fuzzy numbers. Fuzzy numbers have been defined based on their membership functions as below.

Definition 1. [17] A fuzzy number u is a fuzzy set of the real line with the following conditions:
(i) u is normal.
(ii) The support of u is bounded.
(iii) The membership function of u is continuous and convex.

The set of all such fuzzy numbers is represented by E^1. Considering four real numbers $\alpha_1 \leq \alpha_2 \leq \alpha_3 \leq \alpha_4$, the membership function of fuzzy numbers can be introduced as the following form:

$$u(x) = \begin{cases} 0 & if \quad x < \alpha_1, \\ u_1(x) & if \ \alpha_1 \leq x < \alpha_2, \\ 1 & if \quad \alpha_2 \leq x \leq \alpha_3, \\ u_2(x) & if \ \alpha_3 \leq x < \alpha_4, \\ 0 & if \quad \alpha_4 < x, \end{cases}$$

in which $u_1 : [\alpha_1, \alpha_2] \to [0, 1]$ is a nondecreasing function and $u_2 : [\alpha_3, \alpha_4] \to [0, 1]$ is a nonincreasing function. The fuzzy number as the following form, completely characterized by four real numbers $\alpha_1 \leq \alpha_2 \leq \alpha_3 \leq \alpha_4$, is called an *L-R fuzzy number*:

$$u(x) = \begin{cases} 0 & if \quad x < a - \alpha, \\ L(\frac{a-x}{\alpha}) & if \ a - \alpha \leq x < a, \\ 1 & if \ a \leq x \leq b, \\ R(\frac{x-b}{\beta}) & if \ b \leq x \leq b + \beta, \\ 0 & if \ x > b + \beta, \end{cases}$$

in which, $\alpha_1 = a - \alpha$, $\alpha_2 = a$, $\alpha_3 = b$, $\alpha_4 = b + \beta$, and

$$L : [0, 1] \to [0, 1], \quad R : [0, 1] \to [0, 1]$$

are continuous and decreasing shape functions such that $L(0) = R(0) = 1$ and $L(1) = R(1) = 0$. It is mostly denoted in short as $u = (a, b, \alpha, \beta)_{LR}$.

If $L(x) = R(x)$, then u is denoted by $u = (a, b, \alpha, \beta)_L$. Suppose that $u = (a_1, b_1, \alpha_1, \beta_1)_L$ and $v = (a_2, b_2, \alpha_2, \beta_2)_L$ are two fuzzy numbers of type L-L and that $k \in R$. Then the following statements hold:

1) $u + v = (a_1 + a_2, b_1 + b_2, \alpha_1 + \alpha_2, \beta_1 + \beta_2)_L$.

2) $ku(x) = \begin{cases} (ka_1, kb_1, k\alpha_1, k\beta_1)_L & if \ k \geq 0, \\ (kb_1, ka_1, k\beta_1, k\alpha_1)_L & if \ k < 0, \end{cases}$

3) $u - v = (a_1 - b_2, a_2 - b_1, \alpha_1 + \alpha_2, \beta_1 + \beta_2)_L$.

Definition 2. Following [15], we show an arbitrary fuzzy number by an ordered pair of functions $(\underline{u}(r), \overline{u}(r))$; $0 \leq r \leq 1$, that satisfy the following conditions:

1) $\underline{u}(r)$ is a bounded left-continuous nondecreasing over $[0, 1]$.

2) $\overline{u}(r)$ is a bounded left-continuous nonincreasing over $[0, 1]$.

3) $\underline{u}(r) \leq \overline{u}(r)$, $0 \leq r \leq 1$.

If $u = (\underline{u}(r), \overline{u}(r))$ and $v = (\underline{v}(r), \overline{v}(r))$ are two fuzzy numbers, then the following conditions holds

1.
$$\underline{u + v} = \underline{u} + \underline{v}, \qquad \overline{u + v} = \overline{u} + \overline{v}, \tag{1}$$

2.
$$\underline{ku} = k\underline{u}, \qquad \overline{ku} = k\overline{u} \quad if \ \ k \geq 0, \tag{2}$$

3.
$$\underline{ku} = k\overline{u}, \qquad \overline{ku} = k\underline{u} \quad if \ \ k < 0. \tag{3}$$

The parametric form of fuzzy number of type $u = (a, b, \alpha, \beta)_{LR}$ is represented as
$$u = (a - \alpha L^{-1}(r), b + \beta R^{-1}(r)).$$

Definition 3. [18, 3] Let A and B be two arbitrary fuzzy numbers. A distance between A and B is denoted by $D(A, B)$ and defined as below:

$$D(A, B) = \{\int_0^1 (\underline{A}(r) - \underline{B}(r))^2 dr + \int_0^1 (\overline{A}(r) - \overline{B}(r))^2 dr\}^{\frac{1}{2}}. \tag{4}$$

Note that $D(A, B)$ is metric in E^1 and (E^1, D) is a complete space. It is obviously that
$$A = B \Leftrightarrow \underline{A}(r) = \overline{B}(r), \qquad r \in [0, 1]. \tag{5}$$

3 Approximation of general fuzzy number by a given L-R fuzzy number

Let A be an arbitrary fuzzy number. We try to approximate it by a known fuzzy number of type L-R, $N(A) = (a, b, \alpha, \beta)_{LR}$ such that $N(A)$ is the nearest to A with respect to the certain distance in (4). Hence an optimal problem is considered, and then an optimal solution is obtained with an easy method. For this purpose, we denote $N(A)$ as

$$N(A) = (a, a + \xi_1, \xi_2, \xi_3)_{LR},$$

in which $\xi_i \geq 0$ for $i = 1, 2, 3$, and $b = a + \xi_1, \alpha = \xi_2, \beta = \xi_3$. For finding the nearest fuzzy number of type L-R to the arbitrary fuzzy number A, we have to solve the optimization problem as below:

$$\begin{cases} \min D(A, N(A))(a, \xi_1, \xi_2, \xi_3) \\ s.t. \\ \xi_i \geq 0, \quad i = 1, 2, 3, \end{cases} \tag{6}$$

in which

$$D(A, N(A))(a, \xi_1, \xi_2, \xi_3) = \int_0^1 (\underline{A}(r) - a + \xi_1 L^{-1}(r))^2 dr$$
$$+ \int_0^1 (\overline{A}(r) - (a + \xi_2 + \xi_3 R^{-1}(r)))^2 dr.$$

To discuss the existence and uniqueness of the solutions to the optimization problem (6), we need to represent some important inequalities as below:

Lemma 1. Let f and g be two integrable functions from an interval $[a, b]$ to R. Then the following inequalities hold:

1. If g is a nondecreasing (nonincreasing) function and f is a nonincreasing (nondecreasing) function, then

$$\int_a^b f(x)g(x)dx \leq \frac{\int_a^b f(x)dx \int_a^b g(x)dx}{b - a}. \tag{7}$$

2. If g and f are nonincreasing (nondecreasing) functions over $[a, b]$, then

$$\int_a^b f(x)g(x)dx \geq \frac{\int_a^b f(x)dx \int_a^b g(x)dx}{b - a}. \tag{8}$$

3. Regarding to part 1, if f and g are not constant functions over $[a, b]$, then inequality (7) becomes

$$\int_a^b f(x)g(x)dx < \frac{\int_a^b f(x)dx \int_a^b g(x)dx}{b-a}. \tag{9}$$

4. Regarding to part 2, if f and g are not constant functions over $[a, b]$, then inequality (8) becomes

$$\int_a^b f(x)g(x)dx > \frac{\int_a^b f(x)dx \int_a^b g(x)dx}{b-a}. \tag{10}$$

Proof. Since f and g are monotonic functions over $[a, b]$, they are integrable functions. Without loss of generality, assume that g is a nonincreasing and that f is a nondecreasing function over the interval $[a, b]$. For each $x, y \in [a, b]$, if $x \leq y$, then $f(x) \leq f(y)$ and $g(y) \leq g(x)$. Similarly, if $y \leq x$, then $f(x) \geq f(y)$ and i $g(y) \geq g(x)$. Thus for each $x, y \in [a, b]$, we have $(f(x) - f(y))(g(y) - g(x)) \geq 0$. we conclude that

$$\frac{1}{2}\{\int_a^b \int_a^b (f(x) - f(y))(g(y) - g(x))dydx\} \geq 0. \tag{11}$$

Hence

$$\frac{1}{2} \int_a^b \int_a^b (f(x)g(y) - f(x)g(x) - f(y)g(y) + f(y)g(x))dydx \geq 0. \tag{12}$$

Since

$$\int_a^b \int_a^b f(x)g(y)dydx = \int_a^b \int_a^b f(y)g(x)dydx = \int_a^b f(x)dx \int_a^b g(y)dy$$

$$= \int_a^b f(y)dy \int_a^b g(x)dx,$$

$$\int_a^b \int_a^b f(x)g(x)dydx = \int_a^b \int_a^b f(y)g(y)dydx = (b-a) \int_a^b f(y)g(y)dy$$

$$= (b-a) \int_a^b f(x)g(x)dx,$$

then (12) yields the following inequality:

$$\frac{1}{2}\{ \int_a^b f(x)dx \int_a^b g(y)dy - (b-a) \int_a^b f(x)g(x)dx$$

$$- (b-a) \int_a^b f(y)g(y)dy + \int_a^b f(y)dy \int_a^b g(x)dx \} \geq 0.$$

Therefore

$$\int_a^b f(x)dx \int_a^b g(y)dy - (b-a)\int_a^b f(x)g(x)dx \geq 0.$$

We conclude that

$$\int_a^b f(x)dx \int_a^b g(x)dx \geq (b-a)\int_a^b f(x)g(x)dx.$$

It results that

$$\int_a^b f(x)g(x)dx \leq \frac{\int_a^b f(x)dx \int_a^b g(x)dx}{b-a}.$$

Similarity, inequality (8) can be proved. If g is a nonincreasing function and f is a nondecreasing function over the interval $[a,b]$ such that they are not constant functions over $[a,b]$, then inequality (11) becomes

$$\frac{1}{2}\{\int_a^b \int_a^b (f(x) - f(y))(g(y) - g(x))dydx\} > 0.$$

Similar to what was stated in the proof of the part 1, we obtain (9). Inequality (10) is provided in the same manner.   □

Corollary 1. Suppose that $g : [0,1] \to [0,1]$ is integrable and decreasing shape functions such that $g(0) = 1$ and $g(1) = 0$. Then the following properties hold:

1 If f is a nondecreasing continuous function over $[0,1]$, then $\int_0^1 f(x)g(x)dx \leq \int_0^1 f(x)dx \int_0^1 g(x)dx$.

2 If f is a nonincreasing continuous function over $[0,1]$, then $\int_0^1 f(x)g(x)dx \geq \int_0^1 f(x)dx \int_0^1 g(x)dx$.

3 Since g is not constant, $\int_0^1 (g(x))^2 dx > (\int_0^1 (g(x))dx)^2$.

Theorem 1. Let $A = (\overline{A}(r), \underline{A}(r))$ be the parametric form of the given fuzzy number A. The following inequalities hold:

1 $\int_0^1 \underline{A}(r)L^{-1}(r)dr \leq \int_0^1 \underline{A}(r)dr \int_0^1 L^{-1}(r)dr,$

2 $\int_0^1 \overline{A}(r)R^{-1}(r)dr \geq \int_0^1 \overline{A}(r)dr \int_0^1 R^{-1}(r)dr,$

3 $\int_0^1 (R^{-1}(r))^2 dr > (\int_0^1 R^{-1}(r)dr)^2,$

4 $\int_0^1 (L^{-1}(r))^2 dr > (\int_0^1 L^{-1}(r)dr)^2.$

Proof. We know that the following functions satisfy the hypotheses of Lemma 1 and Corollary 1:

a: $L : [0,1] \to [0,1]$ is a decreasing function,

b: $R : [0, 1] \rightarrow [0, 1]$ is a decreasing function,

c: $\underline{A}(r)$ is a nondecreasing function over $[0, 1]$,

d: $\overline{A}(r)$ is a nonincreasing function over $[0, 1]$.

Then the proof of theorem is deduced. $\qquad\qquad\qquad\qquad\square$

For simplicity in computations, we take

$$p = \int_0^1 L^{-1}(r)dr, \qquad p' = \int_0^1 (L^{-1}(r))^2 dr,$$
$$q = \int_0^1 R^{-1}(r)dr, \qquad q' = \int_0^1 (R^{-1}(r))^2 dr. \qquad (13)$$

Also, we take

$$\gamma_1 = \int_0^1 \underline{A}(r)dr, \qquad \gamma_2 = \int_0^1 L^{-1}(r)\underline{A}(r)dr,$$
$$\gamma_3 = \int_0^1 \overline{A}(r)dr, \qquad \gamma_4 = \int_0^1 R^{-1}(r)\overline{A}(r)dr. \qquad (14)$$

The last theorem immediately gives the following results:

$$p' - p^2 > 0, \quad q' - q^2 > 0, \quad \gamma_2 - p\gamma_1 \leq 0, \quad \gamma_4 - q\gamma_3 \geq 0. \qquad (15)$$

Now we prove that the optimization problem (6) has a unique solution, and then the optimal solution is obtained by solving a linear system

Theorem 2. Let A be an arbitrary fuzzy number. Then the nearest fuzzy number of type L-R to A exists and is unique.

Proof. To find the optimal solution for (6), we take

$$\frac{\partial}{\partial a}D(A, N(A))(a, \xi_1, \xi_2, \xi_3) = 0,$$

$$\frac{\partial}{\partial \xi_1}D(A, N(A))(a, \xi_1, \xi_2, \xi_3) = 0,$$

$$\frac{\partial}{\partial \xi_2}D(A, N(A))(a, \xi_1, \xi_2, \xi_3) = 0,$$

$$\frac{\partial}{\partial \xi_3}D(A, N(A))(a, \xi_1, \xi_2, \xi_3) = 0.$$

Hence we obtain the following system:

$$\begin{cases} \int_0^1 (\underline{A}(r) - a + \xi_1 L^{-1}(r))dr + \int_0^1 (\overline{A}(r) - (a + \xi_2 + \xi_3 R^{-1}(r)))dr = 0, \\[2mm] \int_0^1 (\underline{A}(r) - a + \xi_1 L^{-1}(r))L^{-1}(r)dr = 0, \\[2mm] \int_0^1 (\overline{A}(r) - (a + \xi_2 + \xi_3 R^{-1}(r)))dr = 0, \\[2mm] \int_0^1 (\overline{A}(r) - (a + \xi_2 + \xi_3 R^{-1}(r)))R^{-1}(r)dr = 0. \end{cases}$$

The following linear system is obtained:

$$\begin{cases} 2a - \xi_1 \int_0^1 L^{-1}(r)dr + \xi_2 + \xi_3 \int_0^1 R^{-1}(r)dr = \int_0^1 (\overline{A}(r) + \underline{A}(r))dr, \\[2mm] a \int_0^1 L^{-1}(r)dr - \xi_1 \int_0^1 (L^{-1}(r))^2 dr = \int_0^1 \underline{A}(r)L^{-1}(r)dr, \\[2mm] a + \xi_2 + \xi_3 \int_0^1 R^{-1}(r)dr = \int_0^1 \overline{A}(r)dr, \\[2mm] a \int_0^1 R^{-1}(r)dr + \xi_2 \int_0^1 R^{-1}(r)dr + \xi_3 \int_0^1 (R^{-1}(r))^2 dr = \int_0^1 \overline{A}(r)R^{-1}(r)dr. \end{cases} \tag{16}$$

referring to (13) and (14), the linear equations system (16) becomes

$$\begin{pmatrix} 2 & -p & 1 & q \\ p & -p' & 0 & 0 \\ 1 & 0 & 1 & q \\ q & 0 & q & q' \end{pmatrix} \begin{pmatrix} a \\ \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix} = \begin{pmatrix} \gamma_1 + \gamma_3 \\ \gamma_2 \\ \gamma_3 \\ \gamma_4 \end{pmatrix}.$$

We perform elementary row operations on the coefficient matrix and the right-hand side vector, the following system is obtained:

$$\begin{pmatrix} 1 & -p & 0 & 0 \\ p & -p' & 0 & 0 \\ 1 & 0 & 1 & q \\ q & 0 & q & q' \end{pmatrix} \begin{pmatrix} a \\ \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix} = \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \\ \gamma_4 \end{pmatrix}. \tag{17}$$

The coefficient matrix is a block matrix. Based on the definition of the determinate of matrices, it is obvious that

$$det \begin{pmatrix} 1 & -p & 0 & 0 \\ p & -p' & 0 & 0 \\ 1 & 0 & 1 & q \\ q & 0 & q & q' \end{pmatrix} = det \begin{pmatrix} 1 & -p \\ p & -p' \end{pmatrix} det \begin{pmatrix} 1 & q \\ q & q' \end{pmatrix} = (p^2 - p')(q' - q^2) \neq 0.$$

Due to the results in (15), the linear equations system (17) has a unique solution. To ensure that this linear system gives the optimal solution of (6), we need to prove that $\xi_1, i = 1, 2, 3$, are nonnegative. Consider the first and second equations of the linear equations system (17). We have

$$\begin{cases} a - p\xi_1 = \int_0^1 \underline{A}(r))dr, \\[2mm] pa - p'\xi_1 = \int_0^1 \underline{A}(r)L^{-1}(r)dr. \end{cases} \tag{18}$$

To delete the variable a, we multiply the first row by $-p$ and added the results to the second row. The results give ξ_1 as below:

$$\begin{aligned} \xi_1 &= \frac{p\int_0^1 \underline{A}(r)dr - \int_0^1 \underline{A}(r)L^{-1}(r)dr}{p' - p^2} \\[2mm] &= \frac{\int_0^1 L^{-1}(r)dr \int_0^1 \underline{A}(r)dr - \int_0^1 \underline{A}(r)L^{-1}(r)dr}{p' - p^2}, \end{aligned} \tag{19}$$

By subsisting to (19) and using (15), we have

$$\xi_1 = \frac{p\gamma_1 - \gamma_2}{p' - p^2} \geq 0.$$

From the third and forth equations of (17), we have

$$\begin{cases} a + \xi_2 + q\xi_3 = \int_0^1 \overline{A}(r)dr, \\[2mm] qa + q\xi_2 + q'\xi_3 = \int_0^1 \underline{A}(r)R^{-1}(r)dr. \end{cases}$$

Performing elementary row operations on the linear equations system, a and ξ_2 are deleted, and a new linear equation is obtained as below:

$$\begin{aligned} \xi_3 &= \frac{\int_0^1 \overline{A}(r)R^{-1}(r)dr - q\int_0^1 \overline{A}(r)dr}{q' - q^2} \\[2mm] &= \frac{\int_0^1 \overline{A}(r)R^{-1}(r)dr - \int_0^1 R^{-1}(r)dr \int_0^1 \overline{A}(r)dr}{\int_0^1 (R^{-1}(r))^2 dr - (\int_0^1 R^{-1}(r)dr)^2}. \end{aligned}$$

It means that

$$\xi_3 = \frac{\gamma_4 - q\gamma_3}{q' - q^2}.$$

Due to (15), we conclude that $\xi_3 \geq 0$. Consider the first and third linear equations of (17), then perform the elementary operation for deleting a. We have

$$\xi_2 = \int_0^1 \overline{A}(r)dr - \int_0^1 \underline{A}(r)dr - q\xi_3 - p\xi_1.$$

By substituting ξ_1 and ξ_3, we have

$$\xi_2 = \gamma_3 - \gamma_1 - q\frac{\gamma_4 - q\gamma_3}{q' - q^2} - p\frac{p\gamma_1 - \gamma_2}{p' - p^2}.$$

Due to the assumption and results in (15), we conclude that $\xi_2 \geq 0$. Then $N(A) = (a, a + \xi_2, \xi_1, \xi_3)_{LR}$, is the nearest fuzzy number to A. If $\gamma_3 - \gamma_1 < q\frac{\gamma_4 - q\gamma_3}{q' - q^2} - p\frac{p\gamma_1 - \gamma_2}{p' - p^2}$, then $\xi_2 < 0$ and (a, ξ_1, ξ_2, ξ_3) is a global optimal solution to $\min D(A, N(A))(a, \xi_1, \xi_2, \xi_3)$, but $N(A) = (a, a + \xi_2, \xi_1, \xi_3)_{LR}$, is not a fuzzy number. In this case, we use the quadratic penalty method on (6), as

$$\min D(A, N(A))(a, \xi_1, \xi_2, \xi_3) + c_j \sum_{j=1}^{3} \max\{0, -\xi_j\}, \qquad (20)$$

where $c_j \geq 0$. Due to [19], the optimal solution to (20) exists and is obtained as (a, ξ_1, ξ_2, ξ_3), in which $\xi_1, \xi_2, \xi_3 \geq 0$. Then the nearest fuzzy number of type L-R to A is $N(A) = (a, a + \xi_1, \xi_2, \xi_3)_{LR}$. $\qquad \square$

Corollary 2. Let $n > 0$. Then $A = (-(1-r)^n, (1-r)^n)$ is a fuzzy number. Consider $L^{-1} = R^{-1} = 1 - r$.

1: For $0 < n \leq 1$, we have $\xi_2 \geq 0$. Therefore, (17) gives $N(A) = (a, a + \xi_1, \xi_2, \xi_3)_{LR}$, which is the nearest trapezoidal fuzzy number to A with respect to distance (4). For instance, if $A = (-(1-r)^{0.5}, (1-r)^{0.5})$, then $N(A) = (-0.26, 0.53, 0.53, 0.8)_T$.

2: For $1 < n$, we have $\xi_2 < 0$. Therefore, (17) does not give the nearest trapezoidal fuzzy number to A. For instance, if $A = (-(1-r)^2, (1-r)^2)$ we cannot use linear system (17). Using quadratic penalty method, we have $N(A) = (0, 0, 0.73, 0.73)_T$ that is a triangular fuzzy number.

4 Results and examples

In this section, we represent three examples to show the efficiency of our method. The first example uses the fuzzy number of type L-L, the second example uses the trapezoidal fuzzy number, and the third example uses the fuzzy number of type L-R. Finally, we represent a kind of fuzzy number that may not have the nearest fuzzy number of type L-R.

Example 1. Let us consider the fuzzy number

$$A = \begin{cases} 1 - \frac{(x-5)^2}{4}, & 3 \leq x \leq 7, \\ \\ 0 & otherwise. \end{cases}$$

Take $L(x) = R(x) = \sqrt{1-x}$. Then $L^{-1}(r) = R^{-1}(r) = 1 - r^2$. Obviously the parametric form of A is

$$A = (5 - 2\sqrt{1-r}, 5 + 2\sqrt{1-r}).$$

Considering (17), to obtain the entries of the coefficient matrix, we take

$$p = q = \int_0^1 L^{-1}(r)dr = \int_0^1 (1 - r^2)dr = 0.66,$$

$$p' = q' = \int_0^1 (L^{-1}(r))^2 dr = \int_0^1 (1 - r^2)^2 dr = 0.53,$$

and entries of the right-hand side of the linear system (17) are obtained as below:

$$\gamma_1 = \int_0^1 \underline{A}(r)dr = \int_0^1 (5 - 2\sqrt{1-r})dr = 3.67,$$

$$\gamma_2 = \int_0^1 \underline{A}(r)L^{-1}(r)dr = \int_0^1 (5 - 2\sqrt{1-r})(1 - r^2)dr = 2.30,$$

$$\gamma_3 = \int_0^1 \overline{A}(r)dr = \int_0^1 (5 + 2\sqrt{1-r})dr = 6.33,$$

$$\gamma_4 = \int_0^1 \overline{A}(r)R^{-1}(r)dr = \int_0^1 (5 + 2\sqrt{1-r})(1 - r^2)dr = 4.36.$$

Hence the linear system is obtained as follows:

$$\begin{pmatrix} 1 & -0.66 & 0 & 0 \\ 0.66 & -0.53 & 0 & 0 \\ 1 & 0 & 1 & 0.66 \\ 0.66 & 0 & 0.66 & 0.53 \end{pmatrix} \begin{pmatrix} a \\ \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix} = \begin{pmatrix} 3.67 \\ 2.30 \\ 6.33 \\ 4.36 \end{pmatrix}.$$

Solving this linear system, we have $a = 4.54$, $\xi_1 = 1.32$, $\xi_2 = 0.72$, $\xi_3 = 1.60$. Therefore the nearest fuzzy number of type L-R to A is

$$N(A) = (3.18, 4.50, 0.72, 1.60)_L.$$

Figure 1 compares the graph of A and $N(A)$.

In particular, if $L(r) = R(r) = 1 - r$, then $p = q = \int_0^1 (1 - r)dr = \frac{1}{2}$ and $p' = q' = \int_0^1 (1 - r)^2 dr = \frac{1}{3}$. The linear equations system (17) becomes the following system and gives the nearest trapezoidal fuzzy number to the given fuzzy number $A = (\underline{A}(r), \overline{A}(r))$:

$$\begin{pmatrix} 1 & -\frac{1}{2} & 0 & 0 \\ \frac{1}{2} & -\frac{1}{3} & 0 & 0 \\ 1 & 0 & 1 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{3} \end{pmatrix} \begin{pmatrix} a \\ \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix} = \begin{pmatrix} \int_0^1 \underline{A}(r)dr \\ \int_0^1 (1-r)\underline{A}(r)dr \\ \int_0^1 \overline{A}(r)dr \\ \int_0^1 (1-r)\overline{A}(r)dr \end{pmatrix}. \tag{21}$$

Example 2. Let us consider the fuzzy number
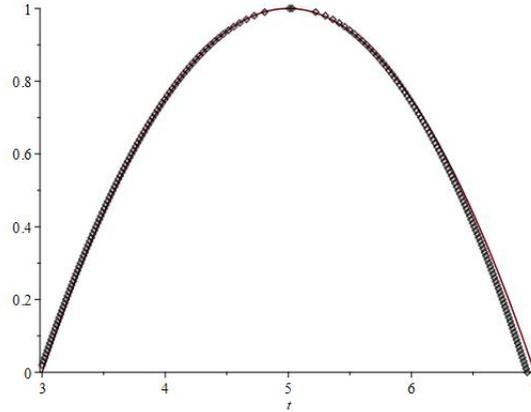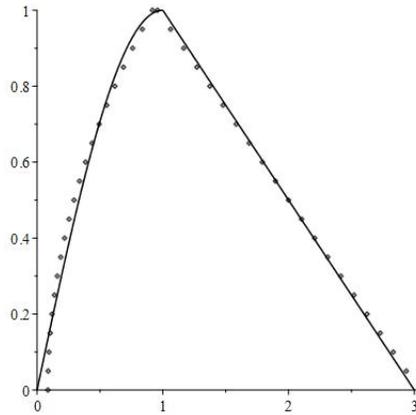
Figure 1: Membership functions of A (line) and $N(A)$ (point), where $N(A)$ is the nearest fuzzy number of type L-R with $L^{-1}(r) = R^{-1}(r) = 1 - r^2$ to A

$$A = \begin{cases} 1 - \frac{(x-5)^2}{4}, & 3 \leq x \leq 7, \\ \\ 0 & otherwise. \end{cases}$$

The parametric form of A is $A = (5 - 2\sqrt{1-r}, 5 + 2\sqrt{1-r})$. For finding the nearest trapezoidal fuzzy number to fuzzy number A, it is enough to take $L(x) = R(x) = 1 - x$ $((L^{-1}(r) = R^{-1}(r) = 1 - r)$ and determine all entries of the right-hand side of linear system (21). Then

$$\gamma_1 = \int_0^1 \underline{A}(r)dr = \int_0^1 (5 - 2\sqrt{1-r})dr = 3.67,$$

$$\gamma_2 = \int_0^1 \underline{A}(r)L^{-1}(r)dr = \int_0^1 (5 - 2\sqrt{1-r})(1-r)dr = 1.70,$$

$$\gamma_3 = \int_0^1 \overline{A}(r)dr = \int_0^1 (5 + 2\sqrt{1-r})dr = 6.33,$$

$$\gamma_4 = \int_0^1 \overline{A}(r)R^{-1}(r)dr = \int_0^1 (5 + 2\sqrt{1-r})(1-r)dr = 3.30.$$

Thus the linear system is obtained as

$$\begin{pmatrix} 1 & -\frac{1}{2} & 0 & 0 \\ \frac{1}{2} & -\frac{1}{3} & 0 & 0 \\ 1 & 0 & 1 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{3} \end{pmatrix} \begin{pmatrix} a \\ \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix} = \begin{pmatrix} 3.67 \\ 1.70 \\ 6.33 \\ 3.30 \end{pmatrix}.$$

By solving this linear system, we have $a = 4.47$, $\xi_1 = 1.60$, $\xi_2 = 1.05$, $\xi_3 = 1.62$. Therefore the nearest trapezoidal fuzzy number to A is $N(A) = (4.47, 4.47 + \xi_1, \xi_2, \xi_3)_T$; that is,

$$N(A) = (4.47, 5.52, 1.60, 1.62)_T.$$

Figure 2 compares the graph of A and $N(A)$.
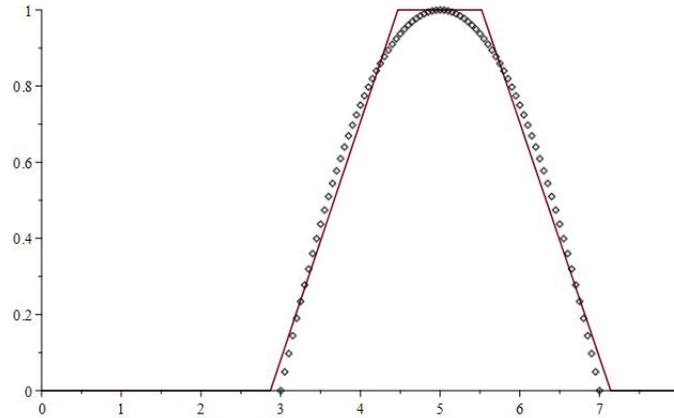
Figure 2: Membership functions of A (point) and $N(A)$ (line), where $N(A)$ is the nearest trapezoidal fuzzy number to A

Example 3. Consider the fuzzy number

$$A = \begin{cases} \sin(\frac{\pi x}{2}), & 0 \le x \le 1, \\ \\ \frac{3-x}{2}, & 1 \le x \le 3, \\ \\ \\ 0, & otherwise. \end{cases}$$

Let $L(x) = \sqrt{1-x}$ and $R(x) = 1 - x$. We want to obtain the nearest fuzzy number of type L-R to A. The parametric form of A is $A = (\frac{2}{\pi} \arcsin(r), 3 - 2r)$. One can easily show that $L^{-1}(r) = 1 - r^2$ and $R^{-1}(r) = 1 - r$. referring to (14), we have $\gamma_1 = \int_0^1 \underline{A}(r)dr = \int_0^1 (\frac{2}{\pi} arcsin(r))dr = 0.363$, $\gamma_2 = \int_0^1 \underline{A}(r)L^{-1}(r)dr = 0.172$, $\gamma_3 = \int_0^1 \overline{A}(r)dr = 2$ and $\gamma_4 = \int_0^1 \overline{A}(r)R^{-1}(r)dr = 1.167$. Hence the linear system is obtained as

$$\begin{pmatrix} 1 & -0.667 & 0 & 0 \\ 0.667 & -0.533 & 0 & 0 \\ 1 & 0 & 1 & 0.500 \\ 0.500 & 0 & 0.500 & 0.333 \end{pmatrix} \begin{pmatrix} a \\ \xi_1 \\ \xi_2 \\ \xi_3 \end{pmatrix} = \begin{pmatrix} 0.363 \\ 0.172 \\ 2.000 \\ 1.167 \end{pmatrix}.$$

Solving this linear system, we have $a = 0.918$, $\xi_1 = 0.832$, $\xi_2 = 0.039$, $\xi_3 = 2.084$. Therefore the nearest fuzzy number of type L-R to A is $N(A) = (0.918, 0.918 + \xi_2, \xi_1, \xi_2)_{LR}$; that is,

$$N(A) = (0.918, 1.147, 0.832, 2.083)_{LR}.$$

Figure 3 compares the graph of A and $N(A)$.

Figure 3: Membership functions of A (line) and $N(A)$ (point), where $N(A)$ is the nearest fuzzy number of type L-R to A with $L^{-1}(r) = 1 - r^2$ and $R^{-1}(r) = 1 - r$

5 Fuzzy linear system

In this section, we focus on the fuzzy linear system as $AX = b$, in which the entries of the right-hand side vector, b, are fuzzy numbers, and entries of the coefficient matrix, A, are real numbers. Using the proposed method, we approximate the fuzzy solutions by L-R fuzzy numbers.

Definition 4. The $n \times n$ linear system

$$\begin{cases} a_{1,1}x_1 + a_{1,2}x_2 + \cdots + a_{1,n}x_n = b_1, \\ a_{2,1}x_1 + a_{2,2}x_2 + \cdots + a_{2,n}x_n = b_2, \\ \vdots \\ a_{n,1}x_1 + a_{n,2}x_2 + \cdots + a_{n,n}x_n = b_n, \end{cases} \tag{22}$$

is called a fuzzy linear system if $b_i = (\underline{b_i}(r), \overline{b_i}(r))$, for $i = 1, 2, \ldots, n$, are fuzzy numbers such that $\xi_{b_j,2} \geq 0$ and $a_{i,j} \in \Re$.

Let $x_j = (\underline{x_j}, \overline{x_j})$ be the fuzzy solutions to (22). Due to (5), (1), (2), and (3), for each i, we have

$$\underline{a_{i,1}x_1} + \underline{a_{i,2}x_2} + \cdots + \underline{a_{i,n}x_n} = \underline{b_i},$$

$$\overline{a_{i,1}x_1} + \overline{a_{i,2}x_2} + \cdots + \overline{a_{i,n}x_n} = \overline{b_i},$$

and then

$$\begin{cases} \sum_{a_{i,j} \geq 0} a_{i,j}\underline{x_j}(r) + \sum_{a_{i,j} \leq 0} a_{i,j}\overline{x_j}(r) = \underline{b_i}(r), & i = 1, 2, \ldots, n, \\[2mm] \sum_{a_{i,j} \geq 0} a_{i,j}\overline{x_j}(r) + \sum_{a_{i,j} \leq 0} a_{i,j}\underline{x_j}(r) = \overline{b_i}(r), & i = 1, 2, \ldots, n. \end{cases} \tag{23}$$

Integrating with respect to r on the interval $[0, 1]$, we have

$$\begin{cases} \sum_{a_{i,j} \geq 0} a_{i,j} \int_0^1 \underline{x_j}(r)dr + \sum_{a_{i,j} \leq 0} a_{i,j} \int_0^1 \overline{x_j}(r)dr = \int_0^1 \underline{b_i}(r)dr, \\ \hspace{9cm} i = 1, 2, \ldots, n, \\[2mm] \sum_{a_{i,j} \geq 0} a_{i,j} \int_0^1 \overline{x_j}(r)dr + \sum_{a_{i,j} \leq 0} a_{i,j} \int_0^1 \underline{x_j}(r)dr = \int_0^1 \overline{b_i}(r)dr, \\ \hspace{9cm} i = 1, 2, \ldots, n. \end{cases} \tag{24}$$

Multiplying both sides of the equations (23) by $L^{-1}(r)$ and integrating, we obtain the following equations:

$$\begin{cases} \sum_{a_{i,j} \geq 0} a_{i,j} \int_0^1 L^{-1}(r)\underline{x_j}(r)dr + \sum_{a_{i,j} \leq 0} a_{i,j} \int_0^1 L^{-1}(r)\overline{x_j}(r)dr \\ \quad = \int_0^1 L^{-1}(r)\underline{b_i}(r)dr, \hspace{3.8cm} i = 1, 2, \ldots, n, \\[2mm] \sum_{a_{i,j} \geq 0} a_{i,j} \int_0^1 L^{-1}(r)\overline{x_j}(r)dr + \sum_{a_{i,j} \leq 0} a_{i,j} \int_0^1 L^{-1}(r)\underline{x_j}(r)dr \\ \quad = \int_0^1 L^{-1}(r)\overline{b_i}(r)dr, \hspace{3.8cm} i = 1, 2, \ldots, n. \end{cases} \tag{25}$$

Multiplying both sides of the equations (23) by $R^{-1}(r)$ and integrating, we obtain the following equations:

$$\begin{cases} \sum_{a_{i,j} \geq 0} a_{i,j} \int_0^1 R^{-1}(r)\underline{x_j}(r)dr + \sum_{a_{i,j} \leq 0} a_{i,j} \int_0^1 R^{-1}(r)\overline{x_j}(r)dr \\ \quad = \int_0^1 R^{-1}(r)\underline{b_i}(r)dr, \hspace{3.8cm} i = 1, 2, \ldots, n, \\[2mm] \sum_{a_{i,j} \geq 0} a_{i,j} \int_0^1 R^{-1}(r)\overline{x_j}(r)dr + \sum_{a_{i,j} \leq 0} a_{i,j} \int_0^1 R^{-1}(r)\underline{x_j}(r)dr \\ \quad = \int_0^1 R^{-1}(r)\overline{b_i}(r)dr, \hspace{3.8cm} i = 1, 2, \ldots, n. \end{cases} \tag{26}$$

We denote $\int_0^1 \underline{x_j}(r)dr$, $\int_0^1 L^{-1}(r)\underline{x_j}(r)dr$, $\int_0^1 \overline{x_j}(r)dr$, $\int_0^1 R^{-1}(r)\overline{x_j}(r)dr$, $\int_0^1 R^{-1}(r)\underline{x_j}(r)dr$, and $\int_0^1 L^{-1}(r)\overline{x_j}(r)dr$, respectively, by $\gamma_{j,1}$, $\gamma_{j,2}$, $\gamma_{j,3}$, $\gamma_{j,4}$, $\gamma_{j,5}$, and $\gamma_{j,6}$. Thus we can rewrite (24), (25), and (26), respectively, as below:

$$\begin{cases} \sum_{a_{i,j} \geq 0} a_{i,j}\gamma_{j,1} + \sum_{a_{i,j} \leq 0} a_{i,j}\gamma_{j,3} = \int_0^1 \underline{b_i}(r)dr, & i = 1, 2, \ldots, n, \\[2ex] \sum_{a_{i,j} \geq 0} a_{i,j}\gamma_{j,3} + \sum_{a_{i,j} \leq 0} a_{i,j}\gamma_{j,1} = \int_0^1 \overline{b_i}(r)dr, & i = 1, 2, \ldots, n, \end{cases} \quad (27)$$

$$\begin{cases} \sum_{a_{i,j} \geq 0} a_{i,j}\gamma_{j,2} + \sum_{a_{i,j} \leq 0} a_{i,j}\gamma_{j,6} = \int_0^1 L^{-1}(r)\underline{b_i}(r)dr, & i = 1, 2, \ldots, n, \\[2ex] \sum_{a_{i,j} \geq 0} a_{i,j}\gamma_{j,6} + \sum_{a_{i,j} \leq 0} a_{i,j}\gamma_{j,2} = \int_0^1 L^{-1}(r)\overline{b_i}(r)dr, & i = 1, 2, \ldots, n, \end{cases} \quad (28)$$

and

$$\begin{cases} \sum_{a_{i,j} \geq 0} a_{i,j}\gamma_{j,5} + \sum_{a_{i,j} \leq 0} a_{i,j}\gamma_{j,4} = \int_0^1 R^{-1}(r)\underline{b_i}(r)dr, & i = 1, 2, \ldots, n, \\[2ex] \sum_{a_{i,j} \geq 0} a_{i,j}\gamma_{j,4} + \sum_{a_{i,j} \leq 0} a_{i,j}\gamma_{j,5} = \int_0^1 R^{-1}(r)\overline{b_i}(r)dr, & i = 1, 2, \ldots, n. \end{cases} \quad (29)$$

The coefficient matrices of the linear equations (27), (28), and (29) are the same. Then it is enough to describe only one of them. Here we describe the linear system (27). Subtracting and adding the first n equations by the second n equations, we obtain two $n \times n$ linear systems as below [1]:

$$I) \begin{cases} a_{1,1}(\gamma_{1,3} + \gamma_{1,1}) + a_{1,2}(\gamma_{2,3} + \gamma_{2,1}) + \cdots + a_{1,n}(\gamma_{n,3} + \gamma_{n,1}) \\ \quad = \int_0^1 \overline{b_1}(r)dr + \int_0^1 \underline{b_1}(r)dr, \\ a_{2,1}(\gamma_{1,3} + \gamma_{1,1}) + a_{2,2}(\gamma_{2,3} + \gamma_{2,1}) + \cdots + a_{2,n}(\gamma_{n,3} + \gamma_{n,1}) \\ \quad = \int_0^1 \overline{b_2}(r)dr + \int_0^1 \underline{b_2}(r)dr, \\ \vdots \\ a_{n,1}(\gamma_{1,3} + \gamma_{1,1}) + a_{n,2}(\gamma_{2,3} + \gamma_{2,1}) + \cdots + a_{n,n}(\gamma_{n,3} + \gamma_{n,1}) \\ \quad = \int_0^1 \overline{b_n}(r)dr + \int_0^1 \underline{b_n}(r)dr, \end{cases} \quad (30)$$

$$II) \begin{cases} a_{1,1}^+(\gamma_{1,3} - \gamma_{1,1}) + a_{1,2}^+(\gamma_{2,3} - \gamma_{2,1}) + \cdots + a_{1,n}^+(\gamma_{n,3} - \gamma_{n,1}) \\ \quad = \int_0^1 \overline{b_1}(r)dr - \int_0^1 \underline{b_1}(r)dr, \\ a_{2,1}^+(\gamma_{1,3} - \gamma_{1,1}) + a_{2,2}^+(\gamma_{2,3} - \gamma_{2,1} + \cdots + a_{2,n}^+(\gamma_{n,3} - \gamma_{n,1}) \\ \quad = \int_0^1 \overline{b_2}(r)dr - \int_0^1 \underline{b_2}(r)dr, \\ \vdots \\ a_{n,1}^+(\gamma_{1,3} - \gamma_{1,1}) + a_{n,2}^+(\gamma_{2,3} - \gamma_{2,1}) + \cdots + a_{n,n}^+(\gamma_{n,3} - \gamma_{n,1}) \\ \quad = \int_0^1 \overline{b_n}(r)dr - \int_0^1 \underline{b_n}(r)dr, \end{cases} \quad (31)$$

in which $a_{i,j}^+ = |a_{i,j}|$. Taking $\gamma_j^c = \gamma_{j,2} + \gamma_{j,1}$ and $\gamma_j^d = \gamma_{j,2} - \gamma_{j,1}$ and solving the linear equations (30) and (31), γ_j^c and γ_j^d are obtained for $j = 1, 2, \ldots, n$, provided that, coefficient matrices are nonsingular. After that, we conclude that $\gamma_{j,1} = 0.5(\gamma_j^c - \gamma_j^d)$ and $\gamma_{j,2} = 0.5(\gamma_j^c + \gamma_j^d)$.

Solving the linear equations (27), (28), and (29), the nearest fuzzy number of type L-R to each x_j are obtained as below:

$$N(x_j) = (a_j, a_j + \xi_{j,2}, \xi_{j,1}, \xi_{j,3})_{LR}, \quad j = 1, 2, \ldots, n,$$

in which

$$\xi_{j,1} = \frac{p\gamma_{j,1} - \gamma_{j,3}}{p' - p^2} \geq 0.$$

$$\xi_{j,2} = \gamma_{j,3} - \gamma_{j.1} - q\frac{\gamma_{j,4} - q\gamma_{j,3}}{q' - q^2} - p\frac{p\gamma_{j,1} - \gamma_{j,2}}{p' - p^2} \geq 0.$$

$$\xi_{j,3} = \frac{p\gamma_{j,1} - \gamma_{j,3}}{p' - p^2} \geq 0.$$

Due to (18), a_j is obtained as

$$a_j = p\xi_{j,1} + \gamma_{j,1}.$$

Example 4. Let us consider the 2×2 fuzzy linear system

$$\begin{cases} 2x_1 + x_2 = (r^2, 2 - r), \\ x_1 - x_2 = (5 - 2\sqrt{1 - r}, 5 + 2\sqrt{1 - r}). \end{cases} \tag{32}$$

For simplicity, suppose that $L^{-1}(r) = R^{-1}(r) = 1 - r$ is linear functions. Then the nearest fuzzy number of type L-R is trapezoidal fuzzy numbers. The exact solutions of system (32) are

$$x_1 = \left(\frac{5r^2 + r + 18}{12} + \sqrt{1 - r}, \frac{-r^2 - 5r + 30}{12} - \sqrt{1 - r}\right),$$

$$x_2 = \left(\frac{-r^2 - 5r - 30}{12} - 3\sqrt{1 - r}, \frac{5r^2 + r - 42}{12} + 3\sqrt{1 - r}\right).$$

One can see that $\underline{x_2}(r) \leq \overline{x_2}(r)$, but $\underline{x_1}(r) \geq \overline{x_1}(r)$. Then x_2 is a fuzzy solution and x_1 is not a fuzzy number. Then it is not a fuzzy solution to (32). Now represent x_1 as $\left(\frac{-r^2 - 5r + 30}{12} - \sqrt{1 - r}, \frac{5r^2 + r + 18}{12} + \sqrt{1 - r}\right)$ that is called weak fuzzy solution. We try to find the nearest trapezoidal fuzzy number to fuzzy solutions x_1 and x_2. Also, since $L^{-1}(r) = R^{-1}(r)$, we need to solve only two linear systems (27) and (28). Solving (27) and (28), the nearest trapezoidal fuzzy numbers to x_1 and x_2 are obtained, respectively, as below:

$$N(x_1) = (1.75, 2.180, 0.306, 0.318)_T,$$

$$N(x_2) = (-3.74, -2.313, 1.975, 1.987)_T.$$

Figures 4 and 5 represent the graphs of the exact fuzzy solutions x_1 and x_2 with their nearest trapezoidal fuzzy numbers.
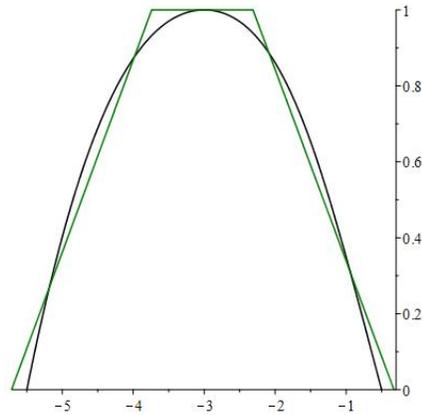

Figure 4: Membership functions of x_1 (curve) and $N(x_1)$ (trapezoidal), where $N(x_1)$ is the nearest fuzzy number of type L-R to x_1 with $L^{-1}(r) = R^{-1}(r) = 1 - r$
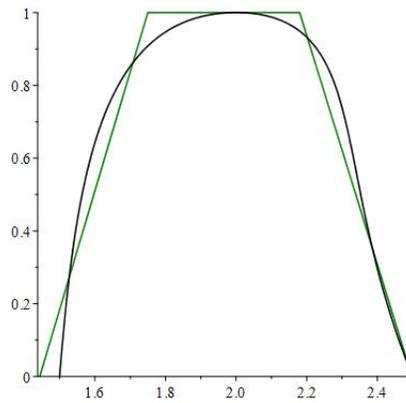

Figure 5: Membership functions of x_2 (curve) and $N(x_2)$ (trapezoidal), where $N(x_2)$ is the nearest trapezoidal fuzzy number to x_2

6 Conclusion

In this study, we focused on the approximate given general fuzzy numbers by fuzzy numbers of type L-R. Fuzzy numbers of type L-R, in particular, trapezoidal fuzzy numbers, play an essential role in the fuzzy environment.

We use a property of linear equations system to obtain the nearest trapezoidal fuzzy number to a given general fuzzy number with respect to the distance formula $D(\cdot, \cdot)$. The presented method is attractive, simple, and can be applied in any way.

References

[1] Abbasabandy, S. and Alavi, M. *A method for solving fuzzy linear systems*, Iran. J. Fuzzy Syst. 2(2) (2005) 37–43.

[2] Abbasbandy, S. and Amirfakhrian, M. *The nearest trapezoidal form of a generalized left right fuzzy number,* Internat. J. Approx. Reason. 43(2) (2006) 166–178.

[3] Abbasbandy, S.and Asady, B. *A note on ''A new approach for defuzzification''[Fuzzy Sets and Systems 111 (2000), no. 3, 351–356; 1 748 552] by M. Ma, A. Kandel and M. Friedman,* Fuzzy Sets Syst. 128 (2002) 131–132.

[4] Abbasbandy, S. and Asady, B. *The nearest trapezoidal fuzzy number to a fuzzy quantity,* Appl. Math. Comput. 156 (2004) 381–386.

[5] Abbasbandy, S. and Hajjari, T. *Weighted trapezoidal approximation-preserving cores of a fuzzy number,* Comput. Math. Appl. 59(9) (2010) 3066–3077.

[6] Amirfakhrian, M. A New Parametric Distance and Using it to Find The Nearest Approximation of A Fuzzy Number, Conference: 11th Iranian Conference of Fuzzy Systems, 2011.

[7] M. Amirfakhrian and S. Bagherian, *The nearest approximation of a fuzzy number by a new parametric distance,* Conference: 40th Annual Iranian Mathematics Conference, 2009.

[8] Ban, A.I and Coroianu, L. *Nearest interval, triangular and trapezoidal approximation of a fuzzy number preserving ambiguity,* Internat. J. Approx. Reason. 53 (5) (2012) 805–836.

[9] Coroianu, L. *Trapezoidal approximations of fuzzy numbers using quadratic programs,*Fuzzy Sets Syst. 417 (2021) 71–92.

[10] Coroianu, L., Gagolewski, M. and Grzegorzewski, P. *Piecewise linear approximation of fuzzy numbers: algorithms, arithmetic operations and stability of characteristics,* Soft Comput. 23 (2019) 9491–9505.

[11] Dubois, D. and Prade, H. *Fuzzy number: an overview,* Fuzzy number: an overview,

[12] T. Hajjari, Approximation of fuzzy numbers via 0.5-Level and Mean core, Int. J. Appl. Math. Stat. (22) (2011) 69–78.

[13] T. Hajjari and H. Mahzoon, *Approximation of a fuzzy number by two main characteristics,* Mathematical Inverse Problems, 1(1) (2014) 101–111.

[14] Ghanbari, M., Allahviranloo, T., Nuraei, R. and Pedrycz, W. *A new effective approximate multiplication operation on LR fuzzy numbers and its application,* Soft Comput. 26(9)(2022) 4103–4113.

[15] Giachetti, R.E. and Young, R.E. *A parametric representation of fuzzy numbers and their arithmetic operators,* Fuzzy Sets Syst. 91 (2) (1997) 185–202.

[16] Guerra, M.L. and Stefanini, L.*Approximate fuzzy arithmetic operations using monotonic interpolations,* Fuzzy Sets Syst. 150 (2005) 5–33.

[17] Klir, G.J., St. Clair, U. and Yuan, B., *Fuzzy set theory: foundations and applications,* Prentice-Hall Inc. 1997.

[18] Ma, M., Kandel, A. and Friedman, M. *Correction to a new approach for the difizification,* "[Fuzzy Sets and Systems 111 (2000) 351-356]." Fuzzy Sets Syst. 128 (1) (2002) 133–134.

[19] Zheng, Q. and Zhang, L. *Global minimization of constrained problems with discontinuous penalty functions, Global optimization, control, and games, III,* Comput. Math. Appl. 37 (1999), no. 4-5, 41–58.

[20] Voxman, W. *Canonical representation of discrete fuzzy numbers,* Fuzzy Set Syst. 118 (3) (2001) 557–466.

[21] Yeh, C.-T. *Existence of interval, triangular, and trapezoidal approximations of fuzzy numbers under a general condition,* Fuzzy Sets Syst. 310 (2017) 1–13.

[22] Zadeh, L.A. *The concept of a linguistic variable and its application to approximate reasoning-I*, Inf. Sci. 8 (1975) 199–249.

[23] Zadeh L.A. *The concept of a linguistic variable and its application to approximate reasoning-II,* Inf. Sci. 8 (1975) , 301–357.

[24] Zadeh L.A. *The concept of a linguistic variable and its application to approximate reasoning-III*, Inf. Sci. 9 (1975) 43–80.

[25] Zhou, J., Yang, F. and Wang, K. *Fuzzy arithmetic on LR fuzzy numbers with applications to fuzzy programming,* J. Intell. Fuzzy Syst. 30 (2016) 71–87.

How to cite this article

A novel integral transform operator and its applications

Sh. Arora and A. Pasrija*

Abstract

The proposed study is focused to introduce a novel integral transform operator, called Generalized Bivariate (GB) transform. The proposed transform includes the features of the recently introduced Shehu transform, ARA transform, and Formable transform. It expands the repertoire of existing Laplace-type bivariate transforms. The primary focus of the present work is to elaborate fashionable properties and convolution theorems for the proposed transform operator. The existence, inversion, and duality of the proposed transform have been established with other existing transforms. Implementation of the proposed transform has been demonstrated by applying it to different types of differential and integral equations. It validates the potential and trustworthiness of the GB transform as a mathematical tool. Furthermore, weighted norm inequalities for integral convolutions have been constructed for the proposed transform operator.

---

*Corresponding author

Received 7 February 2023; revised 25 April 2023; accepted 2 May 2023

Shelly Arora
Department of Mathematics, Punjabi University, Patiala, Punjab-147002, India.
e-mail: aroshelly@pbi.ac.in

Atul Pasrija
Department of Mathematics, Punjabi University, Patiala, Punjab-147002, India.
e-mail: atulpasrijars@pbi.ac.in

1 Introduction

In real-life applications, the study of dynamic relations between individual components leads to different types of differential equations, integral equations, or integro-differential equations [21, 25, 36]. On account of extensive applications, these models crave for efficient techniques to construct their solutions. Integral transform techniques owing to the contribution of Heaveside to operational techniques have emerged as an alternative and bridge between analytic and numerical techniques in solving linear and nonlinear problems. Integral transform techniques are applicable over a wide class of problems, such as time-dependent boundary conditions, where the technique of separation of variables ceases to work. Even in a scenario of impossible analytic evaluation of transform or inverse transform, a wide variety of numerical and asymptotic techniques are now available for their evaluation [7, 10, 11, 19, 23, 28]. This hybrid mixture of techniques preserves some analytic aspects of the system that serves greater physical insight than a purely numerical procedure.

Integral transforms occur in a natural way by virtue of the principle of linear supposition in composing the integral form of the solution of linear differential equations. Integral transforms are one of the mathematical tools that have proved their worth not only for their theoretical interests but also for their accessible features to solve various problems in different fields of science and engineering. In recent work, the widely investigated subject of integral transforms has gained remarkable significance due to its demonstrated applications over quite challenging fractional operators [5, 15, 16, 30, 31]. The fundamental objective of integral transforms is to take one step forward to an easier form of the given problem. For example, an ordinary differential equation with constant coefficients transforms into an algebraic equation of transformed variable, and a partial differential equation (PDE) reduces to another PDE in one less variable. After the manipulation of the solution in the transformed domain, the inverse transform retracts the solution in the original domain. Different types of integral transforms are effectively utilized to obtain the solution of differential, difference, and integral equations. Indeed, the Fourier [9] and Laplace [29, 34] transforms are mostly applied and have been found to have a wide breadth of applications in mathematics, physics, statistics, and engineering sciences. Each of the existing integral transforms admits its strengths and deficiencies, which stimulates the interest to explore enhanced transforms with the arbitrariness of kernel function. Holding significance for centuries, the Fourier and Laplace transform even served as a generator for innumerable Laplace-type transforms with the imposition of specific conditions. The renowned Sumudu transform was introduced in the early 1990s by Watugala [33]. The natural transform [20] was devised in 2008. In 2011, Elzaki [12] framed a new integral transform known as Elzaki transform. In 2013, Atangana and Kilicman [6] established novel transforms for differential equations consisting of some kind of singularities. In recent

---

years, the substantial interest of researchers resulted in many worth mentioning integral transforms, namely, Ramadan Group transform [24], Polynomial transform [8], Yang transform [35], Aboodh transform [1], Mohand transform [2], Rangaing transform [13], Sawi transform [18], HY-transform [4], Shehu transform [22], J-transform [37], ARA transform [27], Formable transform [26], and so on.

In the present work, a new integral transform operator, "Generalized Bivariate transform", has been proclaimed as a generalization of the recently introduced Shehu, ARA, and Formable transforms. Its harmony in the class of Laplace-type transforms marks it as a prime member with inherited advantages of allied integral transforms. By proving fashionable properties along with application to various differential and integral equations, this work is further enhanced by the construction of the weighted norm inequalities for integral convolutions using the proposed transform operator.

2 Formulation of the GB transform

The GB transform of order n of a function $f(\eta)$ is a semi-infinite convergent integral. It can be defined as

$$\mathbb{A}_n[f(\eta)] = \mathcal{P}_n(s,\gamma) = \frac{s}{\gamma^n} \int_0^\infty \eta^{n-1} \exp\left(\frac{-s}{\gamma}\eta\right)f(\eta)\, d\eta, \quad \gamma, s > 0. \tag{1}$$

Equation (1) is equivalent to

$$\mathbb{A}_n[f(\eta)] = \mathcal{P}_n(s,\gamma) = s \int_0^\infty \eta^{n-1} \exp\left(-s\eta\right)f(\gamma\eta)\, d\eta, \quad \gamma, s > 0, \tag{2}$$

over the set of functions

$$\mathcal{F} = \left\{ f(\eta)\colon \text{ there exist } N \in (0,\infty), \eta_i > 0 \text{ for } i = 1, 2; |f(\eta)| < N \exp\left(\frac{|\eta|}{\eta_i}\right), \right.$$
$$\left. \text{if } \eta \in (-1)^i \times [0, \infty) \right\},$$

where s and γ are the variables of the GB transform.

In integral transforms theory, the recovery of a function from its transformed version is a more sophisticated subject than the evaluation of the transform itself, which is referred to as the inversion problem. For a given function, three questions arise at once: Does its inverse transform exist? is it unique? and how to find it? The uniqueness of Laplace-type transforms is determined by integration theory, which implies that a given function holds a unique continuous inverse transform. Moreover, in many cases, finding the inverse results in another transform with a different kernel function.

The inversion of the GB transform is given by

$$
\begin{aligned}
f(\eta) &= \mathbb{A}_{n+1}^{-1}\big[\mathbb{A}_{n+1}[f(\eta)]\big] \\
&= \frac{(-1)^n}{2\pi i} \int_{c-i\infty}^{c+i\infty} \frac{1}{\gamma} e^{\frac{s}{\gamma}\eta} \bigg[(-1)^n \bigg(\frac{1}{s\Gamma(n-1)} \int_0^s (s-x)^{n-1} \\
&\quad \mathbb{A}_{n+1}[f(\eta)](x,\gamma)\,dx + \sum_{k=0}^{n-1} \frac{s^k}{k!} \frac{\partial^k \$(0,\gamma)}{\partial s^k} \bigg) \bigg]\,ds,
\end{aligned}
\tag{3}
$$

where

$$
\$(s,\gamma) = \int_0^\infty e^{\frac{s}{\gamma}\eta} f(\eta)\,d\eta
$$

is the Shehu transform, which is $(n-1)$ times differentiable.

Proof. From the definition of the GB transform, we have

$$
\mathbb{A}_{n+1}[f(\eta)](s,\gamma) = \frac{s}{\gamma^n} \int_0^\infty \eta^n e^{\frac{s}{\gamma}\eta} f(\eta)\,d\eta = s(-1)^n \frac{\partial \$(s,\gamma)}{\partial s^n}.
$$

Thus

$$
\frac{1}{s\Gamma(n-1)} \int_0^s (s-\eta)^{n-1} \mathcal{P}_{n+1}(\eta,\gamma)\,d\eta = (-1)^n \bigg(\$(s,\gamma) - \sum_{k=0}^{n-1} \frac{s^k}{k!}\$(0,\gamma) \bigg).
$$

Therefore,

$$
\frac{(-1)^n}{s\Gamma(n-1)} \int_0^s (s-\eta)^{n-1} \mathcal{P}_{n+1}(\eta,\gamma)\,d\eta + \sum_{k=0}^{n-1} \frac{s^k}{k!}\$(0,\gamma) = \$(s,\gamma).
$$

It follows that

$$
\frac{(-1)^n}{2\pi i} \int_{c-i\infty}^{c+i\infty} \frac{1}{\gamma} e^{\frac{s}{\gamma}\eta} \bigg[(-1)^n \bigg(\frac{1}{s\Gamma(n-1)} \int_0^s (s-x)^{n-1} \mathcal{P}_{n+1}[f(\eta)](x,\gamma)\,dx + \sum_{k=0}^{n-1} \frac{s^k}{k!} \frac{\partial^k \$(0,\gamma)}{\partial s^k} \bigg) \bigg]\,ds
$$

$$
= \frac{(-1)^n}{2\pi i} \int_{c-i\infty}^{c+i\infty} \frac{1}{\gamma} e^{\frac{s}{\gamma}\eta} \Big[(-1)^n \$(s,\gamma) \Big]\,ds
$$

$$
= \frac{(-1)^{2n}}{2\pi i} \int_{c-i\infty}^{c+i\infty} \frac{1}{\gamma} e^{\frac{s}{\gamma}\eta} \$(s,\gamma) \Big]\,ds.
$$

Hence,

$$
\mathbb{A}_{n+1}^{-1}\big[\mathbb{A}_{n+1}[f(\eta)]\big] = \frac{(-1)^{2n}}{2\pi i} \int_{c-i\infty}^{c+i\infty} \frac{1}{\gamma} e^{\frac{s}{\gamma}\eta} \$(s,\gamma) \Big]\,ds = f(\eta).
$$

$\square$

Theorem 1 (Sufficient condition for the existence of the GB transform)**.** If the function $f(\eta)$ is piecewise continuous on every finite interval $0 < \eta < \xi$ and satisfies

$$|\eta^{n-1} f(\eta)| \le K e^{\beta \eta}, \tag{4}$$

then the GB transform exists for all $\frac{s}{\gamma} > \beta$.

Proof. Let ξ be any positive number. This will give

$$\frac{s}{\gamma^n} \int_0^\infty \eta^{n-1} e^{\frac{-s}{\gamma} \eta} f(\eta) d\eta = \frac{s}{\gamma^n} \int_0^\xi \eta^{n-1} e^{\frac{-s}{\gamma} \eta} f(\eta) d\eta + \frac{s}{\gamma^n} \int_\xi^\infty \eta^{n-1} e^{\frac{-s}{\gamma} \eta} f(\eta) d\eta.$$

Since the function is continuous on finite intervals, the first integral on the right-hand side exists. Also, by the hypothesis in (4), the latter integral on the right-hand side converges

$$\left| \frac{s}{\gamma^n} \int_\xi^\infty \eta^{n-1} e^{\frac{s}{\gamma} \eta} f(\eta) \, \eta \right| \le \frac{s}{\gamma^n} \int_\xi^\infty e^{\frac{-s}{\gamma} \eta} K e^{\beta \eta} \, d\eta$$

$$= \lim_{\alpha \to \infty} - \frac{sK}{\gamma^n} \frac{e^{-\xi(\frac{s}{\gamma} - \beta)}}{(\frac{s}{\gamma} - \beta)} \bigg|_0^\alpha = \frac{sK}{\gamma^{n-1}(\frac{s}{\gamma} - \beta)}.$$

Thus the GB transform $\mathbb{A}_n[f(\eta)]$ exists for all $\frac{s}{\gamma} > \beta$. □

There are many functions for which most of variants of the Laplace transform do not exist. The GB transform expands the repertoire of Laplace-type transforms by its applicability over the following functions:

$$\mathbb{A}_2 \left[\frac{1}{\eta} \right] = \frac{s}{\gamma^2} \int_0^\infty \eta^{2-1} e^{\frac{-s}{\gamma} \eta} \frac{1}{\eta} \, d\eta = \frac{1}{\gamma},$$

$$\mathbb{A}_2 \left[2 e^{\eta^2} \cos e^{\eta^2} \right]$$

$$= \frac{s}{\gamma^2} \int_0^\infty 2 \eta e^{\frac{-s}{\gamma} \eta} e^{\eta^2} \cos (e^{\eta^2}) \, d\eta$$

$$= \frac{s}{\gamma^2} \left[e^{\frac{-s}{\gamma} \eta} \sin (e^{\eta^2}) \bigg|_0^\infty + \frac{s}{\gamma} \int_0^\infty e^{\frac{-s}{\gamma} \eta} \sin (e^{\eta^2}) \, d\eta \right] \quad \text{(Integration by parts)}$$

$$= \frac{s}{\gamma^2} \left[- \sin(1) + \mathbb{A}_1[\sin (e^{\eta^2})] \right]. \tag{5}$$

and the latter GB transform exists by Theorem 1. Similarly, $\mathbb{A}_2 \left[2 e^{\eta^2} \sin e^{\eta^2} \right]$ can be obtained as per (5).

Theorem 2 (Uniqueness of the GB transform)**.** Let $f(\eta)$ and $g(\eta)$ be the continuous functions, defined for $\eta \ge 0$ and having the GB transform for

order n, $\mathcal{P}_n(s,\gamma)$, and $\mathcal{Q}_n(s,\gamma)$, respectively. If $\mathcal{P}_n(s,\gamma) = \mathcal{Q}_n(s,\gamma)$, then $f(\eta) = g(\eta)$.

Proof. From the definition of the GB transform of order n, if c is sufficiently large, then the integral expression can be obtained as

$$f(\eta) = \frac{(-1)^n}{2\pi i} \int_{c-i\infty}^{c+i\infty} \frac{1}{u} e^{\frac{s}{\gamma}\eta} [(-1)^n \mathcal{P}_n(s,\gamma)] \, ds.$$

By hypothesis, $\mathcal{P}_n(s,\gamma) = \mathcal{Q}_n(s,\gamma)$ and

$$f(\eta) = \frac{(-1)^n}{2\pi i} \int_{c-i\infty}^{c+i\infty} \frac{1}{u} e^{\frac{s}{\gamma}\eta} [(-1)^n \mathcal{Q}_n(s,\gamma)] \, ds = g(\eta).$$

$\square$

3 Dualities between the GB transform and some integral transforms

In this section, the associative nature of the GB transform with other well-known transforms is illustrated. This association of the GB transform enhances to exploration of other transforms simultaneously under the study of the GB transform.

- GB-ARA duality:

$$\mathcal{P}_n(s,\gamma) = \frac{1}{\gamma^{n-1}} G\left(n, \frac{s}{\gamma}\right), \tag{6}$$
$$\mathcal{P}_n(s,1) = G(n,s),$$

    where [27]

$$\mathcal{G}_n[f(\eta)](s) = G(n,s) = s \int_0^\infty \eta^{n-1} e^{-s\eta} f(\eta) \, d\eta.$$

- GB-Formable duality:

$$\mathcal{P}_n(s,\gamma) = \frac{1}{\gamma^{n-1}} \mathcal{R}[\eta^{n-1} f(\eta)], \tag{7}$$
$$\mathcal{P}_1(s,\gamma) = B(s,\gamma),$$

    where [26]
$$\mathcal{R}[f(\eta)] = B(s,\gamma) = \frac{s}{\gamma} \int_0^\infty e^{-\frac{s}{\gamma}\eta} f(\eta) \, d\eta.$$

- GB-Shehu duality:

$$\mathcal{P}_n(s,\gamma) = \frac{s}{\gamma^n}\mathbb{S}[\eta^{n-1}f(\eta)], \tag{8}$$

$$\mathcal{P}_1(s,\gamma) = \frac{s}{\gamma}V(s,\gamma),$$

where [22]

$$\mathbb{S}[f(\eta)] = V(s,\gamma) = \int_0^\infty e^{-\frac{s}{\gamma}\eta}f(\eta)\,d\eta.$$

- GB-Natural duality:

$$\mathcal{P}_n(s,\gamma) = \frac{s}{\gamma^{n-1}}\mathcal{N}[\eta^{n-1}f(\eta)], \tag{9}$$

$$\mathcal{P}_1(s,\gamma) = sR(s,\gamma),$$

where [20]

$$\mathcal{N}[f(\eta)] = R(s,\gamma) = \frac{1}{\gamma}\int_0^\infty e^{-\frac{s}{\gamma}\eta}f(\eta)\,d\eta.$$

- GB-Aboodh duality:

$$\mathcal{P}_n(s,\gamma) = \frac{s^2}{\gamma^{n+1}}\mathcal{A}[\eta^{n-1}f(\eta)]\left(\frac{s}{\gamma}\right), \tag{10}$$

$$\mathcal{P}_1(s,1) = s^2\mathcal{A}[f(\eta)](s),$$

where [1]

$$\mathcal{A}[f(\eta)](s) = \frac{1}{s}\int_0^\infty e^{-s\eta}f(\eta)\,d\eta.$$

- GB-J duality:

$$\mathcal{P}_n(s,\gamma) = \frac{s}{\gamma^{n+1}}\mathcal{J}[\eta^{n-1}f(\eta)], \tag{11}$$

$$\mathcal{P}_1(s,\gamma) = \frac{s}{\gamma^2}J(s,\gamma),$$

where [37]

$$\mathcal{J}[f(\eta)] = J(s,\gamma) = u\int_0^\infty e^{\frac{-s}{\gamma}\eta}f(\eta)\,d\eta.$$

- GB-Laplace Carson duality:

$$\mathcal{P}_n(s,\gamma) = \frac{1}{\gamma^{n-1}}\mathcal{L}_*[\eta^{n-1}f(\eta)]\left(\frac{s}{\gamma}\right), \tag{12}$$

$$\mathcal{P}_1(s,1) = \mathcal{L}_*[f(\eta)](s),$$

where [3]

$$\mathcal{L}_*[f(\eta)](s) = s \int_0^\infty e^{-s\eta} f(\eta) \, d\eta.$$

- GB-Elzaki duality:

$$\mathcal{P}_n(s,\gamma) = \frac{s^2}{\gamma^{n+1}} \mathcal{E}[\eta^{n-1} f(\eta)] \left(\frac{\gamma}{s}\right),\tag{13}$$

$$\mathcal{P}_1(1,\gamma) = \frac{1}{\gamma^2} \mathcal{E}[f(\eta)](\gamma),$$

where [12]

$$\mathcal{E}[f(\eta)](\gamma) = \gamma \int_0^\infty e^{\frac{-\eta}{\gamma}} f(\eta) \, d\eta.$$

- GB-Sumudu duality:

$$\mathcal{P}_n(s,\gamma) = \frac{1}{\gamma^{n-1}} \mathcal{S}[\eta^{n-1} f(\eta)] \left(\frac{\gamma}{s}\right),\tag{14}$$

$$\mathcal{P}_1(1,\gamma) = \mathcal{S}[f(\eta)](\gamma),$$

where [33]

$$\mathcal{S}[f(\eta)](\gamma) = \frac{1}{\gamma} \int_0^\infty e^{\frac{-\eta}{\gamma}} f(\eta) \, d\eta.$$

4 Properties of the GB transform

In this section, some basic properties such as the linearity property, the shifting in domains, the derivative property, and the convolution property are presented, which enable us to determine the GB transform in applications.

Property 1 (Linearity property)**.** Suppose that $f(\eta)$ and $g(\eta)$ are two functions for which the GB transform exists. Then

$$\mathbb{A}_n \left[\alpha f(\eta) + \beta g(\eta)\right] = \alpha \mathbb{A}_n \left[f(\eta)\right] + \beta \mathbb{A}_n \left[g(\eta)\right],\tag{15}$$

where α and β are nonzero arbitrary constants.

Property 2 (Change of scale)**.** Suppose that the GB transform exists for the given function $f(\alpha\eta)$. Then

$$\mathbb{A}_n \left[f(\alpha\eta)\right] = \mathcal{P}_n(s,\alpha\gamma),\tag{16}$$

where α is an arbitrary constant.

Property 3 (Shifting in s-domain)**.**

$$\mathbb{A}_n[e^{-\alpha\eta} f(\eta)] = \frac{s}{s+\gamma\alpha} \mathcal{P}_n(s+\gamma\alpha, \gamma),\tag{17}$$

where α is an arbitrary constant.

Property 4 (Shifting in η-domain)**.**

$$\mathbb{A}_n[u_\alpha f(\eta - \alpha)] = \frac{e^{\frac{-s}{\gamma}\alpha}}{\gamma^{n-1}}\mathbb{A}_1[(\eta + \alpha)^{n-1}f(\eta)], \qquad (18)$$

where u_α is the unit function and α is an arbitrary constant.

Property 5 (Shifting in n-domain)**.**

$$\mathbb{A}_n[\eta^m f(\eta)] = \gamma^m \, \mathbb{A}_{n+m}[f(\eta)], \qquad (19)$$

where $m \geq 0$ or $n - 1 \geq m$.

The proofs of Properties 1–5 can be easily proved by usual calculus.

Property 6 (GB transform for derivatives)**.** Suppose that $f(\eta), f'(\eta), \ldots,$ $f^{m-1}(\eta)$ are continuous and of exponential order on $[0, \infty)$ while $f^m(\eta)$ is piecewise continuous on $[0, \infty)$. Then

$$\mathbb{A}_n[f^m(\eta)] = (-1)^{n-1}\frac{s}{\gamma}\frac{\partial^{n-1}}{\partial s^{n-1}}\left[\left(\frac{s}{\gamma}\right)^{m-1}\mathbb{A}_1[f(\eta)] - \sum_{k=0}^{m-1}\left(\frac{s}{\gamma}\right)^{m-(k+1)}f^k(0)\right].$$
$$(20)$$

Proof. We have

$$\mathbb{A}_n[f^m(\eta)]$$
$$= \frac{s}{\gamma^n}\int_0^\infty \eta^{n-1}e^{\frac{-s}{\gamma}\eta}f^m(\eta)\, d\eta$$
$$= \frac{1}{\gamma^{n-1}}\mathbb{A}_1[\eta^{n-1}f^m(\eta)]$$
$$= \frac{1}{\gamma^{n-1}}\mathcal{R}[\eta^{n-1}f^m(\eta)] \quad \text{(Duality between GB and Formable transforms)}$$
$$= (-1)^{n-1}s\frac{\partial^{n-1}}{\partial s^{n-1}}\left[\frac{\mathcal{R}[f^m(\eta)]}{s}\right]$$
$$= (-1)^{n-1}s\frac{\partial^{n-1}}{\partial s^{n-1}}\left[\frac{s^{m-1}}{\gamma^{m-1}}\mathcal{R}[f(\eta)] - \sum_{k=0}^{m-1}\frac{s^{m-(k+1)}}{\gamma^{m-(k+1)}}f^k(0)\right]$$
$$= (-1)^{n-1}\frac{s}{\gamma}\frac{\partial^{n-1}}{\partial s^{n-1}}\left[\left(\frac{s}{\gamma}\right)^{m-1}\mathbb{A}_1[f(\eta)] - \sum_{k=0}^{m-1}\frac{s^{m-(k+1)}}{\gamma^{m-(k+1)}}f^k(0)\right].$$
$$\text{(Using the GB-Formable duality)}$$

$$\square$$

Property 7 (GB transform of the convolution). Suppose that $\mathcal{P}_1(s,\gamma)$ and $\mathcal{Q}_1(s,\gamma)$ are the GB transform of order one of the functions $f(\eta)$ and $g(\eta)$, respectively. Then

$$\mathbb{A}_n[f(\eta)*g(\eta)] = (-1)^{n-1}s\gamma\sum_{r=0}^{n-1}C_r^{n-1}\frac{\partial^{n-1-r}}{\partial s^{n-1-r}}\left(\frac{1}{s^2}\right)$$

$$\sum_{k=0}^{r}C_k^r\frac{\partial^{r-k}}{\partial s^{r-k}}\mathcal{P}_1(s,\gamma).\frac{\partial^k}{\partial s^k}\mathcal{Q}_1(s,\gamma), \qquad (21)$$

where the convolution (i.e., $f(\eta)*g(\eta)$) of the functions $f(\eta)$ and $g(\eta)$ is given by the integral

$$f(\eta)*g(\eta) = \int_0^{\eta}f(\eta)g(\eta-\zeta)\,d\zeta.$$

Proof. We have

$\mathbb{A}_n[f(\eta)*g(\eta)]$

$$= \frac{1}{\gamma^{n-1}}\mathbb{A}_1[\eta^{n-1}(f(\eta)*g(\eta))]$$

$$= (-1)^{n-1}s\gamma\frac{\partial^{n-1}}{\partial s^{n-1}}\left[\frac{\mathcal{P}_1(s,\gamma).\mathcal{Q}_1(s,\gamma)}{s^2}\right] \quad \text{(Using the GB-Formable duality)}$$

$$= (-1)^{n-1}s\gamma\sum_{r=0}^{n-1}C_r^{n-1}\frac{\partial^{n-1-r}}{\partial s^{n-1-r}}\left(\frac{1}{s^2}\right)\frac{\partial^r}{\partial s^r}(\mathcal{P}_1(s,\gamma).\mathcal{Q}_1(s,\gamma))$$

$$= (-1)^{n-1}s\gamma\sum_{r=0}^{n-1}C_r^{n-1}\frac{\partial^{n-1-r}}{\partial s^{n-1-r}}\left(\frac{1}{s^2}\right)\sum_{k=0}^{r}C_k^r\frac{\partial^{r-k}}{\partial s^{r-k}}\mathcal{P}_1(s,\gamma).\frac{\partial^k}{\partial s^k}\mathcal{Q}_1(s,\gamma)$$

where C_i^j is the binomial coefficient. $\qquad\square$

Now, computational simplicity of the GB transform is presented by its evaluation for some elementary functions.

Example 1. Consider

$$\mathbb{A}_n[\eta^m e^{\alpha\eta}] = (m+n-1)!\frac{s\gamma^m}{(s-\alpha\gamma)^{m+n}}, \qquad (22)$$

where $m \geq 0$ or $n-1 \geq m$ and α is an arbitrary constant.

Proof. We have

$$\mathbb{A}_{n+1}[\eta^m e^{\alpha\eta}] = \frac{s}{\gamma^{n+1}}\int_0^{\infty}\eta^{m+n}e^{-\left(\frac{s-\alpha\gamma}{\gamma}\right)\eta}\,d\eta$$

$$= \frac{(m+n)}{\gamma^n}\frac{s}{s-\alpha\gamma}\int_0^{\infty}\eta^{m+n-1}e^{-\left(\frac{s-\alpha\gamma}{\gamma}\right)\eta}\,d\eta$$

$$= \frac{(m+n)(m+n-1)}{\gamma^n} \frac{s\gamma}{(s-\alpha\gamma)^2} \int_0^\infty \eta^{m+n-2} e^{-\left(\frac{s-\alpha\gamma}{\gamma}\right)\eta} \, d\eta$$

$$\vdots$$

$$= \frac{(m+n)!}{\gamma^n} \frac{s\gamma^{m+n}}{(s-\alpha\gamma)^{m+n+1}} = (m+n)! \frac{s\gamma^m}{(s-\alpha\gamma)^{m+n+1}}.$$

$\square$

Example 2. Consider

$$\mathbb{A}_n[\sin\alpha\eta] = \frac{s(n-1)!}{2i} \left[\frac{(s+i\alpha\gamma)^n - (s-i\alpha\gamma)^n}{(s^2+\alpha^2\gamma^2)^n} \right], \tag{23}$$

$$\mathbb{A}_n[\cos\alpha\eta] = \frac{s(n-1)!}{2} \left[\frac{(s+i\alpha\gamma)^n + (s-i\alpha\gamma)^n}{(s^2+\alpha^2\gamma^2)^n} \right], \tag{24}$$

$$\mathbb{A}_n[\sinh\alpha\eta] = \frac{s(n-1)!}{2} \left[\frac{(s+\alpha\gamma)^n - (s-\alpha\gamma)^n}{(s^2-\alpha^2\gamma^2)^n} \right], \tag{25}$$

$$\mathbb{A}_n[\cosh\alpha\eta] = \frac{s(n-1)!}{2} \left[\frac{(s+\alpha\gamma)^n + (s-\alpha\gamma)^n}{(s^2-\alpha^2\gamma^2)^n} \right], \tag{26}$$

where α is an arbitrary constant.

Now, using the linearity property given in 1 and basic calculus, above results can be proved easily.

Further applications of the GB transform over some elementary and special functions are given in Table 1.

5 Weighted norm inequalities for integral convolution of the GB transform

Theorem 3. Suppose that $f(\eta)$ and $g(\eta)$ are complex valued continuous functions on $[0,\infty)$ such that the GB transforms

$$\mathcal{P}_\alpha(s,\gamma) = \mathbb{A}_\alpha \left[\frac{f(\eta)}{\Gamma\alpha} \right] \quad and \quad \mathcal{Q}_\beta(s,\gamma) = \mathbb{A}_\beta \left[\frac{g(\eta)}{\Gamma\beta} \right] \tag{27}$$

exist for some $\alpha, \beta > 0$ and $\frac{s}{\gamma} > s_0 \geq 0$.

Then, for any arbitrary $\lambda > 0, p > 1 \left(\frac{1}{p} + \frac{1}{q} = 1 \right), \delta \in [0, min(p,q)]$, and $\xi \in [0,\infty)$, the following inequality holds [14]:

Table 1: GB transform of some special functions

$f(\eta)$	GB transform
1	$\Gamma(n)s^{1-n}$
η	$\Gamma(n+1)\frac{\gamma}{s^n}$
η^m	$\Gamma(m+n)\frac{\gamma^m}{s^{n+m-1}}$
$e^{\alpha\eta}$	$\Gamma(n)\frac{s}{(s-\gamma\alpha)^n}$
$e^{-\alpha\eta}$	$\Gamma(n)\frac{s}{(s+\gamma\alpha)^n}$
$\eta^m e^{\alpha\eta}$	$\Gamma(m+n)\frac{s\gamma^m}{(s-\gamma\alpha)^{m+n}}$
$\sin(\alpha\eta)$	$\left(1+\frac{(\alpha\gamma)^2}{s^2}\right)^{\frac{n}{2}}s^{1-n}\Gamma(n)\sin\left(n\tan^{-1}\left(\frac{\alpha\gamma}{s}\right)\right)$
$\cos(\alpha\eta)$	$\left(1+\frac{(\alpha\gamma)^2}{s^2}\right)^{\frac{n}{2}}s^{1-n}\Gamma(n)\cos\left(n\tan^{-1}\left(\frac{\alpha\gamma}{s}\right)\right)$
$\eta\sin(\alpha\eta)$	$\Gamma(n+1)\frac{\gamma s}{(s^2+\alpha^2\gamma^2)^{\frac{1}{2}(n+1)}}\sin\left((1+n)\tan^{-1}\left(\frac{\alpha\gamma}{s}\right)\right)$
$\eta\cos(\alpha\eta)$	$\Gamma(n+1)\frac{\gamma s}{(s^2+\alpha^2\gamma^2)^{\frac{1}{2}(n+1)}}\cos\left((1+n)\tan^{-1}\left(\frac{\alpha\gamma}{s}\right)\right)$
$\sin(\alpha\eta)-\alpha\eta\cos(\alpha\eta)$	$\Gamma(n)\frac{\alpha s}{(s^2+\alpha\gamma)^{\frac{1}{2}(n+1)}}\left[-\alpha n\cos\left((1+n)\tan^{-1}\left(\frac{\alpha\gamma}{s}\right)\right)+\frac{1}{\gamma}\sqrt{s^2+\alpha^2\gamma^2}\sin\left((1+n)\tan^{-1}\left(\frac{\alpha\gamma}{s}\right)\right)\right]$
$\sin(\alpha\eta)+\alpha\eta\cos(\alpha\eta)$	$\Gamma(n)\frac{\alpha s}{(s^2+\alpha\gamma)^{\frac{1}{2}(n+1)}}\left[\alpha n\cos\left((1+n)\tan^{-1}\left(\frac{\alpha\gamma}{s}\right)\right)+\frac{1}{\gamma}\sqrt{s^2+\alpha^2\gamma^2}\sin\left((1+n)\tan^{-1}\left(\frac{\alpha\gamma}{s}\right)\right)\right]$
$\cos(\alpha\eta)-\alpha\eta\sin(\alpha\eta)$	$\Gamma(n)\frac{\alpha s}{(s^2+\alpha\gamma)^{\frac{1}{2}(n+1)}}\left[\frac{s}{\gamma}\cos\left((1+n)\tan^{-1}\left(\frac{\alpha\gamma}{s}\right)\right)+\alpha(n-1)\sin\left((1+n)\tan^{-1}\left(\frac{\alpha\gamma}{s}\right)\right)\right]$
$\cos(\alpha\eta)+\alpha\eta\sin(\alpha\eta)$	$\Gamma(n)\frac{\alpha s}{(s^2+\alpha\gamma)^{\frac{1}{2}(n+1)}}\left[\frac{s}{\gamma}\cos\left((1+n)\tan^{-1}\left(\frac{\alpha\gamma}{s}\right)\right)+\alpha(n+1)\sin\left((1+n)\tan^{-1}\left(\frac{\alpha\gamma}{s}\right)\right)\right]$
$\sin(\alpha\eta+\beta)$	$\frac{s}{(s^2+\alpha^2\gamma^2)^{\frac{n}{2}}}\Gamma(n)\sin\left(\beta+n\tan^{-1}\left(\frac{\alpha\gamma}{s}\right)\right)$
$\cos(\alpha\eta+\beta)$	$\frac{s}{(s^2+\alpha^2\gamma^2)^{\frac{n}{2}}}\Gamma(n)\cos\left(\beta+n\tan^{-1}\left(\frac{\alpha\gamma}{s}\right)\right)$
$e^{\alpha\eta}\sin(\beta\eta)$	$\Gamma(n)\frac{n}{(s-\alpha\gamma)^n}\left(1+\frac{\beta^2\gamma^2}{(\alpha\gamma-s)^2}\right)^{\frac{-n}{2}}\sin\left(n\tan^{-1}\left(\frac{\beta\gamma}{s-\alpha\gamma}\right)\right)$
$e^{\alpha\eta}\cos(\beta\eta)$	$\Gamma(n)\frac{n}{(s-\alpha\gamma)^n}\left(1+\frac{\beta^2\gamma^2}{(\alpha\gamma-s)^2}\right)^{\frac{-n}{2}}\cos\left(n\tan^{-1}\left(\frac{\beta\gamma}{s-\alpha\gamma}\right)\right)$
$\sinh(\alpha\eta)$	$\frac{1}{2}\frac{s\Gamma(n)}{(s^2-\alpha^2\gamma^2)^n}\left[-(s-\gamma\alpha)^n+(s+\gamma\alpha)^n\right]$
$\cosh(\alpha\eta)$	$\frac{1}{2}\frac{s\Gamma(n)}{(s^2-\alpha^2\gamma^2)^n}\left[(s-\gamma|\alpha|)^n+(s+\gamma|\alpha|)^n\right]$
$e^{\alpha\eta}\sinh(\beta\eta)$	$\Gamma(n)\frac{n}{(s-\alpha\gamma)^n}\left(1-\frac{\beta^2\gamma^2}{(\alpha\gamma-s)^2}\right)^{-n}\left[-\left(1+\frac{\beta\gamma}{\alpha\gamma-s}\right)^n+\left(1-\frac{\beta\gamma}{\alpha\gamma-s}\right)^n\right]$
$e^{\alpha\eta}\cosh(\beta\eta)$	$\Gamma(n)\frac{n}{(s-\alpha\gamma)^n}\left[\left(1+\frac{\beta\gamma}{\alpha\gamma-s}\right)^{-n}+\left(1-\frac{\beta\gamma}{\alpha\gamma-s}\right)^{-n}\right]$
$\delta(\eta-\alpha)$	$\alpha^{n-1}\frac{s}{\gamma^n}\exp\left(\alpha\frac{s}{\gamma}\right)$
$J_0(\alpha\eta)$	$(-1)^{n-1}s\frac{\partial^{n-1}}{\partial s^{n-1}}\left[\frac{1}{\sqrt{s^2-\alpha^2\gamma^2}}\right]$
$\mathcal{U}(\eta-\alpha)$	$\sum_{k=0}^{n-1}\frac{1}{s^k}\left(\frac{\alpha}{\gamma}\right)^{n-k-1}\exp\left(-\frac{\alpha}{\gamma}s\right)$

$$\left|\left|h(\xi\eta)\right|\right|_{[\delta;\,\alpha+\beta,\,\lambda]}\leq\left|\left|f(\xi\eta)\right|\right|_{[p;\,\alpha,\,\beta+\lambda]}\left|\left|g(\xi\eta)\right|\right|_{[q;\,\beta,\,\alpha+\lambda]},\qquad(28)$$

where

$$||(\cdot)||_{[p,\,\alpha,\,\beta]}=\left[\int_0^1\eta^{\alpha-1}(1-\eta)^{\beta-1}|(\cdot)|^p\frac{d\eta}{B(\alpha,\beta)}\right]^{\frac{1}{p}}\qquad(29)$$

and $h(\eta)$, $0\leq\eta<\infty$, is a continuous solution to the integral equation:

$$\mathbb{A}_{\alpha+\beta}\left[\frac{h(\eta)}{\Gamma(\alpha+\beta)}\right]=\frac{1}{s}\,\mathcal{P}_\alpha(s,\gamma)\,\mathcal{Q}_\beta(s,\gamma).\qquad(30)$$

Equivalently, $h(\eta)$ is such function that $\eta^{\alpha+\beta-1}h(\eta)/\Gamma(\alpha+\beta)$ is the convolution of the functions $\eta^{\alpha-1}f(\eta)/\Gamma\alpha$ and $\eta^{\beta-1}g(\eta)/\Gamma\beta$.

Proof. Consider a continuous function $h(\eta)$, such as

$$h(\eta) = \langle f(\eta\delta)g(\eta(1-\delta))\rangle_{(\alpha,\,\beta)}, \qquad \eta \in [0,\infty]. \tag{31}$$

Indeed, the left-hand side of (30) with this function $h(\eta)$ and $\frac{s}{\gamma} > s_0$ yields

$$\frac{1}{\Gamma(\alpha)\Gamma(\beta)}\frac{s}{\gamma^{\alpha+\beta}}\int_0^\infty \eta^{\alpha+\beta-1}e^{\frac{s}{\gamma}\eta}\int_0^1 \delta^{\alpha-1}(1-\delta)^{\beta-1}f(\eta\delta)g(\eta(1-\delta))\,d\delta\,d\eta. \tag{32}$$

Substituting $\nu = \eta\delta \in [0,\eta]$, the change of the order of integration assembles (32) as $\qquad\qquad\square$

$$\frac{1}{\Gamma(\alpha)\Gamma(\beta)}\frac{s}{\gamma^{\alpha+\beta}}\int_0^\infty \nu^{\alpha-1}f(\nu)\int_\nu^\infty e^{\frac{s}{\gamma}\eta}(\eta-\nu)^{\beta-1}g(\eta-\nu)\,d\eta\,d\nu. \tag{33}$$

Another substitution $\mu = \eta - \nu \in [0,\infty)$ reformulates (33) as

$$\frac{1}{\Gamma(\alpha)\Gamma(\beta)}\frac{s}{\gamma^{\alpha+\beta}}\int_0^\infty \nu^{\alpha-1}f(\nu)\int_\nu^\infty e^{-(\mu+\nu)\frac{s}{\gamma}}(\mu)^{\beta-1}g(\mu)\,d\mu\,d\nu$$
$$= \frac{1}{s}\mathcal{P}_\alpha(s,\gamma)\,\mathcal{Q}_\beta(s,\gamma), \tag{34}$$

where $\mathcal{P}_\alpha(s,\gamma)$ and $\mathcal{Q}_\beta(s,\gamma)$ are defined in (27).

6 Applications of the GB transform

In this section, the application of the GB transform is demonstrated for the purpose to solve various Lane–Emden type differential equations, wave-like partial differential equations, and convolution-type integral equations. The success of the newly proposed transform with simplified computation suggests its further implementation to physical problems in sciences and engineering.

Problem 1. Consider the linear Lane–Emden differential equation as

$$u''(\eta) + \frac{2}{\eta}u'(\eta) + u(\eta) = 0, \tag{35}$$

subject to initial conditions

$$u(0) = 1, \quad u'(0) = 0. \tag{36}$$

Solution 1. Application of $\mathbb{A}_2$ on the both sides of (35) implies

$$-\frac{s^2}{\gamma^2}\mathbb{A}'_1[u(\eta)] - \frac{s}{\gamma^2}\mathbb{A}'_1[u(\eta)] + \frac{s}{\gamma^2}u(0) + \frac{1}{\gamma}\left[\frac{s}{\gamma}\mathbb{A}_1[u(\eta)] - \frac{s}{\gamma}u(0)\right]$$
$$+ \frac{\mathbb{A}_1[u(\eta)]}{s} - \mathbb{A}'_1[u(\eta)] = 0.$$

The use of given initial conditions and reordering of terms give

$$\mathbb{A}_1'[u(\eta)] + \frac{\mathbb{A}_1[u(\eta)]}{s} = -\frac{s}{(s^2+\gamma^2)},$$

$$-\gamma s \frac{d}{ds}\left[\frac{\mathbb{A}_1[u(\eta)]}{s}\right] = \frac{s\gamma}{(s^2+\gamma^2)},$$

$$\mathbb{A}_1[\eta u(\eta)] = \frac{s\gamma}{(s^2+\gamma^2)}.$$

The utilization of the inverse GB transform yields

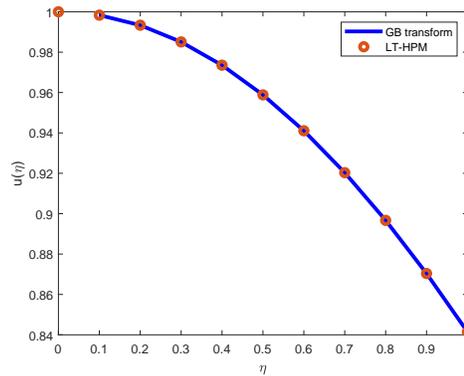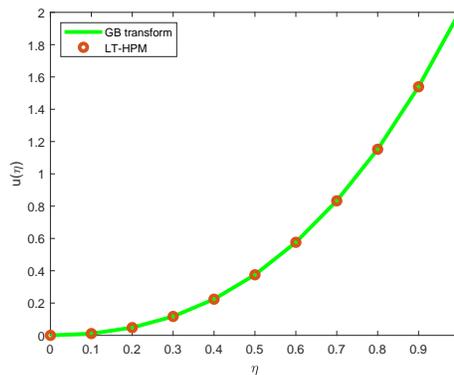$$u(\eta) = \frac{\sin\eta}{\eta},$$

which is an exact solution.

Figure 1: Comparison between the GB transform and LT-HPM solution

In Figure 1, the comparison of solution profiles for problem 1 reveals well agreement of the solution obtained by the GB transform with a series solution of LT-HPM [32].

Problem 2. Consider the linear, nonhomogeneous Lane–Emden differential equation

$$u''(\eta) + \frac{2}{\eta}u'(\eta) + u(\eta) = 6 + 12\eta + \eta^2 + \eta^3 \tag{37}$$

with initial conditions

$$u(0) = 0, \qquad u'(0) = 0. \tag{38}$$

Solution 2. Application of $\mathbb{A}_2$ on the both sides of (37) implies

$$-\frac{s^2}{\gamma^2}\mathbb{A}'_1[u(\eta)] - \frac{s}{\gamma^2}\mathbb{A}'_1[u(\eta)] + \frac{s}{\gamma^2}u(0) + \frac{1}{\gamma}\left[\frac{s}{\gamma}\mathbb{A}_1[u(\eta)] - \frac{s}{\gamma}u(0)\right]$$

$$+ \frac{\mathbb{A}_1[u(\eta)]}{s} - \mathbb{A}'_1[u(\eta)] = \frac{6}{s} + \frac{244}{s^2} + 6\frac{\gamma^2}{s^3} + 24\frac{\gamma^3}{s^4}.$$

Use of given initial conditions and reordering of terms give

$$\mathbb{A}'_1[u(\eta)] + \frac{\mathbb{A}_1[u(\eta)]}{s} = -\frac{\gamma^2}{(s^2 + \gamma^2)}\left[\frac{6}{s} + 24\frac{\gamma}{s^2} + 6\frac{\gamma^2}{s^3} + 24\frac{\gamma^3}{s^4}\right],$$

$$-\gamma s \frac{d}{ds}\left[\frac{\mathbb{A}_1[u(\eta)]}{s}\right] = \frac{\gamma^3}{(s^2 + \gamma^2)}\left[\frac{6}{s} + 24\frac{\gamma}{s^2} + 6\frac{\gamma^2}{s^3} + 24\frac{\gamma^3}{s^4}\right],$$

$$\mathbb{A}_1[\eta u(\eta)] = 6\frac{\gamma^3}{s^3} + 24\frac{\gamma^4}{s^4}.$$

Utilization of the inverse GB transform yields

$$u(\eta) = \eta^2 + \eta^3,$$

which is an exact solution.

Figure 2: Comparison between the GB transform and LT-HPM solution

In Figure 2, the graphical comparison of the solution for Problem 2 reflects the good agreement of the solution obtained by the GB transform and the series solution of LT-HPM [32].

In Problems 1 and 2, the GB transform has been found to be independently efficient in constructing the exact solution of linear Lane–Emden type equations, whereas Laplace transform demands modification of governing equations or other collaborative techniques to drive exact or approximate solutions. The same fact can be stated for the Shehu transform and Formable transform in comparison with the GB transform.

Problem 3. Consider the Bessel differential equation with polynomial coefficients as

$$u''(\eta) + \frac{1}{\eta} u'(\eta) + u(\eta) = 0 \tag{39}$$

with initial data

$$u(0) = 1, \quad u'(0) = 1. \tag{40}$$

Solution 3. Application of $\mathbb{A}_2$ on the both sides of (39) implies

$$-\frac{s^2}{\gamma^2} \mathbb{A}'_1[u(\eta)] - \frac{s}{\gamma^2} \mathbb{A}'_1[u(\eta)] + \frac{s}{\gamma^2} u(0)$$

$$+ \frac{1}{\gamma} \left[\frac{s}{\gamma} \mathbb{A}_1[u(\eta)] - \frac{s}{\gamma} u(0) \right] + \frac{\mathbb{A}_1[u(\eta)]}{s} - \mathbb{A}'_1[u(\eta)] = 0.$$

Use of (40) and reordering of terms give

$$\frac{(-s^2 - \gamma^2)}{\gamma^2} \mathbb{A}'_1[u(\eta)] + \frac{\mathbb{A}_1[u(\eta)]}{s} = 0.$$

The solution of the above equation yields

$$\mathbb{A}_1[u(\eta)] = \frac{\alpha s}{\sqrt{s^2 + \gamma^2}}.$$

Utilization of the inverse GB transform yields

$$u(\eta) = \alpha J_0(\eta).$$

Use of the initial data provides

$$u(\eta) = J_0(\eta),$$

where J_0 is the Bessel function.

Problem 4. Consider the wave-like partial differential equation

$$\frac{\partial^2 u(\xi, \eta)}{\partial \eta^2} = \frac{\partial^2 u(\xi, \eta)}{\partial \xi^2} + u(\xi, \eta) \tag{41}$$

with boundary conditions

$$u(0, \eta) = \cosh(\eta) \quad \text{and} \quad \frac{\partial u(0, \eta)}{\partial \xi} = 1 \tag{42}$$

and initial conditions

$$u(\xi, 0) = \sin(\xi) + 1 \quad and \quad \frac{\partial u(\xi, 0)}{\partial \eta} = 1. \tag{43}$$

Solution 4. Application of the GB transform of order one and the given initial conditions to (41) and (42) give

$$\frac{\partial^2}{\partial \xi^2} \mathcal{P}_1(\xi, s, \gamma) + \mathcal{P}_1(\xi, s, \gamma)\left(1 - \frac{s^2}{\gamma^2}\right) + (\sin(\xi) + 1)\frac{s^2}{\gamma^2} = 0 \qquad (44)$$

with boundary conditions

$$\mathcal{P}_1(0, s, \gamma) = \frac{s}{s^2 - \gamma^2} \quad \text{and} \quad \frac{\partial}{\partial \xi} \mathcal{P}_1(0, s, \gamma) = 1. \qquad (45)$$

Here to obtain the solution of (44), the homotopy perturbation technique proposed by He [17] has been applied. In the view of (44), a perturbation equation can be readily constructed by embedding homotopy parameter $\theta \in [0, 1]$ as

$$(1-\theta)\left[\frac{\partial^2 \mathcal{P}_1}{\partial \xi^2} - \frac{\partial^2 \mathcal{P}*_{1,0}}{\partial \xi^2}\right] + \theta\left[\frac{\partial^2 \mathcal{P}_1}{\partial \xi^2} + \left(1 - \frac{s^2}{\gamma^2}\right)\mathcal{P}_1 + \frac{s^2}{\gamma^2}(\sin(\xi) + 1)\right] = 0. \quad (46)$$

Assume that the solution of the (46) can be expanded as

$$\mathcal{P}_1(\xi, s, \gamma) = \mathcal{P}_{1,0}(\xi, s, \gamma) + \theta \mathcal{P}_{1,1}(\xi, s, \gamma) + \theta^2 \mathcal{P}_{1,2}(\xi, s, \gamma) + \cdots. \qquad (47)$$

Substitution of (47) into (46) and equating the coefficients of identical powers of θ serve the system as

$$\theta^0: \qquad \frac{\partial^2 \mathcal{P}_{1,0}}{\partial \xi^2} - \frac{\partial^2 \mathcal{P}*_{1,0}}{\partial \xi^2} = 0,$$

$$\mathcal{P}_{1,0}(0, s, \gamma) = \frac{s^2}{s^2 - \gamma^2}, \quad and \quad \frac{\partial}{\partial \xi}\mathcal{P}_{1,0}(0, s, \gamma) = 1,$$

$$\theta^1: \frac{\partial^2 \mathcal{P}_{1,1}}{\partial \xi^2} + \left(1 - \frac{s^2}{\gamma^2}\right)\mathcal{P}_{1,0} + (1 + \sin(\xi))\frac{s^2}{\gamma^2} = 0, \qquad (48)$$

$$\mathcal{P}_{1,1}(0, s, \gamma) = 0, \qquad and \qquad \frac{\partial}{\partial \xi}\mathcal{P}_{1,1}(0, s, \gamma) = 0,$$

$$\vdots$$

$$\theta^j: \frac{\partial^2 \mathcal{P}_{1,1}}{\partial \xi^2} + \left(1 - \frac{s^2}{\gamma^2}\right)\mathcal{P}_{j-1} = 0,$$

$$\mathcal{P}_{1,j}(0, s, \gamma) = 0, \qquad and \qquad \frac{\partial}{\partial \xi}\mathcal{P}_{1,j}(0, s, \gamma) = 0.$$

Utilizing the freedom of initialization, set initial approximation as

$$\mathcal{P}_{1,0}(\xi, s, \gamma) = \mathcal{P}*_{1,0}(\xi, s, \gamma) = \xi + \frac{s^2}{s^2 - \gamma^2}, \tag{49}$$

which satisfies the obtained boundary conditions in (45).

Substitution of preceding components will drive the rest of components of the expanded solution in (47) as

$$\mathcal{P}_{1,1}(\xi, s, \gamma) = \left(\frac{\xi^3}{6} - \xi + \sin(\xi)\right)\frac{s^2}{\gamma^2} - \frac{\xi^3}{6},$$

$$\mathcal{P}_{1,2}(\xi, s, \gamma) = \left(\frac{\xi^5}{120} - \frac{\xi^3}{6} + \xi - \sin(\xi)\right)\frac{s^4}{\gamma^4} - \left(\frac{-\xi^5}{60} + \frac{\xi^3}{6} - \xi + \sin(\xi)\right)\frac{s^2}{\gamma^2} + \frac{\xi^5}{120}, \tag{50}$$

$$\vdots$$

Therefore, the expression for the expanded solution can be written as

$$\mathcal{P}_1(\xi, s, \gamma) = \xi + \frac{s^2}{s^2 - \gamma^2} + \left(\frac{\xi^3}{6} - \xi + \sin(\xi)\right)\frac{s^2}{\gamma^2} - \frac{\xi^3}{6}$$

$$+ \left(\frac{\xi^5}{120} - \frac{\xi^3}{6} + \xi - \sin(\xi)\right)\frac{s^4}{\gamma^4}$$

$$- \left(\frac{-\xi^5}{60} + \frac{\xi^3}{6} - \xi + \sin(\xi)\right)\frac{s^2}{\gamma^2} + \frac{\xi^5}{120}. \tag{51}$$

Utilizing the inverse GB transform over (51) with observation that $\mathbb{A}_1^{-1}\left[u(\eta)\frac{s^{n+1}}{\gamma^{n+1}}\right] = u(\eta)\frac{d^n \delta(\eta)}{d\eta^n}$ in which $\delta(\eta)$ is the Dirac delta function, that is, universally zero except at the origin. Thus, corresponding terms will get vanished, and the solution will reduce to

$$u(\xi, \eta) = \xi + \cosh(\eta) - \frac{\xi^3}{6} + \frac{\xi^5}{120} + \cdots \simeq \cosh(\eta) + \sin(\xi),$$

which is an exact solution.

Problem 5. Consider Abel's integral equation as

$$t = \int_0^\eta \frac{1}{\sqrt{\eta - \xi}} u(\xi) \, d\xi. \tag{52}$$

Solution 5. Application of the GB transform of order one and its convolution property to (52) give

$$\frac{\gamma}{s} = \frac{\gamma}{s}\Gamma\left(\frac{-1}{2}+1\right)\left(\frac{u}{s}\right)^{\frac{-1}{2}}\mathbb{A}_1[u(\eta)],$$

$$\frac{1}{\sqrt{\pi}}\left(\frac{u}{s}\right)^{\frac{-1}{2}} = A_1[u(\eta)].$$

Utilization of the inverse GB transform yields

$$\frac{2}{\pi}\eta^{\frac{1}{2}} = u(\eta),$$

which is an exact solution.

Problem 6. Consider the convolution type Volterra integral equation of first kind:

$$\sin\eta = \int_0^\eta u(\eta-\xi)u(\eta)\,d\xi. \tag{53}$$

Solution 6. Application of the GB transform of order one and its convolution property to (53) give

$$\frac{s\gamma}{s^2+\gamma^2} = \frac{\gamma}{s}\mathbb{A}_1^2[u(\eta)],$$

$$\mathbb{A}_1[u(\eta)] = \frac{s}{\sqrt{s^2+\gamma^2}}.$$

Utilization of the inverse GB transform yields

$$u(\eta) = J_0(\eta),$$

which is an exact solution.

Problem 7. Consider convolution type Volterra integral equation of second kind:

$$u(\eta) = \eta + \int_0^\eta u(\xi)\sin(\eta-\xi)\,d\xi. \tag{54}$$

Solution 7. Application of the GB transform of order one and its convolution property to (54) will derive

$$\mathbb{A}_1[u(\eta)] = \frac{\gamma}{s} + \frac{\gamma}{s}\mathbb{A}_1[u(\eta)]\mathbb{A}_1[\sin\eta].$$

Thus,

$$\mathbb{A}_1[u(\eta)] = \frac{\gamma}{s} + \frac{\gamma^3}{s^3}.$$

Utilization of the inverse GB transform yields

$$u(\eta) = \eta + \frac{\eta^3}{6},$$

which is an exact solution.

7 Conclusion

In this paper, a new integral transform operator called the GB transform has been presented along with sufficient conditions for its existence. The explanation of the duality of the GB transform with other transforms enhanced it as a generalized version of members in the class of Laplace-type transforms. For theoretical interest, the present work proved the worth of the GB transform with essential properties, viz., uniqueness, linearity, convolution, and so on. In view of applicability, the accessible features of the proposed transform are demonstrated in solving Lane–Emden type, wave-like, and convolution-type equations. In addition, the construction of weighted norm inequalities for integral convolution with the GB transform extends the scope of its study for the future also. In the future, we intend to apply the GB transform over fractional equations and will propose their bounds.

References

[1] Aboodh, K.S. *The new integral transform 'Aboodh transform'*, Glob. J. Pure Appl. Math. 9(1) (2013), 35–43.

[2] Aggarwal, S., Gupta, A.R. and Kumar, D. *Mohand transform of error function*, Int. J. Res. Advent Technol. 7(5) (2019), 224–231.

[3] Aggarwal, S., Gupta, A.R., Sharma, S.D., Chauhan, R. and Sharma, N. *Mahgoub transform (Laplace-Carson transform) of error function*, International Journal of Latest Technology in Engineering, Management & Applied Science 8(4) (2019), 92–98.

[4] Ahmadi, S.A.P., Hosseinzadeh, H. and Cherati, A.Y. *A new integral transform for solving higher order linear ordinary Laguerre and hermite differential equations*, Int. J. Appl. Comput. Math. 5(5) (2019), 1–7.

[5] Akinyemi, L. and Iyiola, O.S. *Exact and approximate solutions of time-fractional models arising from physics via Shehu transform*, Math. Methods Appl. Sci. 43(12) (2020), 7442–7464.

[6] Atangana, A. and Kilicman, A. *A novel integral operator transform and its application to some f FODE and FPDE with some kind of singularities*, Math. Probl. Eng. (2013), Art. ID 531984, 7 pp.

[7] Babolian, E., Biazar, J. and Vahidi, A. *A new computational method for Laplace transforms by decomposition method*, Appl. Math. Comput. 150(3) (2004), 841–846.

[8] Barnes, B. *Polynomial integral transform for solving differential equations*, Eur. J. Pure Appl. Math. 9(2) (2016), 140–151.

[9] Bochner, S., Chandrasekharan, K. and Chandrasekharan, K. *Fourier transforms*, Princeton University Press, 1949.

[10] Davies, B. and Martin, B. *Numerical inversion of the Laplace transform: a survey and comparison of methods*, J. Comput. Phys. 33(1) (1979), 1–32.

[11] Djebali, R., Mebarek-Oudina, F. and Rajashekhar, C. *Similarity solution analysis of dynamic and thermal boundary layers: further formulation along a vertical flat plate*, Physica Scripta 96(8) (2021), 085206.

[12] Elzaki, T. M. *The new integral transform Elzaki transform*, Glob. J. Pure Appl. Math. 7(1) (2011), 57–64.

[13] Filipinas, J.L.D.C. and Convicto, V.C. *On another type of transform called Rangaig transform*, International Journal 5(1) (2017), 42–48.

[14] Grinshpan, A.Z. *Weighted norm inequalities for convolutions, differential operators, and generalized hypergeometric functions*, Integral Equations Operator Theory 75(2) (2013), 165–185.

[15] Gupta, V.G., Shrama, B. and Kiliçman, A. *A note on fractional Sumudu transform*, J. Appl. Math. (2010), Art. ID 154189, 9 pp.

[16] Haroon, F., Mukhtar, S. and Shah, R. *Fractional view analysis of Fornberg– Whitham equations by using Elzaki transform*, Symmetry 14(10) (2022), 2118.

[17] He, J.-H. *Homotopy perturbation technique*, Comput. Methods Appl. Mech. Engrg. 178(3-4) (1999), 257–262.

[18] Higazy, M. and Aggarwal, S. *Sawi transformation for system of ordinary differential equations with application*, Ain Shams Eng. J. 12(3) (2021), 3173–3182.

[19] Khan, M., Gondal, M.A., Hussain, I. and Vanani, S.K. *A new comparative study between homotopy analysis transform method and homotopy perturbation transform method on a semi infinite domain*, Math. Comput. Model. 55(3-4) (2012), 1143–1150.

[20] Khan, Z.H. and Khan, W.A. *N-transform properties and applications*, NUST J. Eng. Sci. 1(1) (2008), 127–133.

[21] Kumar, S., Kumar, A., Kumar, D., Singh, J. and Singh, A. *Analytical solution of Abel integral equation arising in astrophysics via Laplace transform*, J. Egyptian Math. Soc. 23(1) (2015), 102–107.

[22] Maitama, S. and Zhao, W. *New integral transform: Shehu transform a generalization of Sumudu and Laplace transform for solving differential equations*, arXiv preprint arXiv:1904.11370 (2019).

[23] Mullineux, N. and Reed, J. *Numerical inversion of integral transforms*, Computers & Mathematics with Applications 3(4) (1977), 299–306.

[24] Ramadan, M.A., Raslan, K.R., El-Danaf, T.S. and Hadhoud, A.R. *On a new general integral transform: some properties and remarks*, J. Math. Comput. Sci. 6(1) (2016), 103–109.

[25] Raza, J., Mebarek-Oudina, F. and Ali Lund, L. *The flow of magnetised convective Casson liquid via a porous channel with shrinking and stationary walls*, Pramana 96(4) (2022) 229.

[26] Saadeh, R.Z. and Ghazal, B.F. *A new approach on transforms: Formable integral transform and its applications*, Axioms 10(4) (2021), 332.

[27] Saadeh, R., Qazza, A. and Burqan, A. *A new integral transform: Ara transform and its properties and applications*, Symmetry 12(6) (2020), 925.

[28] Sadefo Kamdem, J. *Generalized integral transforms with the homotopy perturbation method*, J. Math. Model. Algorithms Oper. Res. 13(2) (2014), 209–232.

[29] Schiff, J L. *The Laplace transform: theory and applications*, Springer Science & Business Media, 1999.

[30] Shah, K., Khalil, H. and Khan, R.A. *Analytical solutions of fractional order diffusion equations by natural transform method*, Iran. J. Sci. Technol. Trans. A Sci. 42(3) (2018), 1479–1490.

[31] Silva, F.S., Moreira, D.M. and Moret, M.A. *Conformable Laplace transform of fractional differential equations*, Axioms 7(3) (2018), 55.

[32] Tripathi, R. and Mishra, H.K. *Homotopy perturbation method with Laplace transform (LT-HPM) for solving Lane–Emden type differential equations (LETDEs)*, SpringerPlus 5(1) (2016), 1–21.

[33] Watugala, G. *Sumudu transform: a new integral transform to solve differential equations and control engineering problems*, Internat. J. Math. Ed. Sci. Tech. 24(1) (1993), 35–43.

[34] Widder, D.V. *Laplace transform (PMS-6)*, Princeton university press, 2015.

[35] Yang, X.-J. *A new integral transform with an application in heat-transfer problem*, Ther. Sci. 20(3) (2016), 677–681.

[36] Yüzbaı, ., Sezer, M. and Kemancı, B. *Numerical solutions of integro-differential equations and application of a population model with an improved Legendre method*, Appl. Math. Model. 37(4) (2013), 2086–2101.

[37] Zhao, W. and Maitama, S. *Beyond sumudu transform and natural transform*: *transform properties and applications*, J. Appl. Anal. Comput. 10(4) (2020), 1223–1241.

How to cite this article

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