



Iranian Journal of Numerical Analysis and Optimization

Volume 11 , Number 2

Summer 2021

Serial Number: 20

In the Name of God

Iranian Journal of Numerical Analysis and Optimization (IJNAO)

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

Volume 11, Number 2, Summer 2021

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

Publisher: Faculty of Mathematical Sciences, Ferdowsi University of Mashhad

Published by: Ferdowsi University of Mashhad Press

Printing Method: Electronic

Address: Iranian Journal of Numerical Analysis and Optimization

Faculty of Mathematical Sciences, Ferdowsi University of Mashhad

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

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

E-mail: ijnao@um.ac.ir

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

This journal is indexed by:

- Scopus database of Elsevier (from 31/07/2020)
- Zentralblatt
- ISC
- SID
- Civilica
- Magiran
- DOAJ
- OAJI
- CiteFactor
- AcademicKeys
- COPE
- Mendeley
- Academia.edu
- LinkedIn
- The Journal granted the Scientific-Research degree by the Iranian Ministry of Science, Research, and Technology.

Iranian Journal of Numerical Analysis and Optimization

Volume 11, Number 2, Summer 2021

Ferdowsi University of Mashhad - Iran

©2021 All rights reserved. Iranian Journal of Numerical Analysis and Optimization

Iranian Journal of Numerical Analysis and Optimization

Director

M. H. Farahi

Editor-in-Chief

Ali R. Soheili

Managing Editor

M. Gachpazan

EDITORIAL BOARD

Abbasbandi, S.*

(Numerical Analysis)

Imam Khomeini International University,
Iran.

e-mail: abbasbandy@ikiu.ac.ir

Area, I.*

(Numerical Analysis)

Universidade de Vigo, Spain.

e-mail: area@uvigo.es

Babolian, E.*

(Numerical Analysis)

Kharazmi University, Iran.

e-mail: babolian@khu.ac.ir

Dehghan, M.*

(Numerical Analysis)

Amirkabir University of Technology, Iran.

e-mail: mdehghan@aut.ac.ir

Effati, S.*

(Optimal Control & Optimization)

Ferdowsi University of Mashhad, Iran.

e-mail: s-effati@um.ac.ir

Emrouznejad, A.*

(Operations Research)

Aston University, UK.

e-mail: a.emrouznejad@aston.ac.uk

Farahi, M. H.*

(Optimal Control & Optimization)

Ferdowsi University of Mashhad, Iran.

e-mail: farahi@um.ac.ir

Gachpazan, M.**

(Numerical Analysis)

Ferdowsi University of Mashhad, Iran.

e-mail: gachpazan@um.ac.ir

Ghanbari, R.**

(Operations Research)

Ferdowsi University of Mashhad, Iran.

e-mail: rghanbari@um.ac.ir

Hadizadeh Yazdi, M.**

(Numerical Analysis)

Khaje-Nassir-Toosi University of
Technology, Iran.

e-mail: hadizadeh@kntu.ac.ir

Hojjati, GH. R.*

(Numerical Analysis)

University of Tabriz, Iran.

e-mail: ghojjati@tabrizu.ac.ir

Hong, J.*

(Scientific Computing)

Chinese Academy of Sciences (CAS), China.

e-mail: hjl@lsec.cc.ac.cn

Khojasteh Salkuyeh, D.*

(Numerical Analysis)

University of Guilan, Iran.

e-mail: khojasteh@guilan.ac.ir

Lohmander, P.*

(Optimization)

Swedish University of Agricultural Sciences,
Sweden.

e-mail: Peter@Lohmander.com

Lopez-Ruiz, R.**

(Complexity, nonlinear models)

University of Zaragoza, Spain.

e-mail: rilopez@unizar.es

Mahdavi-Amiri, N.*

(Optimization)

Sharif University of Technology, Iran.

e-mail: nezamm@sina.sharif.edu

Salehi Fathabadi, H.*

(Operations Research)

University of Tehran, Iran.

e-mail: hsalehi@ut.ac.ir

Soheili, Ali R.*

(Numerical Analysis)

Ferdowsi University of Mashhad, Iran.

e-mail: soheili@um.ac.ir

Soleimani Damaneh, M.*

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

University of Tehran, Iran.

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

Toutounian, F.*

(Numerical Analysis)

Ferdowsi University of Mashhad, Iran.

e-mail: toutouni@um.ac.ir

Türkyılmazoğlu, M.*

(Applied Mathematics)

Hacettepe University, Turkey.

e-mail: turkyilm@hacettepe.edu.tr

Kamyad, A.V.*

(Optimal Control & Optimization)

Ferdowsi University of Mashhad, Iran.

e-mail: vahidian@um.ac.ir

Xu, Z.*

(Decision Making)

Sichuan University, China.

e-mail: xuzeshui@263.net

Vasagh, Z.

(English Text Editor)

Ferdowsi University of Mashhad, Iran.

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

* Full Professor

** Associate Professor

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

Letter from the Editor-in-Chief

I would like to welcome you to the Iranian Journal of Numerical Analysis and Optimization (IJNAO). This journal is published two issues per year and supported by the Faculty of Mathematical Sciences at the Ferdowsi University of Mashhad. Faculty of Mathematical Sciences with three centers of excellence and three 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 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.

Ali R. Soheili

Contents

Trainable fourth-order partial differential equations for image noise removal	235
N. Khoeiniha, S.M. Hosseini and R. Davoudi	
Exponentially fitted tension spline method for singularly perturbed differential difference equations	261
M.M. Woldaregay and G.F. Duressa	
New class of hybrid explicit methods for numerical solution of optimal control problems	283
M. Ebadi, I. Malih Maleki and A. Ebadian	
The strict complementarity in linear fractional optimization	305
M. Mehdiloo, K. Tone and M.B. Ahmadi	
Solving quantum optimal control problems by wavelets method	333
M. Rahimi, S. M.Karbassi and M.R. Hooshmandasl	
Singularly perturbed robin type boundary value problems with discontinuous source term in geophysical fluid dynamics	351
B.M. Abagero, G.F. Duressa and H.G. Debela	
Two new approximations to Caputo–Fabrizio fractional equation on non-uniform meshes and its applications	365
Z. Soori and A. Aminataei	
Application of Newton–Cotes quadrature rule for nonlinear Hammerstein integral equations	385
A. Shahsavaran	
Investigating a claim about resource complexity measure	401

H.R. Yousefzadeh

A new algorithm for solving linear programming problems with bipolar fuzzy relation equation constraints	407
---	------------

S. Aliannezhadi and A. Abbasi Molai

Review of the strain-based formulation for analysis of plane structures, Part I: Formulation of basics and the existing elements	437
---	------------

M. Rezaiee-Pajand, N. Gharaei-Moghaddam and M. Ramezani

Review of the strain-based formulation for analysis of plane structures, Part II: Evaluation of the numerical performance	485
--	------------

M. Rezaiee-Pajand, N. Gharaei-Moghaddam and M. Ramezani



Trainable fourth-order partial differential equations for image noise removal

N. Khoeiniha*, S.M. Hosseini and R. Davoudi

Abstract

Image processing by partial differential equations (PDEs) has been an active topic in the area of image denoising, which is an important task in computer vision. In PDE-based methods for unprocessed image processing, the original image is considered as the initial value for the PDE and the solution of the equation is the outcome of the model. Despite the advantages of using PDEs in image processing, designing and modeling different equations for various types of applications have always been a challenging and interesting problem. In this article, we aim to tackle this problem by introducing a fourth-order equation with flexible and trainable coefficients, and with the help of an optimal control problem, the coefficients are determined; therefore the proposed model adapts itself to each particular application. At the final stage, the image enhancement is performed on the noisy test image and the performance of our proposed method is compared to other PDE-based models.

AMS subject classifications (2020): 35G25; 68U10; 90C90.

Keywords: Partial Differential Equations; Image Processing; Image Denoising; Optimal Control.

*Corresponding author

Received 14 December 2020; revised 8 February 2021; accepted 6 April 2021

Negin Khoeiniha

Department of Applied Mathematics, Faculty of Mathematical Sciences, Tarbiat Modares University, Tehran, Iran. e-mail: Neginkhoeiniha@modares.ac.ir

Mohammad Hosseini

Department of Applied Mathematics, Faculty of Mathematical Sciences, Tarbiat Modares University, Tehran, Iran. e-mail: hossei_m@modares.ac.ir

Ramtin Davoudi

Department of Applied Mathematics, Faculty of Mathematical Sciences, Tarbiat Modares University, Tehran, Iran. e-mail: R.davoudi@modares.ac.ir

1 Introduction

Image denoising is an important preliminary step in many computer vision and image processing problems that can affect further processing, and has been an open research area for a long time [28, 31, 38]. By considering the image as a function I in \mathbb{R}^2 , the aim of denoising an image is to extract the clear image I from the noisy image I_0 that has been degraded by the model $I_0 = I + n$, in which n is a noise function.

Various approaches for image denoising have been developed, such as methods based on nonlocal means filters [4, 18], wavelets [8, 25], Perona and Malik (PM) [29], block-matching, and three-dimensional filtering [6, 7, 26], sparse representation [46], adaptive image filtering [1, 2], bilateral filtering [35, 10], and so on.

In contrast to the discrete methods used in signal processing in the past, which were the basis of digital image processing, in the past decades, continuous methods based on partial differential equations (PDEs) have become an effective tool.

The first efforts of using PDEs in computer vision and image processing tasks date back to the 1960s [11, 15]; however, these methods did not attract much attention. In the 1980s, the importance of multi-scale descriptions of images has been recognized. The importance of scale-space filtering in images was introduced by Witkin [41] and then developed by Koenderink [17]. The idea of their approach is to generate a family of scaled images $I(x, y, t)$ by convolving the original image $I_0(x, y)$ with a Gaussian kernel $G(x, y, t)$ of variance t as follows:

$$I(x, y, t) = I_0(x, y) \otimes G(x, y, t).$$

The time variable t is said to be the scale parameter, and the higher the scale gets, it results in an image with reduced resolution; therefore, the noise vanishes.

It was pointed out by Koenderink [17] that, $I(x, y, t)$ is the fundamental solution to the heat conduction equation as below, with the original image $I_0(x, y)$ as the initial condition and the conductance coefficient c , a real number:

$$I_t = c(I_{xx} + I_{yy}).$$

Although the outcome of this equation is a less noisy image, the main disadvantage of this approach is the fact that while smoothing, it blurs important image features; in other words, the Gaussian scale-space does not respect the edges of objects in an image. Therefore, the scaled image loses important details after the process.

In 1992, a method introduced by Perona and Malik on anisotropic diffusion [29] drew great attention to PDE methods in image enhancement. In fact, they designed the following equation, known as PM second-order PDE,

to achieve a good balance between noise removal and retaining the edges of objects in the image:

$$I_t = c(x, y, t)(I_{xx} + I_{yy}).$$

In the above PDE, the conductance coefficient function, $c(x, y, t)$, is chosen to decrease smoothing around the edges and reduce noise in other areas. Following their model, many equations have been introduced on this basis; see for instance [16, 39, 5].

Although the PM second-order PDE and its variances were successfully applied in image denoising problems; these methods cause blocky effects on the processed image. In [43], it is noted that the PM PDE is a second-order model, and this feature guarantees its ability to reconstruct images with discontinuities but is responsible for the blocky effect [45].

In recent years, many efforts have been made to resolve this problem in the PM method, such as introducing different coefficients, different equations for processing, raising the order of the equation, and so on. One of the innovative efforts to tackle the problems of second-order approaches was inspired by learning-based methods in machine learning [19]. Afterward, some second-order learning PDEs (L-PDEs) were introduced that adopt a technique called PDE-based optimal control [14] and use training images to learn the coefficients in the PDE, and once they are computed, then the PDE is obtained and can be applied to test images.

In learning PDEs proposed in [19, 22, 23], the coefficient functions are assumed to minimize a functional subject to some PDE constraints that are designed according to the problem.

In these classes of approaches, the structure of the problem is considered, and the learned coefficients will form the PDE according to the training images. Thus, the most effort in obtaining a PDE is to prepare some input/output training image pairs.

One of the other successful strategies in enhancing the performance of second-order PDEs for image denoising is to design higher-order equations in replace of quadratic ones [13, 21, 24]. In 2000, in a method proposed by You and Kaveh, known as YK fourth-order PDE [42], the following equation was introduced:

$$\frac{\partial I}{\partial t} = -\nabla^2 (c_1 (|\nabla^2 I|) |\nabla^2 I|), \quad (1)$$

where ∇^2 denotes the Laplacian operator and conductance coefficient $c_1(\cdot)$ in YK fourth-order PDE, is a function of the absolute value of the Laplacian of the image intensity, that is,

$$c_1 (|\nabla^2 I|) = \frac{1}{1 + (|\nabla^2 I|/k)^2}, \quad (2)$$

in which $k > 0$ is the Laplacian threshold. The YK fourth-order PDE attempts to remove noise and preserve edges by approximating an observed image with a piecewise planar image [42].

In recent years, more PDE-based models have been developed, improving the performance of the YK method. In 2016, Zhang and Ye [47] proposed the adaptive fourth-order PDE for noise removal

$$\frac{\partial u}{\partial t} = -\text{div}^2(h(\nabla u, D^2 u) \frac{D^2 u}{|D^2 u|}) + \mu \text{div}((1 - h(\nabla u, D^2 u)) \frac{\nabla u}{|\nabla u|}),$$

where $h(\nabla u, D^2 u)$ is an edge detector chosen by

$$h(\nabla u, D^2 u) = \frac{1}{1 + k_1 |\nabla(G_\sigma \otimes u)|^2 + k_2 |D^2(G_\sigma \otimes u)|^2 + \gamma}, \quad (3)$$

in which $\gamma > 0$ is a small constant to guarantee $h < 1$, k_1 and k_2 are positive constants, and $\mu > 0$ and $\lambda > 0$ are parameters balancing the contribution of the three terms in the objective function.

Later in 2018, Siddig et al. [33] introduced the following equation:

$$\begin{cases} \frac{\partial u}{\partial t} + D_{ij}^2 \left(\frac{\alpha(u) D_{ij}^2 u}{|D_{ij}^2 u|} \right) = 0, & (x, t) \in \Omega_T = (0, T) \times \Omega, \\ u(x, t) = 0, & (x, t) \in (0, T) \times \partial\Omega, \\ \frac{\partial u}{\partial \vec{n}} = 0, & (x, t) \in (0, T) \times \partial\Omega, \\ u(x, 0) = u_0(x), & x \in \Omega, \end{cases}$$

where $\alpha(u) = \frac{1}{\sqrt{1 + |G_\sigma \otimes \nabla u|^2}}$ and $G_\sigma(x)$ is the Gaussian filter with σ as its parameter. In these models (Siddig's and Zhang's PDEs), they tend to overcome the disadvantages of YK and TV model while keeping the pros of PDE algorithms.

The theoretical analysis in [9, 12], shows the advantages of fourth-order equations over second-order ones. First, fourth-order diffusion reduces fluctuation faster than second-order ones. Second, since the second-order PDE will evolve toward a piecewise constant approximation in smooth regions, unlike second-order PDE, fourth-order PDE will evolve toward and settle down to a piecewise smooth image if the image support is infinite [44].

In this article, motivated by the previous efforts on solving second-order equations' problems, we propose a fourth-order PDE with trainable coefficients for image denoising to enhance the performance of the second-order L-PDEs. To mention two advantages of this PDE, one should note the advantage of using fourth-order PDEs, and also the proposed PDE uses image differential invariants up to fourth-order, which can detect more image features compared to those of second-order PDEs. The results obtained from our scheme are compared to the state-of-the-art previous models, such as second-order learning-based PDEs [22], YK fourth-order PDE [42], and a

more recent fourth-order PDEs, Siddig [33], and Zhang's [47], on this topic. The experimental results show that, in terms of subjective and objective measures, the proposed model can outperform the pre-existing methods.

The rest of this article is organized as follows. Section 2 presents the trainable fourth-order PDE with a detailed description of the solution of the optimal control problem. Experimental results are presented in Section 3, and the article is concluded in Section 4.

2 Proposed model

In this section, first, the formulation of the fourth-order PDE in our model for noise reduction is presented and then an optimal control problem will be introduced to assist in completing the PDE. Finally, the optimality conditions will be investigated for the recommended model, which helps us to extract enough information to solve the problem.

The framework of this article is based on the fourth-order noise removal PDE models, and inspired by second-order learning PDEs mentioned in the previous section we will consider flexible trainable coefficients for the proposed PDE.

In this work, the typical notations in image processing and optimal control literature are used. We denote f as the input image and I as the desired output image. Moreover, Ω is an open bounded region of \mathbb{R}^2 , $\partial\Omega$ is its boundary, the spatial variable (x, y) belongs to Ω , $t \in (0, T_f)$ is the temporal variable, $\Omega \times (0, T_f)$ is named Q , and Γ is its boundary. We then define the set η as bellow and $|p|$ is the length of string p in which $p \in \eta$:

$$\eta = \{0, x, y, xy, xx, yy, xxx, xxy, xyy, yyy, xxxx, xxxy, xxyy, yyyy, yyyy\}$$

Changing the viewpoints of an object causes a geometric deformation, which can be modeled by a two-dimensional Euclidean, similarity, affine, or projective transformation group. Many methods have been designed to correctly recognize planar objects, by extracting image invariant features for the action of various two-dimensional transformation groups, and differential invariants are one of them. We can express image differential invariants as the functions of image partial derivatives, and they are widely used to describe local image structures. Therefore, to construct our PDE, we used a set of differential invariants up to fourth-order. We will use a set of differential invariants that are more widely used.

Thus, the main reason for using these types of partial derivatives to form our equation is to create a model that can cover more image features compared to second-order ones and therefore be more sensitive to edges in the image. Furthermore, these differential invariants stay unchanged under various two-dimensional transformations like rotation and translation [27]. Consider

the fourth-order PDE below:

$$\begin{cases} \frac{\partial I}{\partial t} = \mathbf{u}(t)^T \text{diff}(I), & \text{in } Q, \\ I(x, y, t) = 0, & \text{on } \Gamma, \\ I(x, y, 0) = f, & \text{in } \Omega, \end{cases} \quad (4)$$

in which $\mathbf{u}(t) = [u_0(t) \cdots u_6(t)]^T$ is the coefficient function and $\text{diff}(\cdot)$ includes the differential parts defined as

$$\begin{aligned} \text{diff}_1(I) &= I, \\ \text{diff}_2(I) &= I_y^2 + I_x^2, \\ \text{diff}_3(I) &= I_{xx}I_{yy} - I_{xy}^2, \\ \text{diff}_4(I) &= I_{xx}I_y^2 - 2I_{xy}I_xI_y + I_{yy}I_x^2, \\ \text{diff}_5(I) &= I_{xx}I_xI_y - I_{xy}I_y^2 - I_{xy}I_x^2 + I_{yy}I_xI_y, \\ \text{diff}_6(I) &= I_{xxx}I_x^3 + 3I_{xxy}I_yI_x^2 + 3I_{xyy}I_y^2I_x + I_{yyy}I_y^3, \\ \text{diff}_7(I) &= I_{xxx}I_x^4 + 4I_{xyy}I_y^3I_x + 6I_{xxy}I_x^2I_y^2 + 4I_{xxy}I_yI_x^3 + I_{yyy}I_y^4. \end{aligned} \quad (5)$$

The initial value for this equation is f , the noisy image, and the output $I(\cdot)$ is the desired clear image. To obtain a workable equation, we need to determine the coefficient function $\mathbf{u}(t)$, which is used to control the evolution of I . We aim to attain a flexible system that can train the coefficients; then the function $\mathbf{u}(t)$ would vary in different problems and be adapted according to the given training images.

As mentioned previously, by using differential invariants in our model, the PDE system (4) is rotationally invariant and the coefficient functions must be independent of (x, y) so that (4) is shift-invariant as well.

Proposition 1. Considering equation (4), the control functions $\{u_i\}_{i=0}^6$ must be independent of (x, y) .

Proof. Considering

$$D(I, \mathbf{u}) := \mathbf{u}(t)^T \text{diff}(I), \quad (6)$$

we then rewrite $D(I, \mathbf{u}) = \tilde{D}(I, x, y, t)$. Therefore, it is enough to prove the independence of \tilde{D} of (x, y) .

Due to the translation invariance of (4), when we change $f(x, y)$ to $f(x - x', y - y')$, $I(x, y)$ changes to $I(x - x', y - y')$ and we have

$$\frac{\partial I(x - x', y - y')}{\partial t} = \tilde{D}(I(x - x', y - y'), x, y, t).$$

Then, by the conversion $x - x' = x$ and $y - y' = y$, we get

$$\frac{\partial I(x, y)}{\partial t} = D(I(x, y), x + x', y + y', t).$$

On the other hand,

$$\begin{aligned} \frac{\partial I(x, y)}{\partial t} &= D(I(x, y), x, y, t), \\ \Rightarrow D(I, x + x', y + y', t) &= D(I, x, y, t) \quad \text{for all } (x', y'). \end{aligned}$$

Hence, D is independent of (x, y) , which indicates the independence of the function \mathbf{u} . \square

2.1 Extracting the coefficient function

Attempting to solve the problem (4), we need to determine the coefficients $\mathbf{u}(t)$, which is done by an optimal control problem. Therefore we recollect some necessary definitions in this field. We begin by assuming $I \in H_0^1(\Omega)$, defined to be the closure of the infinitely differentiable functions compactly supported in Ω in $H_1(\Omega)$.

Definition 1. Considering the optimal control problem

$$\begin{aligned} &\min J(I, u) \\ &s.t. \ T(I, u) = 0 \\ &u \in U_{ad} \end{aligned} \tag{7}$$

where $J : H_0^1(\Omega) \times L^2(\Omega) \rightarrow \mathbb{R}$ is called the objective function of the optimal control problem and $U_{ad} \in L^2(\Omega)$ is the admissible set, a nonempty set in \mathbb{R}^m .

A vector $\bar{u} \in U_{ad}$ is called an optimal control for (7), if $J(\bar{I}, \bar{u}) \leq J(I, u)$ for each $u \in U_{ad}$ and \bar{I} is called the optimal state associated with \bar{u} .

Definition 2 ([36]). Let U and V be normed vector spaces, let $U_{ad} \subseteq U$, and let $f : U_{ad} \rightarrow V$, and consider some $u \in U_{ad}$. Let $h \in U$. If $u + th \in U_{ad}$ for sufficiently small $t > 0$, and the limit

$$\delta f(u, h) = \lim_{t \rightarrow 0} \frac{f(u + th) - f(u)}{t},$$

exists in V , then $\delta f(u, h)$ is called the directional derivative of F at u in the direction h . If the directional derivative $\delta f(u, h)$ exists for each h , then the map

$$\delta f(u, \cdot) : U \rightarrow V, h \mapsto \delta f(u, h),$$

is called the first variation of f at u . In addition, if the first variation constitutes a bound linear operator, then it is called the Gateaux derivative of f .

Definition 3 ([36]). Let X and Y be Banach spaces and let $A : X \rightarrow Y$ be a bounded linear operator. The map

$$A^* : Y^* \rightarrow X^*, A^*(f) = f \circ A,$$

is called the adjoint of A . For the optimal control problem (7), the adjoint equation is of the form

$$D_y T(I, u)^T \lambda = \nabla_y J(y, u), \quad (8)$$

and its solution λ is the adjoint state.

Definition 4 ([36]). The function $L(I, u, \lambda) : H_0^2(\Omega) \times L^2(\Omega) \times H_0^2(\Omega) \rightarrow \mathbb{R}$ defined as bellow is called the Lagrangian function for the problem (7):

$$L(I, u, \lambda) := J(I, u) - \langle T(I, u), \lambda \rangle \quad (9)$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product defined on $H_0^2(\Omega)$.

Our motivation here is the optimal heating problem, which uses an optimal control problem to get the desired output from a heat equation. For instance, in a transient heating problem, the goal is to reach a certain temperature \bar{y} by controlling the heating elements [30]. Likewise, our objective in this article is to approximate the target clean noise-free image \tilde{I} by controlling the coefficient function. Therefore, a PDE-based optimal control problem will be presented. As the first step, we need to prepare some input/output training images (f_k, \tilde{I}_k) , where f_k is the noisy image and \tilde{I}_k is the clear image. The final output of the equation needs to be close to the ground truth; thus the coefficient functions must minimize the following functional:

$$J(\{I_k\}_{k=1}^K, \mathbf{u}) = \frac{1}{2} \sum_{k=1}^K \int_{\Omega} (I_k(T_f) - \tilde{I}_k)^2 d\Omega + \frac{1}{2} \sum_{i=0}^6 \alpha_i \int_0^{T_f} u_i^2(t) dt, \quad (10)$$

where K is the number of the training images, $I_k(T_f)$ is the output image determined from (4) at time $t = T_f$ when the initial value is f_k , and α_i are positive weighting parameters¹. The first term of the functional J requires the final output of our PDE to be close to the ground truth and the second term is for regularization so that this optimal control problem is well-posed and counteracts the tendency of the control to become locally bounded as J approaches its minimum [36].

We have the following optimal control problem with PDE constrains:

$$\min_{\mathbf{u}} J(\{I_k\}_{k=1}^K, \mathbf{u}); \quad s.t. \quad \begin{cases} \frac{\partial I_k}{\partial t} = \mathbf{u}(t)^T \text{diff}(I_k), & \text{in } Q, \\ I_k(x, y, t) = 0, & \text{on } \Gamma, \\ I_k(x, y, 0) = f_k, & \text{in } \Omega. \end{cases} \quad (11)$$

In addition, we consider an inequality constraint $0 \leq \mathbf{u}(t)$ for the problem (11) to make sure that the coefficients are all positive. In the following, the

¹ in this article, we fix $\alpha_i = 10^{-7}$, $i = 0, \dots, 6$.

necessary conditions for this problem (11) will be examined, so that we would be able to solve the problem by proper numerical methods.

Lemma 1. Suppose that $J(I, u) = \int_Q g(u, I) dQ$ is of class C^1 , where g is a smooth function. Then considering the function $D(I, \mathbf{u})$ in (6) as

$$\frac{DJ}{Du} = D_I^*(I, u)(\lambda) + g_I^*(I, u)(T_f),$$

where D_I^* and g_I^* are the adjoint operators of D_I and g_I , respectively, $\frac{DJ}{Du}$ is the Gateaux derivative of J with respect to the control u , and λ is the adjoint state of (4).

Proof. We first define an operator ψ , mapping u to the solution of (4), and set $I = \psi(u)$. Now consider

$$d = \lim_{\varepsilon \rightarrow 0} \frac{\psi(u + \varepsilon \delta u) - \psi(u)}{\varepsilon},$$

where δu is the perturbation of u .

We have

$$\begin{aligned} \left(\frac{DJ}{Du}, \delta u \right)_Q &= \lim_{\varepsilon \rightarrow 0} \frac{J(I, u + \varepsilon \delta u) - J(I, u)}{\varepsilon} \\ &= \lim_{\varepsilon \rightarrow 0} \frac{J(\psi(u + \varepsilon \delta u), u + \varepsilon \delta u) - J(\psi(u), u)}{\varepsilon} \\ &= \lim_{\varepsilon \rightarrow 0} \frac{J(I + \varepsilon d + o(\varepsilon), u + \varepsilon \delta u) - J(I, u)}{\varepsilon} \\ &= \int_Q (g_I(u, I)(d) + g_I(u, I)(\delta u)) dQ \\ &= (g_I(u, I)(d), 1)_Q + (g_I(u, I)(\delta u), 1)_Q \\ &= (g_I^*(u, I)(1), d)_Q + (g_I^*(u, I)(1), \delta u)_Q. \end{aligned} \quad (12)$$

If we take λ as the adjoint state of (11), then from the PDE constrains of (11), we have

$$((\psi(u))_t, \lambda)_Q = (D(u, \psi(u)), \lambda)_Q \quad (13)$$

and

$$((\psi(u + \varepsilon \delta u))_t, \lambda)_Q = (D(u + \varepsilon \delta u, \psi(u + \varepsilon \delta u)), \lambda)_Q. \quad (14)$$

Now by subtracting (14) with (13), we have

$$\begin{aligned} &((\psi(u + \varepsilon \delta u) - \psi(u))_t, \lambda)_Q \\ &= (D(u + \varepsilon \delta u, \psi(u + \varepsilon \delta u)) - D(u, \psi(u)), \lambda)_Q \\ &= (\varepsilon D_u(u, I)(\delta u) + D_I(u, I)(\psi(u + \varepsilon \delta u) - \psi(u)), \lambda) + o(\varepsilon). \end{aligned} \quad (15)$$

In this step, we divide both sides with ε , and let $\varepsilon \rightarrow 0$. Therefore

$$(d_t, \lambda)_Q = (D_u(u, I)(\delta u), \lambda)_Q + (D_I(u, I)(d), \lambda)_Q. \quad (16)$$

Integrating by parts implies

$$(d, \lambda)_\Omega|_0^T - (d, \lambda_t)_Q = (\delta u, D_u^*(u, I)(\lambda))_Q + (d, D_I^*(u, I)(\lambda)). \quad (17)$$

Now, since λ is the adjoint equation of the problem, we have

$$(d, \lambda)_\Omega|_0^T = 0, (g_I^*(u, I)(T), d)_Q = (-\lambda_t - D_I^*(u, I)(\lambda), d), \quad (18)$$

and combining it with (16) yields

$$(g_I^*(u, I)(T), d)_Q = (-\lambda_t - D_I^*(u, I)(\lambda), d) = (\delta u, D_u^*(u, I)(\lambda)). \quad (19)$$

Therefore we arrive at

$$\frac{DJ}{Du} = g_I^*(u, I)(T) + D_u^*(u, I)(\lambda).$$

□

Now, we investigate the conditions that the optimal vectors \bar{u} and \bar{I} must satisfy to be then determined, using numerical methods.

Theorem 1. Consider the optimal control problem (11), and suppose that

$$U_{ad} = \{\mathbf{u} \in L_2(\Omega); \mathbf{u}_a \leq \mathbf{u}\} \quad (20)$$

is the admissible control set, where $U_{ad} \subseteq O \subseteq L^2(\Omega)$, O is an open subset of $L^2(\Omega)$. In our case, $\mathbf{u}_a = 0$ and $J : H_0^1(\Omega) \times L_2(\Omega) \rightarrow \mathbb{R}$ is of class C^1 . If $\bar{\mathbf{u}} \in U_{ad}$ is an optimal control for (11) in the sense of Definition 1 with corresponding state \bar{I} and adjoint state $\bar{\lambda}$, then $\bar{\mathbf{u}} = (\bar{u}_i)_{0 \leq i \leq 6}$ satisfies for each $i \in \{0, \dots, 6\}$

$$\bar{u}_i = u_{a,i}, \quad \text{where } L_u(\bar{I}, \bar{u}, \bar{\lambda}) > 0. \quad (21)$$

In addition, by introducing the extended Lagrange function

$$L(I, u, \lambda, \mu_a) = L(I, u, \lambda) + \langle u_a - u, \mu_a \rangle, \quad (22)$$

and letting

$$\mu_a := \max\{0, L_u(I, u, \lambda)\},$$

the 4-tuples $(\bar{I}, \bar{u}, \bar{\lambda}, \mu_a) \in (H_0^2(\Omega), L^2(\Omega), H_0^2(\Omega), L^2(\Omega))$ satisfies

$$\frac{\partial J}{\partial u_i} = \alpha_i u_i - \sum_{k=1}^K \int_{\Omega} \lambda_k \text{diff}_i(I_k) d\Omega - \mu_a, \quad i = 0, \dots, 6. \quad (23)$$

Proof. First note that, (20) implies that U_{ad} is convex. We have the following optimality system which can be used to determine the optimal control \bar{u} and \bar{I} :

$$\begin{aligned} \nabla_{\lambda} L(\bar{y}, \bar{u}, \bar{\lambda}) &= 0, \\ \nabla_I L(\bar{y}, \bar{u}, \bar{\lambda}) &= 0, \\ \nabla_u L(\bar{y}, \bar{u}, \bar{\lambda}) &= 0, \\ \langle L_u(\bar{I}, \bar{u}, \bar{\lambda}), u - \bar{u} \rangle &\geq 0, \text{ for each } u \in U_{ad}. \end{aligned} \quad (24)$$

Every solution to the general form of optimal control (7) must satisfy this system [36].

Slightly rearranging the fourth optimal condition, we get

$$\langle L_u(\bar{I}, \bar{u}, \bar{\lambda}), \bar{u} \rangle_{L^2(\Omega)} \leq \langle L_u(\bar{I}, \bar{u}, \bar{\lambda}), u \rangle_{L^2(\Omega)}.$$

It implies that \bar{u} is the solution to the optimal problem

$$\min_{u \in U_{ad}} \langle L_u(\bar{I}, \bar{u}, \bar{\lambda}), u \rangle_{L^2(\Omega)} = \min \sum_{i=0}^6 L_u(\bar{I}, \bar{u}_i, \bar{\lambda}) u_i. \quad (25)$$

The components u_i can be varied independently because of the form of U_{ad} , such that the sum in (25) attains its min if and only if each summand is minimal. Therefore,

$$L_u(\bar{I}, \bar{u}_i, \bar{\lambda}) \bar{u}_i = \min_{u \in U_{ad}} L_u(\bar{I}, \bar{u}_i, \bar{\lambda}) u_i. \quad (26)$$

Now, (21) is a direct result of (26).

Moreover, according to the definition of $L(I, u, \lambda, \mu_a)$ in (22), we have

$$\begin{aligned} \nabla_{\lambda} L(I, u, \lambda, \mu_a) &= \nabla_{\lambda} L(\bar{I}, \bar{u}, \bar{\lambda}), \\ \nabla_I L(I, u, \lambda, \mu_a) &= \nabla_I L(\bar{I}, \bar{u}, \bar{\lambda}), \\ \nabla_u L(I, u, \lambda, \mu_a) &= \nabla_u L(\bar{I}, \bar{u}, \bar{\lambda}) - \mu_a. \end{aligned} \quad (27)$$

Therefore, we continue the proof with the old form of Lagrangian for simplicity. By Definition 4, the Lagrangian function for the optimal control problem (11) is as follows:

$$L(\{I_k\}_{k=1}^K, \mathbf{u}) = J(\{I_k\}_{k=1}^K, \mathbf{u}) + \sum_{k=1}^K \int_Q \lambda_k ((I_k)_t - D(I_k, \mathbf{u})) dQ, \quad (28)$$

where the multiplier λ_k is the adjoint state. It can be seen that the first optimality condition $\nabla_{\lambda} L(\bar{y}, \bar{u}, \bar{\lambda}) = 0$ is the PDE constraint in (11).

To find the adjoint state λ , we first perturb $D(I_k)$ with respect to I :

$$\begin{aligned}
& D(I_k + \varepsilon \delta I_k, \mathbf{u}) - D(I_k, \mathbf{u}) \\
&= \varepsilon \cdot \left(\frac{\partial D}{\partial (I_k)_x} \frac{\partial (\delta I_k)}{\partial x} + \cdots + \frac{\partial D}{\partial (I_k)_{yyyy}} \frac{\partial^4 (\delta I_k)}{\partial y^4} \right) + O(\varepsilon) \\
&= \varepsilon \cdot \sum_{p \in \eta} \sigma_p(I_k) \frac{\partial^{|p|} (\delta I_k)}{\partial p} + O(\varepsilon),
\end{aligned} \tag{29}$$

in which

$$\sigma_p(I_k) = \frac{\partial D(I_k, \mathbf{u})}{\partial I_k} = \mathbf{u}^T(t) \cdot \frac{\partial (\text{diff}(I_k))}{\partial (I_k)_p} = \sum_{i=0}^6 u_i \frac{\partial \text{diff}_i(I_k)}{\partial (I_k)_p}. \tag{30}$$

Besides, for having $L_{I_k}(I, u, \lambda)$, we perturb I_k in L :

$$\begin{aligned}
L_{I_k}(I, u, \lambda) &= \lim_{\varepsilon \rightarrow 0} \frac{L(\dots, I_k + \varepsilon \delta I_k, \dots) - L(\dots, I_k, \dots)}{\varepsilon} \\
&= \lim_{\varepsilon \rightarrow 0} \left(\frac{1}{2 \cdot \varepsilon} \int_{\Omega} ((I_k + \varepsilon \delta I_k)(x, y, T_f) - \tilde{I}_k(x, y))^2 d\Omega \right. \\
&\quad \left. - \frac{1}{2} \int_{\Omega} (I_k(x, y, T_f) - \tilde{I}_k(x, y))^2 d\Omega \right. \\
&\quad \left. + \int_Q \lambda_k [((I_k + \varepsilon \delta I_k)_t - D(I_k + \varepsilon \delta I_k)) - ((I_k)_t - D(I_k))] dQ \right) \\
&= \int_{\Omega} (I_k(x, y, T_f) - \tilde{I}_k(x, y)) \delta I_k(x, y, T_f) d\Omega \\
&\quad + \int_Q \lambda_k (\delta I_k(x, y, T_f))_t dQ - \int_Q \lambda_k \sum_{p \in \eta} \sigma_p \frac{\partial^{|p|} (\delta I_k)}{\partial p} dQ + O(\varepsilon),
\end{aligned}$$

and δI_k should satisfy

$$\begin{aligned}
\delta I_k &= 0 \quad \text{on } \Gamma, \\
\delta I_k &= 0 \quad \text{in } \Omega.
\end{aligned}$$

Due to the boundary and initial conditions of I_k , if we assume that $\lambda_k = 0$ on Γ , then upon integration by parts (with respect to t in $(I_k)_t$ and spatial variables in $\partial^{|p|}(\delta I_k)$), we have

$$\begin{aligned}
L_{I_k}(I, u, \lambda) &= \int_{\Omega} (I_k(x, y, T_f) - \tilde{I}_k(x, y)) \delta I_k(x, y, T_f) d\Omega + \int_{\Omega} (\lambda_k \cdot \delta I_k)(x, y, T_f) d\Omega \\
&\quad - \int_Q (\lambda_k)_t \delta I_k dQ - \int_Q \sum_{p \in \eta} (-1)^{|p|} \frac{\partial^{|p|} (\sigma_p(I_k) \lambda_k)}{\partial p} \cdot \delta I_k dQ
\end{aligned}$$

$$= \int_Q [(\lambda_k + I_k - \tilde{I}_k)\delta(t - T_f) - (\lambda_k)_t - \sum_{p \in \eta} (-1)^{|p|} \frac{\partial^{|p|}(\sigma_p \lambda_k)}{\partial p}] \delta I_k dQ + O(\varepsilon),$$

where $\delta(\cdot)$ is the Dirac function. Finally, by considering the second optimality condition $L_u(I, u, \lambda) = 0$, our argument yields the following system as the adjoint equation for λ_k :

$$\begin{cases} \frac{\partial \lambda_k}{\partial t} + \sum_{p \in \eta} (-1)^{|p|} (\sigma_p \lambda_k)_p = 0, & \text{in } Q, \\ \lambda_k = 0, & \text{on } \Gamma, \\ \lambda_k(x, y, T_f) = \tilde{I}_k - I_k(T_f), & \text{in } \Omega, \end{cases} \quad (31)$$

for $k = 1, \dots, K$.

Now by perturbing u_i in L , we get $L_{u_i}(I, u, \lambda)$

$$\begin{aligned} L_{u_i}(I, u, \lambda) &= \lim_{\varepsilon \rightarrow 0} \frac{L(\dots, u_i + \varepsilon \delta u_i, \dots) - L(\dots, u_i, \dots)}{\varepsilon} \\ &= \lim_{\varepsilon \rightarrow 0} \left(\frac{1}{2\varepsilon} \alpha_i \int_0^T (u_i + \varepsilon \delta u_i)^2(t) dt - \frac{1}{2} \alpha_i \int_0^T u_i^2(t) dt \right. \\ &\quad \left. - \sum_{k=1}^K \int_Q \lambda_k((u_i + \varepsilon \delta u_i)(t) - u_i^t(t)) \operatorname{diff}(I) dQ \right) \\ &= \int_0^T (\alpha_i u_i \delta u_i)(t) dt - \int_0^T \left(\sum_{k=1}^K \int_\Omega \lambda_k \operatorname{diff}(I_k) d\Omega \right) \delta u_i dt. \end{aligned}$$

Therefore, we get

$$L_{u_i}(I, u, \lambda) = \alpha_i u_i - \sum_{k=1}^K \int_\Omega \lambda_k \operatorname{diff}_i(I_k) d\Omega, \quad i = 0, \dots, 6,$$

where the function λ_k is the solution to (31). Now we get back to the Lagrangian (22) of our model, we have

$$L_{u_i}(I, u, \lambda, \mu_a) = \alpha_i u_i - \sum_{k=1}^K \int_\Omega \lambda_k \operatorname{diff}_i(I_k) d\Omega - \mu_{a,i}, \quad i = 0, \dots, 6,$$

From Lemma 1, we have $\frac{DJ}{Du} = L_u(I, u, \lambda)$. Therefore we arrive at

$$\frac{DJ}{Du_i} = \alpha_i u_i - \sum_{k=1}^K \int_\Omega \lambda_k \operatorname{diff}_i(I_k) d\Omega - \mu_{a,i}, \quad i = 0, \dots, 6. \quad (32)$$

□

In summary, we have the Gateaux derivative of the object function J with respect to the control u ,

$$\frac{DJ}{Du_i} = \alpha_i u_i - \sum_{k=1}^K \int_{\Omega} \lambda_k \text{diff}_i(I_k) d\Omega - \mu_{a,i}, \quad i = 0, \dots, 6, \quad (33)$$

where the adjoint state λ is the solution to the PDE bellow:

$$\begin{cases} \frac{\partial \lambda_k}{\partial t} + \sum_{p \in \eta} (-1)^{|p|} (\sigma_p \lambda_k)_p = 0, & \text{in } Q, \\ \lambda_k = 0, & \text{on } \Gamma, \\ \lambda_k(x, y, T_f) = \tilde{I}_k - I_k(T_f), & \text{in } \Omega. \end{cases} \quad (34)$$

For more details and a more mathematically precise explanation, the interested reader is referred to [36, 20]. Now that we have extracted enough information from the optimality conditions, we can use numerical methods to determine $\bar{\mathbf{u}}$ for (11).

Now, we need to find the initial $\mathbf{u}(t)$ as the starter of the numerical method. At each time step, $\frac{\partial I_k(t)}{\partial t}$ is expected to be $\frac{\tilde{I}_k - I_k(t)}{T_f - t}$ so that I_k tends to the expected output \tilde{I}_k . On the other hand, we have $\frac{\partial I_k(t)}{\partial t} = L(I_k, \mathbf{u})$. Next, we aim to find $\mathbf{u}(t)$ to minimize

$$\sum_{k=1}^K \int_{\Omega} \left(L(I_k, \mathbf{u}) - \frac{\partial I_k(t)}{\partial t} \right)^2 d\Omega = \sum_{k=1}^K \int_{\Omega} [p_k(t)^T \mathbf{u}(t) - d_k(t)]^2 d\Omega.$$

Therefore, the initial $\mathbf{u}(t)$ can be obtained by solving the following system:

$$P(t)\mathbf{u}(t) = \mathbf{d}(t), \quad (35)$$

where we have $P(t) = \sum_{k=1}^K \int_{\Omega} p_k(t) p_k(t)^T d\Omega$ and $\mathbf{d}(t) = \sum_{k=1}^K \int_{\Omega} p_k(t) d_k(t)^T d\Omega$, in which $p_k(t) = \text{diff}(I_k)$ and $d_k(t) = \frac{\tilde{I}_k - I_k(t)}{T_f - t}$.

Finally, with the help of the above $\mathbf{u}(t)$ and by considering the Gateaux derivative (33), we can perform Conjugate Gradient (CG) method [32, 34] to solve the optimal control problem (11).

Our PDE framework is summarized in Algorithm 1. Once the coefficients are learned, equation (4) is completed and can be applied on test images, where the noisy image f is the input as the initial value and the output $I(T_f)$ is the desired denoised image.

Algorithm 1 The framework to learn PDEs for image restoration

Input: Training image pairs (f_k, \tilde{I}_k) , $k = 1, \dots, K$, T_f .

- 1: Initialize $\mathbf{u}(t)$, $t \in [0, T_f)$, by solving (35).
 - 2: **while** not converged **do**
 - 3: Compute $\frac{\partial J}{\partial u_i}$, $i = 0, \dots, 6$ using (33).
 - 4: Perform conjugate gradient method using $\frac{\partial J}{\partial u_i}$ [32, 34].
 - 5: **end while**
 - 6: **return** Coefficient functions $\mathbf{u}(t)$, $t \in [0, T_f)$.
 - 7: Solve (4) using the result from previous step.
-

3 Experimental Results

In this section, the application of our proposed PDE for image denoising is demonstrated. The experiments are performed on gray-scale images. We compare the results obtained from our PDE with the evolutionary previous works, PM second-order PDE [29], YK fourth-order PDE [42], second-order L-PDE [22], Zhang [47], and Siddig's [33] fourth-order PDE approaches for image denoising. For selecting the parameters in PM and YK method, we chose the values that produce the most appealing results and the best result is chosen as the output of the method. For all experiments, the grid sizes $h = 1$ and $T_f = 5$ are considered. All our experiments are performed in MATLAB R2018b on an Intel(R) Core(TM) i5, 2.40 GHz, 8 GB RAM laptop.

For a quantitative comparison between the previous methods and our scheme, the peak signal-to-noise ratio (PSNR)[44] and structural similarity index (SSIM)[37] of the processed images are considered. It should be noted that PSNR is most easily defined via the mean squared error (MSE). Given a noise-free $m \times n$ monochrome image I and its noisy approximation f , MSE is defined as:

$$MSE = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} (I(i, j) - f(i, j))^2.$$

Then PSNR is obtained as follows:

$$\begin{aligned} PSNR &= 10 \cdot \log_{10} \left(\frac{\max^2 I}{\sqrt{MSE}} \right) \\ &= 20 \cdot \log_{10}(\max I) - 10 \cdot \log_{10}(MSE). \end{aligned} \tag{36}$$

Here, $\max I$ is the maximum possible pixel value of the image. It can easily be seen that, the closer an image is to its ground truth, the higher its PSNR is. Given similar assumptions, SSIM is defined as

$$SSIM = \frac{(2\bar{I}\bar{f} + C_1)(2\sigma_{If} + C_2)}{(\mu_I^2 + \mu_f^2 + C_1)(\sigma_I^2 + \sigma_f^2 + C_2)},$$

in which \bar{I} and \bar{f} are the mean values of I and f , σ_I and σ_f are the standard deviation of them, σ_I^2 and σ_f^2 are their variances, and σ_{If} is the variance of I and f . For stabilizing the division with the weak denominator, $C_1 = (K_1L)^2$ and $C_2 = (K_2L)^2$ are two variables and L is the dynamic range of pixel values, $K_1 = 0.01$ and $K_2 = 0$. Like PSNR, the closer an image is to its ground truth, the closer the SSIM will be to one.

For choosing images in our experiment, first, we used a group of images containing 50 clear images and their noisy version, with additive white zero mean, Gaussian white noise with the variance of 0.01, in various sizes, categorized in ten different classes including animals, sea, personal picture, city view, nature images, and so on. These images are gathered from SIPI image database [40], Berkeley image database [3], and Matlab's image library. Four pictures in each class were used for training the PDE, and one was used as the test image and the same test image was used as the initial image for the other denoising methods.

In Figures 1 and 2, the performance of our proposed method and the other five denoising models are shown. First, "Lena" picture with size 512×512 , Figure 1(a) is the original clean image, Figure 1(b) is a noisy version of it with additive white noise followed by Figures 1(c-g) showing the denoised images by PM, YK, 2nd L-PDE, Zhang's's PDE, and Siddig's model, respectively, and Figure 1(h) is the denoised image with our proposed method. It can be seen that compared to the other five methods, our fourth-order PDE demonstrates better results in preserving edges while decreasing the noise of the image. In Figure 2, the "City" image with size 850×480 is displayed followed by Figure 2(b), its noisy version, and on Figures 2(c-g), the enhanced form of it by the Pm, YK, 2nd order L-PDE, Zhang's model, and Siddig's approach is shown, and in Figure 2(h), the result of our denoising model is illustrated. It is vivid that the recommended PDE reduces noise without blurring the image features or leaving any speckles.

The PSNR and SSIM values collected from denoising ten images with additive white Gaussian noise with all the mentioned models are presented in Table 1. Comparing the values obtained from discussed methods in each image shows that our proposed scheme has the best performance among the other methods, which is indicated by the higher PSNR and SSIM values.

In addition, to better investigate the performance of our suggested method on more complex noises, the second group consisting of 50 images, in five separate classes with unknown noise, meaning a combination of zero-mean Gaussian white noise, Poisson noise, and the salt & pepper noise ($d = 0.1$) is added to the images, such as aerial and texture images are chosen. This group of images is also gathered from SIPI and Berkeley image databases and pictures are taken by a Canon EOS 750D camera. Table 2 represents the results of performing all the methods on our five image classes, which

can be seen that our suggested model results in better outcomes. In training the L-PDEs in [22] and our model for denoising the images with unknown noise, we used more training images to learn the coefficients due to the more complex noise in each image. In Figures 3 and 4, the noisy test image “Aerial” and “Tile” are shown with their denoised versions and it is observed that our model is superior in decreasing noise while preserving the image features.

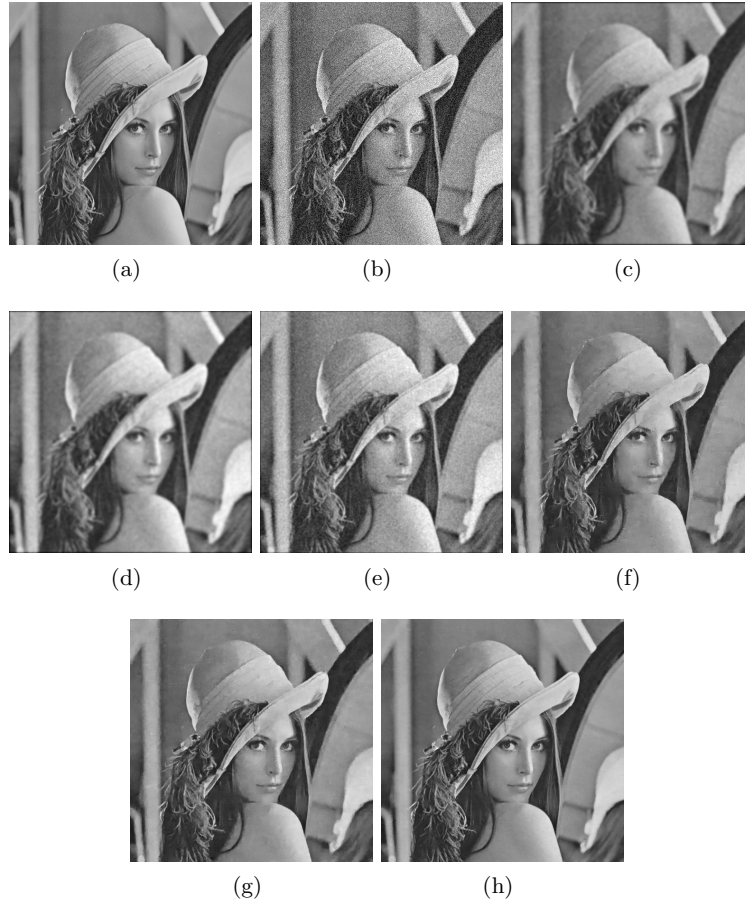


Figure 1: Comparison of proposed scheme with five methods on “Lena” image with added white Gaussian noise. (a) Original image; (b) Noisy image; (c) PM second-order PDE; (d) YK fourth-order PDE; (e) second-order L-PDE; (f) Zhang’s fourth-order PDE; (g) Siddig’s model; (h) proposed fourth-order PDE

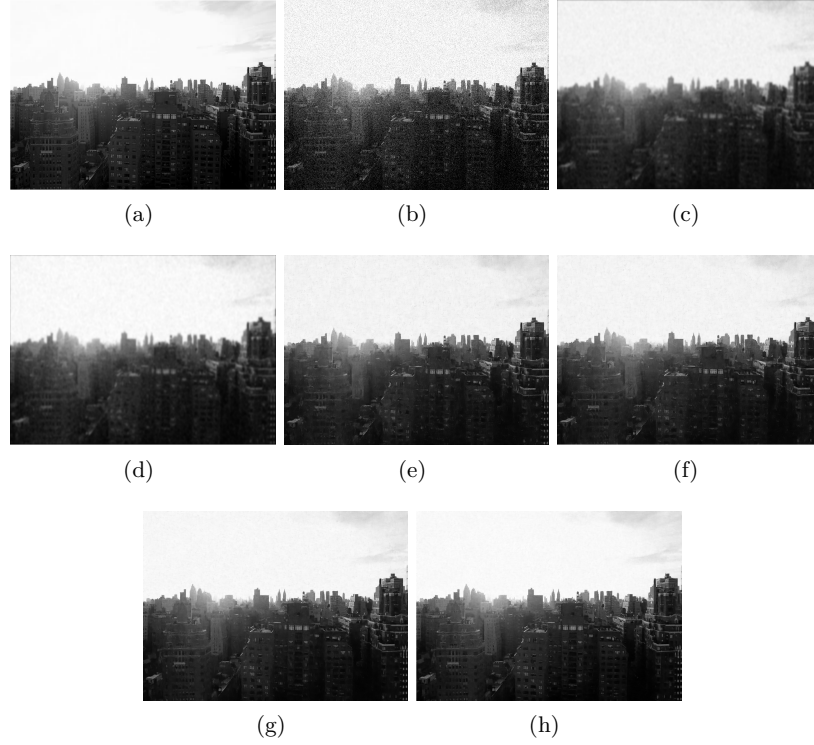


Figure 2: Comparison of proposed scheme with five methods on “City” image with added white Gaussian noise. (a) Original image; (b) Noisy image; (c) PM second-order PDE; (d) YK fourth-order PDE; (e) second-order L-PDE; (f) Zhang’s fourth-order PDE; (g) Siddig’s model; (h) proposed fourth-order PDE

In Figure 3, a noisy aerial image with unknown noise is shown and its denoised forms are Figures 3(b-g). Even with the complex unknown noise, the performance of our proposed quadratic model is better than the other competitors. In Figure 4, a noisy tile image is presented. As expected from the PSNR and SSIM values in Table 2, the recommended learning-based PDE still yields better results in comparison to the other contestants.

Finally, we compare the average PSNR values resulted from all the models, in Figure 5. Comparing the average values obtained from each method on both image groups with Gaussian and unknown noise, we can see that the average values gained from our proposed PDE are larger than the other approaches on both types of noises, which indicates that our scheme outperforms other models in both image categories.

Table 1: PSNR and SSIM results for ten test images with additive white Gaussian noise.

	PM	YK	2nd L-PDE	Zhang	Siddig	Proposed
Lena	19.4755	20.8096	23.8938	27.7531	28.5738	30.4873
	0.7527	0.8031	0.8093	0.8069	0.8421	0.9184
City	20.3853	20.7052	24.2705	28.2597	29.6089	29.1105
	0.6908	0.7288	0.8041	0.8472	0.8133	0.8563
Text	18.0646	21.9834	23.9050	28.0671	28.5636	31.2598
	0.7093	0.7313	0.8527	0.8530	0.8600	0.8697
Woman	19.8594	20.7346	23.0504	29.4682	30.1833	30.4666
	0.6551	0.6995	0.8211	0.8873	0.8237	0.9030
Eiffel	19.1024	21.4389	24.1924	28.0046	30.1453	30.2713
	0.6694	0.6890	0.7806	0.8074	0.8497	0.8901
Sea	20.2217	21.0962	23.9878	27.2673	27.5078	31.0694
	0.7589	0.7854	0.8030	0.8364	0.7890	0.8751
Nightcity	19.8372	21.9698	25.2569	28.5490	28.6472	29.3441
	0.7048	0.7589	0.8497	0.7573	0.8628	0.8638
Nature	20.2255	22.3523	24.6548	29.7069	29.6567	30.4155
	0.6508	0.7001	0.7945	0.7886	0.8349	0.9153
Flowers	19.3536	20.9230	25.2612	28.9910	29.0531	29.9513
	0.6461	0.6934	0.7914	0.8199	0.8276	0.8785
Trees	20.3819	22.1760	25.8949	28.6375	28.8983	30.5744
	0.6835	0.7412	0.7364	0.8536	0.8619	0.9143

Table 2: PSNR and SSIM results for five test images with real unknown noise.

	PM	YK	2nd L-PDE	Zhang	Siddig	Proposed
Truck	16.1163	17.2961	18.6548	19.9679	21.0791	22.5288
	0.7001	0.7058	0.7413	0.7538	0.8103	0.8526
Aerial	15.9888	18.1222	18.9908	19.5223	20.9317	22.9611
	0.6846	0.7140	0.7851	0.7987	0.8159	0.8794
Personal	16.6055	18.3664	19.0258	20.6046	21.6548	23.7669
	0.6422	0.6988	0.7423	0.7642	0.7941	0.8682
Texture	15.4097	17.0267	18.9516	19.5583	20.9513	22.0304
	0.7085	0.7631	0.7970	0.8155	0.8289	0.8437
Tile	16.2849	17.9806	18.9823	20.7140	21.1675	23.6656
	0.6827	0.7125	0.7681	0.8155	0.8468	0.8549

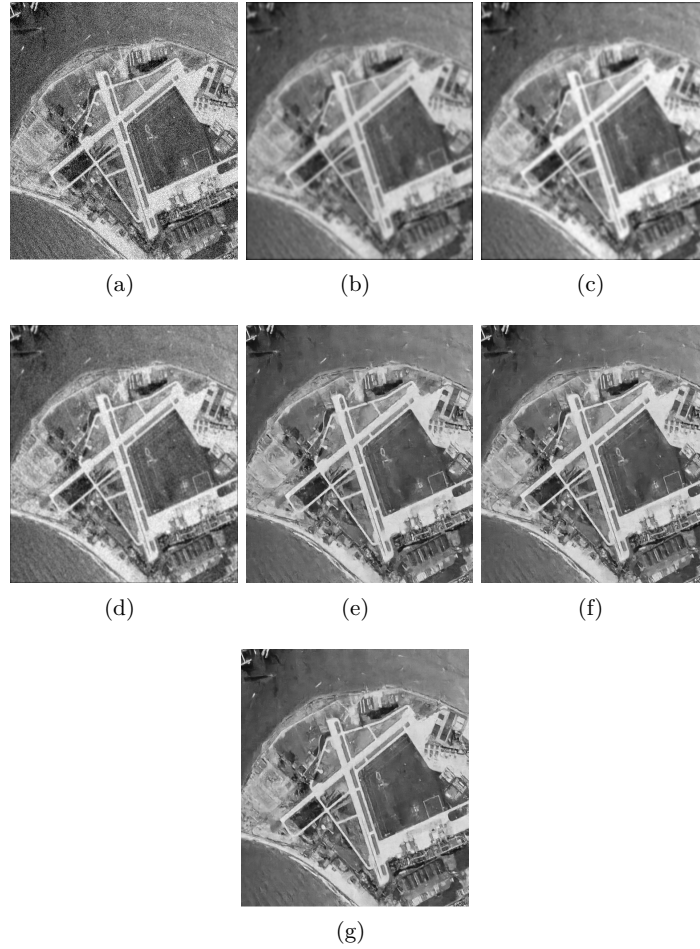


Figure 3: Comparison of proposed scheme with five methods on “Aerial” image with real unknown noise. (a) Noisy image; (b) PM second-order PDE; (c) YK fourth-order PDE; (d) second-order L-PDE; (e) Zhang’s fourth-order PDE; (f) Siddig’s model; (g) proposed fourth-order PDE

4 Conclusion

In this article, we proposed a fourth-order PDE for image denoising with trainable coefficients asserted by an optimal control problem. The experimental results show improvements in performance in comparison to previous approaches in the field of PDE-based methods for image denoising. Compared

to previous methods, our scheme results in better outcomes, clearer image with higher PSNR and SSIM. For future studies, we want to strengthen our approach in the following areas. First, design a more general equation to adapt better to different problems, more precisely, having an optimal control problem with controls depending on spatial variables as well. Second, to introduce a region-based PDE with trainable coefficients, which leads to having two controls for each region. Finally, finding a more suitable numerical method for solving the optimal control problem to have a faster and more accurate numerical result.

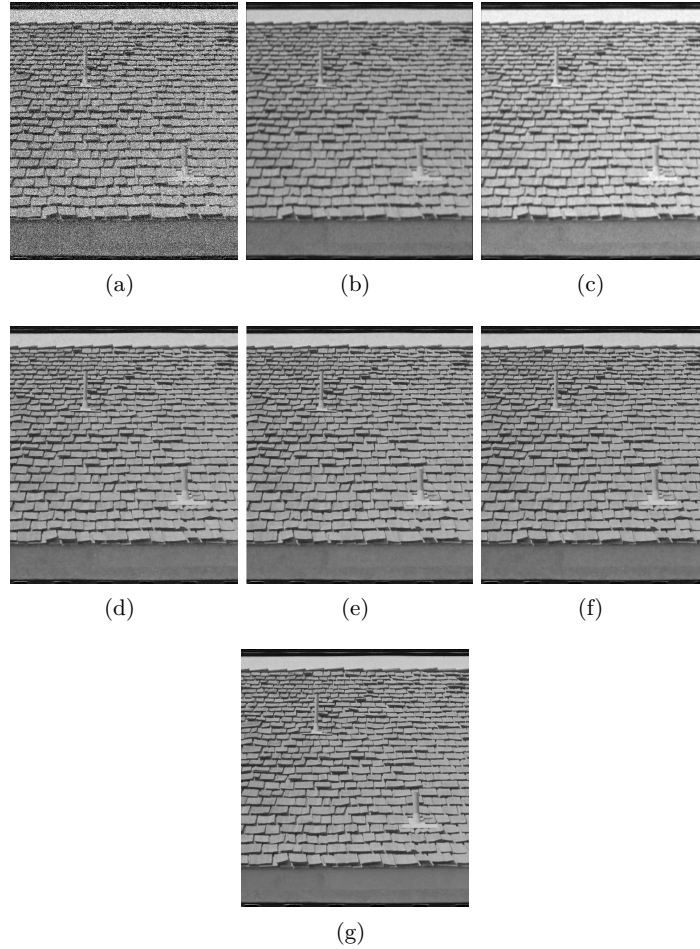


Figure 4: Comparison of proposed scheme with five methods on “Tile” image with real unknown noise. (a) Noisy image; (b) PM second-order PDE; (c) YK fourth-order PDE; (d) second-order L-PDE; (e) Zhang’s fourth-order PDE; (f) Siddig’s model; (g) proposed fourth-order PDE

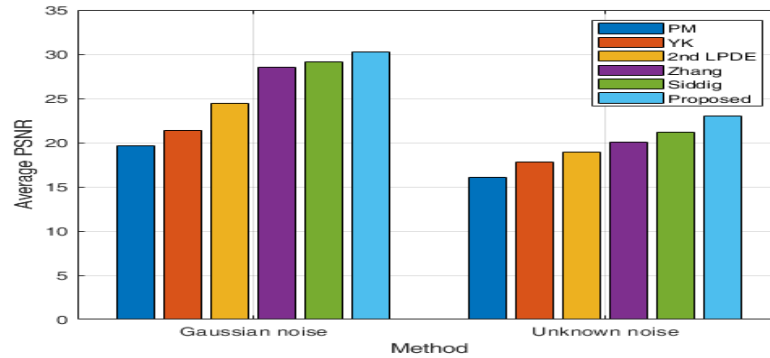


Figure 5: Comparison of average PSNR values in four methods

Acknowledgements

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

References

1. Awate, S.P., and Whitaker, R.T. *Unsupervised, information-theoretic, adaptive image filtering for image restoration*. IEEE Transactions on pattern analysis and machine intelligence 28, 3 (2006), 364–376.
2. Barash, D. *Fundamental relationship between bilateral filtering, adaptive smoothing, and the nonlinear diffusion equation*. IEEE Transactions on Pattern Analysis and Machine Intelligence 24, 6 (2002), 844–847.
3. Besser, H. *Visual access to visual images: the UC berkeley image database project*. Library Trends. 38, 1990.
4. Buades, A., Coll, B., and Morel, J.-M. *A review of image denoising algorithms, with a new one*. Multiscale Model. Simul. 4 (2005), 490–530.
5. Chen, Q., Montesinos, P., Sun, Q.S., and Xia, D.S. *Ramp preserving Perona–Malik model*. Signal Processing, 90 (6) (2010), 1963–1975.
6. Dabov, K., Foi, A., Katkovnik, V., and Egiazarian, K. *Image denoising with block-matching and 3d filtering*. In Image Processing: Algorithms

- and Systems, Neural Networks, and Machine Learning (2006), vol. 6064, International Society for Optics and Photonics, p. 606414.
7. Danielyan, A., Katkovnik, V., and Egiazarian, K. *BM3D frames and variational image deblurring*. IEEE Trans. Image Process. 21(4) (2012), 1715–1728.
 8. Dautov, Ç.P., and Özerdem, M.S. *Wavelet transform and signal denoising using wavelet method*. 26th Signal Processing and Communications Applications Conference (SIU), Izmir, Turkey, 2018, pp. 1-4.
 9. Didas, S., Weickert, J., and Burgeth, B. *Properties of higher order non-linear diffusion filtering*. J. Math. Imaging Vision 35 (3) (2009), 208–226.
 10. Elad, M. *On the origin of the bilateral filter and ways to improve it*. IEEE Trans. Image Process. 11 (10) (2002), 1141–1151.
 11. Gabor, D. *Information theory in electron microscopy*. Lab Invest. 14 (1965), 801–807.
 12. Greer, J.B., and Bertozzi, A.L. *Traveling wave solutions of fourth order PDEs for image processing*. SIAM J. Math. Anal. 36 (1) (2004), 38–68.
 13. Hajiaboli, M.R. *An anisotropic fourth-order diffusion filter for image noise removal*. Int. J. Comput. Vis. 92 (2) (2011), 177–191.
 14. Hinze, M., Pinnau, R., Ulbrich, M., and Ulbrich, S. *Optimization with PDE constraints, Mathematical Modelling: Theory and Applications*. 23. Springer, New York, 2009.
 15. Jain, A.K. *Partial differential equations and finite-difference methods in image processing, part 1: Image representation*. J. Optim. Theory Appl. 23 (1) (1977), 65–91.
 16. Kichenassamy, S. *The Perona-Malik paradox*. SIAM J. Appl. Math. 57 (5) (1997), 1328–1342.
 17. Koenderink, J.J. *The structure of images*. Biol. Cybernet. 50 (5) (1984), 363–370.
 18. Li, M. *An improved non-local filter for image denoising*. 1–4.
 19. Lin, Z., Zhang, W., and Tang, X. *Learning partial differential equations for computer vision*. Peking Univ., Chin. Univ. of Hong Kong 2008.
 20. Lions, J.L. *Optimal control of systems governed by partial differential equations* Translated from the French by S. K. Mitter. Die Grundlehren der mathematischen Wissenschaften, Band 170 Springer-Verlag, New York-Berlin 1971.

21. Liu, Q., Yao, Z., and Ke, Y. *Entropy solutions for a fourth-order nonlinear degenerate problem for noise removal*. Nonlinear Anal. 67 (6) (2007), 1908–1918.
22. Liu, R., Lin, Z., Zhang, W., and Su, Z. *Learning PDEs for image restoration via optimal control*. Daniilidis K., Maragos P., Paragios N. (eds) Computer Vision – ECCV 2010. ECCV 2010. 115–128, Lecture Notes in Computer Science, vol 6311. Springer, Berlin, Heidelberg.
23. Liu, R., Zhong, G., Cao, J., Lin, Z., Shan, S., and Luo, Z. *Learning to diffuse: A new perspective to design pdes for visual analysis*. IEEE transactions on pattern analysis and machine intelligence 38 (12) (2016), 2457–2471.
24. Lysaker, M., Lundervold, A., and Tai, X.-C. *Noise removal using fourth-order partial differential equation with applications to medical magnetic resonance images in space and time*. IEEE Transactions on image processing 12 (12) (2003), 1579–1590.
25. Malfait, M., and Roose, D. *Wavelet-based image denoising using a Markov random field a priori model*. IEEE Transactions on image processing 6 (4) (1997), 549–565.
26. Milanfar, P. *A tour of modern image filtering: New insights and methods, both practical and theoretical*. IEEE signal processing magazine 30 (1) (2012), 106–128.
27. Mo, H., and Li, H. *Image differential invariants*. arXiv:1911.05327 (2019).
28. Own, C., Tsai, H., Yu, P., and Lee, Y. *Adaptive type-2 fuzzy median filter design for removal of impulse noise*. NSIP 2005. Abstracts. IEEE-Eurasip Nonlinear Signal and Image Processing, 2005., Sapporo, Japan, 2005, Imaging Sci. J. 54 (1) (2006), 3–18.
29. Perona, P., and Malik, J. *Scale-space and edge detection using anisotropic diffusion*. IEEE Transactions on pattern analysis and machine intelligence 12 (7) (1990), 629–639.
30. Philip, P. *Optimal control of partial differential equations*. Lecture Notes, Ludwig-Maximilians-Universität, Germany (2009).
31. Rudin, L.I., Osher, S., and Fatemi, E. *Nonlinear total variation based noise removal algorithms*. Experimental mathematics: computational issues in nonlinear science (Los Alamos, NM, 1991). Phys. D 60 (1-4) (1992), 259–268.
32. Shewchuk, J.R. *An introduction to the conjugate gradient method without the agonizing pain*. School of computer science. Carnegie Mellon University, Pittsburgh, PA 15213 (1994), 10.

33. Siddig, A., Guo, Z., Zhou, Z., and Wu, B. *An image denoising model based on a fourth-order nonlinear partial differential equation*. Comput. Math. Appl. 76 (5) (2018), 1056–1074.
34. Stoer, J., and Bulirsch, R. *Introduction to numerical analysis*, vol. 12. Springer Science & Business Media, 2013.
35. Tomasi, C., and Manduchi, R. *Bilateral filtering for gray and color images*. Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271), Bombay, India, 1998, 839–846.
36. Tröltzsch, F. *Optimal control of partial differential equations: theory, methods, and applications*, Translated from the 2005 German original by Jürgen Sprekels. Graduate Studies in Mathematics, 112. American Mathematical Society, Providence, RI, 2010.
37. Wang, Z., Bovik, A.C., Sheikh, H.R., and Simoncelli, E.P. *Image quality assessment: from error visibility to structural similarity*. IEEE transactions on image processing 13 (4) (2004), 600–612.
38. Wang, Y., Chen, W., Zhou, S., Yu, T., and Zhang, Y. *Mtv: modified total variation model for image noise removal*. IEE Electronics Letters, 47 (10) (2011), 592–594.
39. Wang, Y., Guo, J., Chen, W., and Zhang, W. *Image denoising using modified Perona–Malik model based on directional Laplacian*. Signal Processing 93 (9) (2013), 2548–2558.
40. Weber, A.G. *The USC-SIPI image database version 5*. USC-SIPI Report 315, 1 (1997).
41. Witkin, A.P. *Scale-space filtering*. Readings in Computer Vision (1987) 329–332.
42. You, Y.-L., and Kaveh, M. *Fourth-order partial differential equations for noise removal*. IEEE Transactions on Image Processing 9 (10) (2000), 1723–1730.
43. You, Y.-L., Xu, W., Tannenbaum, A., and Kaveh, M. *Behavioral analysis of anisotropic diffusion in image processing*. IEEE Transactions on Image Processing 5 (11) (1996), 1539–1553.
44. Zeng, W., and Lu, X. *A robust variational approach to super-resolution with nonlocal tv regularisation term*. Imaging Sci. J. 61 (2) (2013), 268–278.
45. Zeng, W., Lu, X., and Tan, X. *A local structural adaptive partial differential equation for image denoising*. Multimed. Tools Appl. 74 (3) (2015), 743–757.

46. Zhang, M., and Desrosiers, C. *Image denoising based on sparse representation and gradient histogram*. IET Image Processing 11 (1) (2016), 54–63.
47. Zhang, X., and Ye, W. *An adaptive fourth-order partial differential equation for image denoising*. Comput. Math. Appl. 74 (10) (2017), 2529–2545.



Exponentially fitted tension spline method for singularly perturbed differential difference equations

M.M. Woldaregay* and G.F. Duressa

Abstract

In this article, singularly perturbed differential difference equations having delay and advance in the reaction terms are considered. The highest-order derivative term of the equation is multiplied by a perturbation parameter ε taking arbitrary values in the interval $(0, 1]$. For the small value of ε , the solution of the equation exhibits a boundary layer on the left or right side of the domain depending on the sign of the convective term. The terms with the shifts are approximated by using the Taylor series approximation. The resulting singularly perturbed boundary value problem is solved using an exponentially fitted tension spline method. The stability and uniform convergence of the scheme are discussed and proved. Numerical examples are considered for validating the theoretical analysis of the scheme. The developed scheme gives an accurate result with linear order uniform convergence.

AMS subject classifications (2020): 65L11; 65L03; 65L70.

Keywords: Differential difference; Exponentially fitted; Singularly perturbed problem; Tension spline; Uniform convergence.

1 Introduction

A large number of mathematical models have appeared in different areas of science and engineering that take into account not just the present state of a physical system but also its past history. These models are described by

*Corresponding author

Received 9 January 2021; revised 18 March 2021; accepted 21 April 2021

Mesfin Mekuria Woldaregay

Department of Applied Mathematics, Adama Science and Technology University, Adama, Ethiopia. e-mail: mesfin.mekuria@astu.edu.et

Gemechis File Duressa

Department of Mathematics, Jimma University, Jimma, Ethiopia. e-mail: gam-meef@gmail.com

certain classes of functional differential equations often called delay differential equations or differential difference equations (DDEs). The class of DDEs with characteristics of delay/advance and singularly perturbed behavior is known as singularly perturbed differential difference equations (SPDDEs). The DDEs with delay or advance term play an important role in modeling many real life phenomena in bioscience, control theory, economics, and engineering [5]. Some applications are the mathematical modeling of population dynamics and epidemiology [19], physiological kinetics [4], blood cell production [25], and so on.

In SPDDEs model process, the evaluation not only depends on the current state of the system but also includes the past history. A number of model problems in science and engineering take the forms of SPDDEs [32]; we list a few of them: neuron variability model in computational neuroscience, optimal control theory problems, and model describing the motion of sunflower.

For the perturbation parameter, when ε tends to zero, the smoothness of the solution of the singularly perturbed problems deteriorates and it forms boundary layer; see [6, 27]. In the case where ε is very small, standard numerical methods such as FDM, FEM, and collocation method lead to oscillations in the computed solutions. To handle the oscillation, a large number of mesh points are required, which is not practical; see [32].

The solution methods of SPDDEs have received great attention in recent years because of their wide applications. It is of theoretical and practical interest to consider numerical methods for such problems [40]. Adilaxmi et al. [1, 2] proposed the exponentially fitted nonstandard FDM and integration method using a nonpolynomial interpolating function. In articles [21, 22, 23, 24], Lange and Miura developed asymptotic methods for solving a class of SPDDEs. The authors extend the matched asymptotic method initially developed for solving BVPs to obtain an approximate solution for SPDDEs. In articles [9, 11, 12, 13], Kadalbajoo and Sharma developed ε -uniform numerical methods using fitted mesh techniques. Swamy, Phaneendra, and Reddy [35] used the exponentially fitted Galerkin method for treating the problem. The authors in [36, 37] developed a fourth-order FDM with an exponential fitting factor. Melesse, Tiruneh, and Deresechite [26] used the initial value technique to treat the problem. They showed the applicability of the scheme by considering different examples. Ranjan and Prasad [31] used the modified fitted FDM for solving the problem. Sirisha, Phaneendra, and Reddy [34] developed a finite difference scheme using the procedure of domain decomposition. Kumar and Sharma [20] applied the B-spline collocation method to approximate the solution of the SPDDEs. Mohapatra and Natesan [28] applied the fitted mesh FDM using the equidistributed grid technique. In [40], Woldaregay and Duressa developed the exponentially fitted FDM with Richardson extrapolation techniques.

Different authors in [3, 7, 8, 14, 15, 16, 17, 30] have applied the tension spline method for treating singularly perturbed reaction diffusion or convection diffusion problems. To the best of the authors' knowledge, the

exponentially fitted tension spline method has not been developed for treating SPDDEs. Developing uniformly convergent schemes is an active research area [33]. This motivates us to develop an accurate and uniformly convergent scheme using the exponentially fitted tension spline method.

Notations: N denotes the number of mesh interval in the discretization, C denotes a positive constant independent of ε and N , and the norm $\|\cdot\|$ denotes the supremum norm.

2 Continuous problem

Consider a class of SPDDE of the form

$$-\varepsilon u''(x) + a(x)u'(x) + \alpha(x)u(x-\delta) + \omega(x)u(x) + \beta(x)u(x+\eta) = f(x), \quad x \in \Omega, \quad (1)$$

with the interval conditions

$$\begin{aligned} u(x) &= \phi(x), & x &\in [-\delta, 0], \\ u(x) &= \psi(x), & x &\in [1, 1+\eta], \end{aligned} \quad (2)$$

where $\Omega = (0, 1)$, $\varepsilon \in (0, 1]$ is the singular perturbation parameter, and δ and η are delay and advance parameters satisfying $\delta, \eta < \varepsilon$. The functions $a(x), \alpha(x), \omega(x), \beta(x), f(x), \phi(x)$, and $\psi(x)$ are assumed to be sufficiently smooth and bounded for the existence of unique solution. The coefficient functions $\alpha(x), \omega(x)$, and $\beta(x)$ are assumed to satisfy

$$\alpha(x) + \omega(x) + \beta(x) \geq \alpha + \omega + \beta =: \theta > 0, \quad \text{for all } x \in \bar{\Omega},$$

where the constants α, ω , and β are lower bounds of $\alpha(x), \omega(x)$, and $\beta(x)$, respectively.

In the case when $\delta, \eta = 0$, equations (1)–(2) reduce to a singularly perturbed boundary value problem, in which for small ε , it exhibits boundary layer. The layer is maintained for $\delta, \eta \neq 0$ but sufficiently small.

2.1 Properties of the continuous solution

In the case when $\delta, \eta < \varepsilon$, using Taylor's series approximation for the terms with deviating, the argument is appropriate [38]. Using the Taylor series approximation, we approximate

$$\begin{aligned} u(x-\delta) &\approx u(x) - \delta u'(x) + (\delta^2/2)u''(x) + O(\delta^3), \\ u(x+\eta) &\approx u(x) + \eta u'(x) + (\eta^2/2)u''(x) + O(\eta^3). \end{aligned} \quad (3)$$

Replacing (3) in (1) gives

$$-c_\varepsilon u''(x) + p(x)u'(x) + q(x)u(x) = f(x), \quad x \in \Omega, \quad (4)$$

with the boundary conditions

$$u(0) = \phi(0), \quad u(1) = \psi(1), \quad (5)$$

where $c_\varepsilon = \varepsilon^2 - (\delta^2/2)\alpha - (\eta^2/2)\beta$, $p(x) = a(x) - \delta\alpha(x) + \eta\beta(x)$, and $q(x) = \alpha(x) + \beta(x) + \omega(x)$. For small values of δ and η , (1)–(2) and (4)–(5) are asymptotically equivalent, since the difference between these two equations is $O(\delta^3, \eta^3)$. The differential operator L is denoted for the differential equation in (4) and defined as

$$Lu(x) = -c_\varepsilon u''(x) + p(x)u'(x) + q(x)u(x).$$

The problem in (4)–(5) exhibits the regular boundary layer of thickness $O(c_\varepsilon)$, and the position of the boundary layer depends on the conditions: If $p(x) < 0$, then the left boundary layer exists, and if $p(x) > 0$, then the right boundary layer exists. In the case when $p(x), x \in \Omega$, the change sign interior layer will exist [10].

The problem obtained by setting $c_\varepsilon = 0$ in (4)–(5) is called the reduced problem and given as

$$\begin{aligned} p(x)u'_0(x) + q(x)u_0(x) &= f(x), \quad \text{for all } x \in \Omega, \\ u_0(0) &= \phi(0), \quad u_0(1) \neq \psi(1). \end{aligned} \quad (6)$$

For the right boundary layer case, it does not satisfy the right boundary condition, and for the left boundary layer case, it is given as

$$\begin{aligned} p(x)u'_0(x) + q(x)u_0(x) &= f(x), \quad \text{for all } x \in \Omega, \\ u_0(0) &\neq \phi(0), \quad u_0(1) = \psi(1). \end{aligned} \quad (7)$$

For small values of c_ε , the solution $u(x)$ of (4)–(5) is very close to the solution $u_0(x)$ of (6) or (7).

Lemma 1 (The maximum principle [40]). For sufficiently smooth function z on Ω , satisfying $z(0) \geq 0$, $z(1) \geq 0$, and $Lz(x) \geq 0$, for all $x \in \Omega$, implies that $z(x) \geq 0$, for all $x \in \bar{\Omega}$.

Lemma 2 (Stability estimate). The solution $u(x)$ of the continuous equation (4)–(5) satisfies the bounded

$$|u(x)| \leq \theta^{-1} \|f\| + \max\{|\phi(0)|, |\psi(1)|\}. \quad (8)$$

Proof. It is proved by the construction of barrier function and using the maximum principle. Let us define barrier functions $\vartheta^\pm(x)$ as $\vartheta^\pm(x) = \theta^{-1} \|f\| + \max\{|\phi(0)|, |\psi(1)|\} \pm u(x)$. On the boundary points, we obtain

$$\begin{aligned}\vartheta^\pm(0) &= \theta^{-1}\|f\| + \max\{\phi(0), \psi(1)\} \pm u(0) \geq 0, \\ \vartheta^\pm(1) &= \theta^{-1}\|f\| + \max\{\phi(0), \psi(1)\} \pm u(1) \geq 0.\end{aligned}$$

On the differential operator, we have

$$\begin{aligned}L\vartheta^\pm(x) &= -c_\varepsilon \vartheta_\pm''(x) + p(x)\vartheta_\pm'(x) + q(x)\vartheta_\pm(x) \\ &= -c_\varepsilon(0 \pm u''(x)) + p(x)(0 \pm u'(x)) + q(x)(\theta^{-1}\|f\| + \max\{\phi(0), \psi(1)\} \\ &\quad \pm u(x)) \\ &= q(x)(\theta^{-1}\|f\| + \max\{\phi(0), \psi(1)\}) \pm f(x) \\ &\geq 0, \quad \text{since } q(x) \geq \theta > 0.\end{aligned}$$

By using the hypothesis of the maximum principle, we obtain $\vartheta^\pm(x) \geq 0$, for all $x \in \bar{\Omega}$, which implies the required bound. \square

In the next lemma, we obtain a bound for the derivatives of solution.

Lemma 3. Derivatives of the solutions of the problem in (4)–(5) satisfy the bound

$$|u^{(k)}(x)| \leq C(1 + c_\varepsilon^{-k} \exp(\frac{-p^*x}{c_\varepsilon})), \quad x \in \Omega, \quad 0 \leq k \leq 4,$$

for the left boundary layer problem and

$$|u^{(k)}(x)| \leq C(1 + c_\varepsilon^{-k} \exp(\frac{-p^*(1-x)}{c_\varepsilon})), \quad x \in \Omega, \quad 0 \leq k \leq 4,$$

for the right boundary layer problem.

Proof. See [6]. \square

3 Numerical scheme formulation

In this article, an exponentially fitted tension spline method is proposed for solving equations (1)–(2). The exponential fitting factor is used to hinder the influence of the perturbation parameter in the boundary layer region. The theory of the asymptotic method is used for developing the exponential fitting factor. We consider and treat the left and right boundary layer problems separately.

3.1 Exponentially fitted tension spline method

Let $0 = x_0 < x_1 < x_2 < \dots < x_N = 1$ be a uniform partition for $[0, 1]$ such that $x_i = ih$, $i = 0, 1, 2, \dots, N$. A function $S(x, \tau) = S(x)$ is a class of $C^2(\bar{\Omega})$, which interpolates $u(x)$ at the mesh points x_i depending on the parameter τ , reduces to cubic spline in $\bar{\Omega}$ as $\tau \rightarrow 0$, and is termed as a parametric cubic spline function [17, 3]. In $[x_i, x_{i+1}]$, the spline function $S(x)$ satisfies the differential equation

$$S''(x) - \tau S(x) = [S''(x_i) - \tau S(x_i)] \frac{x_{i+1} - x}{h} + [S''(x_{i+1}) - \tau S(x_{i+1})] \frac{x - x_i}{h}, \quad (9)$$

where $S(x_i) = u(x_i)$ and $\tau > 0$ is termed as a cubic spline in compression. Solving the linear second-order differential equation in (9) and determining the arbitrary constants from the interpolation conditions $S(x_{i+1}) = u(x_{i+1})$, $S(x_i) = u(x_i)$, we get

$$S(x) = \frac{h^2}{\lambda^2 \sinh \lambda} \left[M_{i+1} \sinh\left(\frac{\lambda(x - x_i)}{h}\right) + M_i \sinh\left(\frac{\lambda(x_{i+1} - x)}{h}\right) \right] - \frac{h^2}{\lambda^2} \left[\left(M_{i+1} - \frac{\lambda^2}{h^2} u(x_{i+1})\right) \left(\frac{x - x_i}{h}\right) + \left(M_i - \frac{\lambda^2}{h^2} u(x_i)\right) \left(\frac{x_{i+1} - x}{h}\right) \right], \quad (10)$$

where $\lambda = h\tau^{1/2}$ and $M_j = u''(x_j)$ for $j = i \pm 1, i$.

Now, differentiating (10) and letting $x \rightarrow x_i$, on the interval $[x_i, x_{i+1}]$, we obtain

$$S'(x_i^+) = \frac{u(x_{i+1}) - u(x_i)}{h} - \frac{h}{\lambda^2} \left[M_{i+1} \left(1 - \frac{\lambda}{\sinh \lambda}\right) + M_i (\lambda \coth \lambda - 1) \right], \quad (11)$$

and on the interval $[x_{i-1}, x_i]$, we obtain

$$S'(x_i^-) = \frac{u(x_i) - u(x_{i-1})}{h} + \frac{h}{\lambda^2} \left[M_i (\lambda \coth \lambda - 1) + M_{i-1} \left(1 - \frac{\lambda}{\sinh \lambda}\right) \right]. \quad (12)$$

Equating the left- and right-hand derivatives at x_i gives

$$\begin{aligned} & \frac{u(x_i) - u(x_{i-1})}{h} + \frac{h}{\lambda^2} \left[M_i (\lambda \coth \lambda - 1) + M_{i-1} \left(1 - \frac{\lambda}{\sinh \lambda}\right) \right] \\ &= \frac{u(x_{i+1}) - u(x_i)}{h} - \frac{h}{\lambda^2} \left[M_{i+1} \left(1 - \frac{\lambda}{\sinh \lambda}\right) + M_i (\lambda \coth \lambda - 1) \right], \end{aligned} \quad (13)$$

$i = 1, 2, \dots, N - 1.$

Rearranging, we obtain the tridiagonal system

$$\lambda_1 M_{i-1} + 2\lambda_2 M_i + \lambda_1 M_{i+1} = \frac{u(x_{i-1}) - 2u(x_i) + u(x_{i+1}))}{h^2}, \quad i = 1, 2, \dots, N-1, \quad (14)$$

where $\lambda_1 = \frac{1}{\lambda^2}(\frac{\lambda}{\sinh \lambda} - 1)$ and $\lambda_2 = \frac{1}{\lambda^2}(1 - \lambda \coth \lambda)$.

The condition of continuity given in (14) ensures the continuity of the first-order derivatives of the spline $S(x)$ at interior nodes.

Now, substituting $-c_\varepsilon M_j = f(x_j) - p(x_j)u'(x_j) - q(x_j)u(x_j)$ for $j = i-1, i$ and $i+1$, we obtain

$$\begin{aligned} L^h u(x_i) \equiv & -\frac{c_\varepsilon}{h^2} [u(x_{i-1}) - 2u(x_i) + u(x_{i+1}))] + \lambda_1 [p(x_{i-1})u'(x_{i-1}) \\ & + q(x_{i-1})u(x_{i-1})) + 2\lambda_2 [p(x_i)u'(x_i) + q(x_i)u(x_i)] \\ & + \lambda_1 [p(x_{i+1})u'(x_{i+1}) + q(x_{i+1})u(x_{i+1}))] \\ & = \lambda_1 f(x_{i-1}) + 2\lambda_2 f(x_i) + \lambda_1 f(x_{i+1}) + T_1(h), \quad i = 1, 2, \dots, N-1, \end{aligned} \quad (15)$$

where $T_1(h)$ is the truncation error in the above discretization, one can see the detail in [3], and it is given by

$$T_1(h) = \frac{h^4}{3}(-2\lambda_1 + \lambda_2)p(x_i)u'''(\zeta_i) + \frac{h^4}{12}(1 - 12\lambda_1)p(x_i)c_\varepsilon u^{(4)}(\zeta_i) + O(h^6), \quad (16)$$

for any choice of λ_1 and λ_2 whose sum is $1/2$, except $\lambda_1 = 1/12$ and $\lambda_2 = 5/12$. For the choice $\lambda_1 = 1/12, \lambda_2 = 5/12$, we have

$$T_1(h) = \frac{c_\varepsilon h^6}{240} u^{(6)}(\zeta_i), \quad \zeta_i \in [x_{i-1}, x_{i+1}]. \quad (17)$$

Next, we use the left-shifted, central, and right-shifted finite difference approximation as

$$\begin{aligned} u'(x_{i-1}) &= \frac{-3u(x_{i-1}) + 4u(x_i) - u(x_{i+1}))}{2h} + O(h), \\ u'(x_i) &= \frac{u(x_{i+1}) - u(x_{i-1}))}{2h} + O(h^2), \quad \text{and} \\ u'(x_{i+1}) &= \frac{3u(x_{i+1}) - 4u(x_i) + u(x_{i-1}))}{2h} + O(h). \end{aligned} \quad (18)$$

Substituting (18) in (15) leads to

$$\begin{aligned} Lu_i \equiv & -\frac{c_\varepsilon}{h^2} [u_{i-1} - 2u_i + u_{i+1}] + \lambda_1 [p(x_{i-1})\left(\frac{-3u_{i-1} + 4u_i - u_{i+1}}{2h}\right) + q(x_{i-1})u_{i-1}] \\ & + 2\lambda_2 [p(x_i)\left(\frac{u_{i+1} - u_{i-1}}{2h}\right) + q(x_i)u_i] + \lambda_1 [p(x_{i+1})\left(\frac{3u_{i+1} - 4u_i + u_{i-1}}{2h}\right) \\ & + q(x_{i+1})u_{i+1}] = \lambda_1 f(x_{i-1}) + 2\lambda_2 f(x_i) + \lambda_1 f(x_{i+1}) + T_2(h), \\ & i = 1, 2, \dots, N-1, \end{aligned} \quad (19)$$

where u_i denotes the approximation of $u(x_i)$ in the above discretization and $T_2(h) = O(h) + T_1(h)$.

3.1.1 Case I: Left boundary layer problem

In this case, the boundary layer occurs on the left side of the domain. From the theory of singular perturbations, the zeros-order asymptotic solution of (4)–(5) is given as [29]

$$u(x) = u_0(x) + \frac{p(0)}{p(x)}(\phi(0) - u_0(0)) \exp\left(-\int_0^x \left(\frac{p(x)}{c_\varepsilon(x)} - \frac{q(x)}{p(x)}\right) dx\right) + O(c_\varepsilon). \quad (20)$$

Using the Taylor series about $x = 0$ for $p(x)$ and $q(x)$ and simplifying give

$$u(x) = u_0(x) + (\phi(0) - u_0(0)) \exp(-p(0)x) + O(c_\varepsilon), \quad (21)$$

where u_0 is the solution of the reduced problem. The domain $[0, 1]$ is discretized into N equal number of subintervals, each of length h . Let $0 = x_0 < x_1 < x_2 < \dots < x_N = 1$ be the points such that $x_i = ih$, $i = 0, 1, 2, \dots, N$.

Considering h small enough, the discretized form of (21) becomes

$$u(ih) \simeq u_i = u_0(ih) + (\phi(0) - u_0(0)) \exp(-p(0)(i\rho)), \quad (22)$$

where $\rho = h/c_\varepsilon$ and $h = 1/N$. Similarly, we write

$$\begin{aligned} u_{i+1} &= u_0((i+1)h) + (\phi(0) - u_0(0)) \exp(-p(0)((i+1)\rho)), \\ u_{i-1} &= u_0((i-1)h) + (\phi(0) - u_0(0)) \exp(-p(0)((i-1)\rho)). \end{aligned} \quad (23)$$

In order to handle the influence of the perturbation parameter, the exponentially fitting factor σ_1 is multiplied on the term containing c_ε as

$$\begin{aligned} L_L^h u_i &\equiv -\frac{c_\varepsilon \sigma_1}{h^2} [u_{i-1} - 2u_i + u_{i+1}] + \lambda_1 [p(x_{i-1}) \left(\frac{-3u_{i-1} + 4u_i - u_{i+1}}{2h}\right) \\ &\quad + q(x_{i-1})u_{i-1}] + 2\lambda_2 [p(x_i) \left(\frac{u_{i+1} - u_{i-1}}{2h}\right) + q(x_i)u_i] \\ &\quad + \lambda_1 [p(x_{i+1}) \left(\frac{3u_{i+1} - 4u_i + u_{i-1}}{2h}\right) + q(x_{i+1})u_{i+1}] \\ &= \lambda_1 f(x_{i-1}) + 2\lambda_2 f(x_i) + \lambda_1 f(x_{i+1}) + T_3(h), \quad i = 1, 2, \dots, N-1. \end{aligned} \quad (24)$$

Multiplying both sides of (24) by h , denoting $c_\varepsilon/h = \rho$, and taking the limit as $h \rightarrow 0$, we obtain

$$\begin{aligned}
& - \lim_{h \rightarrow 0} \frac{\sigma_1}{\rho} [u_{i-1} - 2u_i + u_{i+1}] + \lambda_1 \lim_{h \rightarrow 0} [p((i-1)h)(-3u_{i-1} + 4u_i - u_{i+1})] \\
& + 2\lambda_2 \lim_{h \rightarrow 0} [p(ih)(u_{i+1} - u_{i-1})] + \lambda_1 \lim_{h \rightarrow 0} [p((i+1)h)(3u_{i+1} - 4u_i + u_{i-1})] = 0.
\end{aligned} \tag{25}$$

The expression in (25) simplifies to

$$\lim_{h \rightarrow 0} \frac{c_\varepsilon \sigma_1}{\rho} [u_{i-1} - 2u_i + u_{i+1}] = \lim_{h \rightarrow 0} (\lambda_1 + \lambda_2) [p(0)(u_{i+1} - u_{i-1})]. \tag{26}$$

Using the results in (22) and (23), we obtain

$$\begin{aligned}
\lim_{h \rightarrow 0} [u_{i+1} - 2u_i + u_{i-1}] &= (\phi(0) - u_0(0)) \exp(-p(0)i\rho) \\
&\quad \times [\exp(-p(0)\rho) - 2 + \exp(p(0)\rho)], \\
\lim_{h \rightarrow 0} [u_{i+1} - u_{i-1}] &= (\phi(0) - u_0(0)) \exp(-p(0)i\rho) \\
&\quad \times [\exp(-p(0)\rho) - \exp(p(0)\rho)].
\end{aligned} \tag{27}$$

Using the result in (27) into (26), we obtain the exponential fitting factor as

$$\sigma_1 = p(0)\rho(\lambda_1 + \lambda_2) \coth(p(0)\frac{\rho}{2}). \tag{28}$$

Hence, the required finite difference scheme becomes

$$\begin{aligned}
L_L^h u_i &\equiv - \frac{c_\varepsilon \sigma_1}{h^2} [u_{i-1} - 2u_i + u_{i+1}] + \lambda_1 [p(x_{i-1}) (\frac{-3u_{i-1} + 4u_i - u_{i+1}}{2h}) \\
&\quad + q(x_{i-1})u_{i-1}] + 2\lambda_2 [p(x_i) (\frac{u_{i+1} - u_{i-1}}{2h}) + q(x_i)u_i] \\
&\quad + \lambda_1 [p(x_{i+1}) (\frac{3u_{i+1} - 4u_i + u_{i-1}}{2h}) + q(x_{i+1})u_{i+1}] \\
&= \lambda_1 f(x_{i-1}) + 2\lambda_2 f(x_i) + \lambda_1 f(x_{i+1}) + T_3(h), \quad i = 1, 2, \dots, N-1,
\end{aligned} \tag{29}$$

with the boundary conditions $u_0 = \phi(0)$ and $u_N = \psi(1)$. The bound of truncation error $T_3(h)$ is derived in Theorem 1.

3.1.2 Case II: Right boundary layer problem

In this case, the boundary layer is on the right side of the domain. From the theory of singular perturbations, the zeros-order asymptotic solution of (4)–(5) is given as [29]

$$u(x) = u_0(x) + \frac{p(1)}{p(x)} (\psi(1) - u_0(1)) \exp\left(-\int_x^1 \left(\frac{p(x)}{c_\varepsilon} - \frac{q(x)}{p(x)}\right) dx\right) + O(c_\varepsilon). \tag{30}$$

Using the Taylor series about $x = 1$ for $p(x)$ and $q(x)$ and simplifying, we obtain

$$u(x) = u_0(x) + (\psi(1) - u_0(1)) \exp(-p(1)(1-x)) + O(c_\varepsilon), \quad (31)$$

where u_0 is the solution of the reduced problem. The domain $[0, 1]$ is discretized into N equal number of subintervals, each of length h . Let $0 = x_0 < x_1 < x_2 < \dots < x_N = 1$ be the points such that $x_i = ih$, $i = 0, 1, 2, \dots, N$.

Considering h is small enough, the discretized form of (21) becomes

$$u(ih) \simeq u_i = u_0(ih) + (\psi(1) - u_0(1)) \exp(-p(1)(1/c_\varepsilon - i\rho)), \quad (32)$$

where $\rho = h/c_\varepsilon$ and $h = 1/N$. Similarly, we write

$$\begin{aligned} u_{i+1} &= u_0((i+1)h) + (\psi(1) - u_0(1)) \exp(-p(1)(1/c_\varepsilon - (i+1)\rho)), \\ u_{i-1} &= u_0((i-1)h) + (\psi(1) - u_0(1)) \exp(-p(1)(1/c_\varepsilon - (i-1)\rho)). \end{aligned} \quad (33)$$

Using the similar procedure as the left boundary layer case, we obtain the exponential fitting factor as

$$\sigma_2 = p(1)\rho(\lambda_1 + \lambda_2) \coth\left(\frac{p(1)\rho}{2}\right). \quad (34)$$

Hence, the required finite difference scheme becomes

$$\begin{aligned} L_R^h u_i &\equiv -\frac{c_\varepsilon \sigma_2}{h^2} [u_{i-1} - 2u_i + u_{i+1}] + \lambda_1 [p(x_{i-1}) \left(\frac{-3u_{i-1} + 4u_i - u_{i+1}}{2h} \right) \\ &\quad + q(x_{i-1})u_{i-1}] + 2\lambda_2 [p(x_i) \left(\frac{u_{i+1} - u_{i-1}}{2h} \right) + q(x_i)u_i] \\ &\quad + \lambda_1 [p(x_{i+1}) \left(\frac{3u_{i+1} - 4u_i + u_{i-1}}{2h} \right) + q(x_{i+1})u_{i+1}] \\ &= \lambda_1 f(x_{i-1}) + 2\lambda_2 f(x_i) + \lambda_1 f(x_{i+1}) + T_3(h), \quad i = 1, 2, \dots, N-1, \end{aligned} \quad (35)$$

with the boundary conditions $u_0 = \phi(0)$ and $u_N = \psi(1)$.

3.2 Stability and uniform convergence

In this section, we discuss the uniform stability and convergence for the right boundary layer problems. Similarly, one can do it for the left boundary layer case. First, we prove the discrete comparison principle for the scheme in (35) for the existence of the unique discrete solution.

We observe that the nonzero entries of the coefficient matrix of $L_R^h u_i$ are given by

$$\begin{aligned}
a_{i,i-1} &= -\frac{c_\varepsilon \sigma_2}{h^2} - \lambda_1 \left(\frac{3p(x_{i-1})}{2h} - \frac{p(x_{i+1})}{2h} + q(x_{i-1}) \right) - \lambda_2 \frac{p(x_i)}{h}, \\
a_{i,i} &= \frac{2c_\varepsilon \sigma_2}{h^2} - \lambda_1 \left(\frac{2p(x_{i-1})}{h} - \frac{2p(x_{i+1})}{h} \right) + q(x_i), \\
a_{i,i+1} &= -\frac{c_\varepsilon \sigma_2}{h^2} - \lambda_1 \left(\frac{p(x_{i-1})}{2h} + \frac{3p(x_{i+1})}{2h} + q(x_{i+1}) \right) + \lambda_2 \frac{p(x_i)}{h}.
\end{aligned} \tag{36}$$

For each $i = 1, 2, \dots, N-1$ and for arbitrary values of h and c_ε , we have $a_{i,i-1} < 0$, $a_{i,i+1} < 0$, and $a_{i,i} > 0$.

Lemma 4 (Discrete comparison principle). Assume that, for a mesh function u_i , there exists a comparison function v_i such that $L^h u_i \leq L^h v_i$, $i = 1, 2, \dots, N-1$ and if $u_0 \leq v_0$ and $u_N \leq v_N$, then $u_i \leq v_i$, $i = 0, 1, 2, \dots, N$.

Proof. The matrix associated with operator L_R^h is of size $(N+1) \times (N+1)$ and where for $i = 1$ and $i = N-1$, the terms involving u_0 and u_N have been moved to the right-hand side. It is easy to see that the matrix of coefficients is diagonally dominant and has nonpositive off-diagonal entries. Hence, the matrix is an irreducible M matrix. See the details of proof in [18]. \square

Lemma 5 (Discrete uniform stability estimate). The solution of the discrete scheme in (29) satisfies the bound

$$|u_i| \leq \theta^{-1} \|L_R^h u_i\| + \max\{|u_0|, |u_N|\}. \tag{37}$$

Proof. Let $r = \theta^{-1} \|L_R^h u_i\| + \max\{u_0, u_N\}$, and define the barrier function ϑ_i^\pm by $\vartheta_i^\pm = r \pm u_i$. On the boundary points, we obtain

$$\begin{aligned}
\vartheta_0^\pm &= r \pm u_0 = \theta^{-1} \|L_R^h u_i\| + \max\{u_0, u_N\} \pm \phi(0) \geq 0, \\
\vartheta_N^\pm &= r \pm u_N = \theta^{-1} \|L_R^h u_i\| + \max\{u_0, u_N\} \pm \psi(1) \geq 0.
\end{aligned}$$

On the discretized spatial domain x_i , $0 < i < N$, we obtain

$$\begin{aligned}
L_R^h \vartheta_i^\pm &\equiv -\frac{c_\varepsilon \sigma_1}{h^2} [(r \pm u_{i-1}) - 2(r \pm u_i) + (r \pm u_{i+1})] \\
&\quad + \lambda_1 [p(x_{i-1}) \left(\frac{-3(r \pm u_{i-1}) + 4(r \pm u_i) - (r \pm u_{i+1})}{2h} \right) \\
&\quad + q(x_{i-1})(r \pm u_{i-1})] + 2\lambda_2 [p(x_i) \left(\frac{(r \pm u_{i+1}) - (r \pm u_{i-1})}{2h} \right) \\
&\quad + q(x_i)(r \pm u_i)] + \lambda_1 [p(x_{i+1}) \left(\frac{(r \pm u_{i+1}) - 4(r \pm u_i) + 3(r \pm u_{i-1})}{2h} \right) \\
&\quad + q(x_{i+1})(r \pm u_{i+1})] \\
&= [\lambda_1 q(x_{i-1}) + 2\lambda_2 q(x_i) + \lambda_1 q(x_{i+1})] (\theta^{-1} \|L_R^h u_i\| + \max\{u_0, u_N\}) \\
&\quad \pm [\lambda_1 f(x_{i-1}) + 2\lambda_2 f(x_i) + \lambda_1 f(x_{i+1})] \geq 0, \quad \text{since } q_i \geq \theta > 0.
\end{aligned}$$

From Lemma 4, we obtain $\vartheta_i^\pm \geq 0$, for all $x_i \in \bar{\Omega}^N$. Hence, the required bound is obtained. \square

Now for $z > 0$, C_1 and C_2 are constants, and we have

$$C_1 \frac{z^2}{z+1} \leq z \coth(z) - 1 \leq C_2 \frac{z^2}{z+1}, \quad \text{and} \quad c_\varepsilon \frac{(h/c_\varepsilon)^2}{h/c_\varepsilon + 1} = \frac{h^2}{h + c_\varepsilon}, \quad (38)$$

giving that

$$|c_\varepsilon [p(1)\rho(\lambda_1 + \lambda_2) \coth(\frac{p(1)\rho}{2}) - 1] D^+ D^- u(x_i)| \leq \frac{Ch^2}{h + c_\varepsilon} \|u''(x_i)\|, \quad (39)$$

since $\lambda_1 + \lambda_2 \leq 1/2$. We obtain the bound

$$\begin{aligned} |c_\varepsilon [u''(x_i) - \sigma D^+ D^- u(x_i)]| &= |c_\varepsilon [p(1)\rho(\lambda_1 + \lambda_2) \coth(\frac{p(1)\rho}{2}) - 1] D^+ D^- u(x_i) \\ &\quad + c_\varepsilon (u''(x_i) - D^+ D^- u(x_i))| \\ &\leq \frac{Ch^2}{h + c_\varepsilon} \|u''(x_i)\| + Cc_\varepsilon h^2 \|u^{(4)}(x_i)\|. \end{aligned} \quad (40)$$

Now, let us denote the right-shifted, central, and left-shifted finite differences, respectively, as

$$\begin{aligned} D^R u(x_i) &= \frac{u_{i-1} - 4u_i + 3u_{i+1}}{2h}, \quad D^0 u(x_i) = \frac{u_{i+1} - u_{i-1}}{2h}, \quad \text{and} \\ D^L u(x_i) &= \frac{-3u_{i-1} + 4u_i - u_{i+1}}{2h}. \end{aligned}$$

Using Taylor's series approximation, we obtain the bound

$$\begin{aligned} |u'(x_{i-1}) - D^L u(x_{i-1})| &\leq Ch \|u''(\zeta)\|, \\ |u'(x_i) - D^0 u(x_i)| &\leq Ch^2 \|u'''(\zeta)\|, \quad \text{and} \\ |u'(x_{i+1}) - D^R u(x_{i+1})| &\leq Ch \|u''(\zeta)\|, \end{aligned} \quad (41)$$

where $\|u''(\zeta)\| = \max_{x_0 \leq x_i \leq x_N} |u''(x_i)|$ and $\|u'''(\zeta)\| = \max_{x_0 \leq x_i \leq x_N} |u'''(x_i)|$.

The next theorem gives the truncation error bound for the proposed scheme.

Theorem 1. Let $u(x_i)$ and u_i be the solution of (4)–(5) and (29), respectively. Then, the following error estimate holds:

$$|Lu(x_i) - L_R^h u_i| \leq Ch \left(1 + c_\varepsilon^{-3} \exp\left(-\frac{p^*(1-x_i)}{c_\varepsilon}\right)\right). \quad (42)$$

Proof. Consider the truncation error bound in the above discretization

$$\begin{aligned}
|Lu(x_i) - L_R^h u_i| &\leq |Lu(x_i) - L^h u(x_i)| + |L^h u(x_i) - L_R^h u_i| \\
&\leq \|T_1(h)\| + |c_\varepsilon u''(x_i) - c_\varepsilon \sigma D^+ D^- u(x_i)| \\
&\quad + |u'_{i-1} - D^L u(x_i)| + |u'_i - D^0 u(x_i)| \\
&\quad + |u'_{i+1} - D^R u(x_i)|.
\end{aligned} \tag{43}$$

Using the bounds in (40), (41), and (43) and using

$$T_1(h) \leq Ch^4(-2\lambda_1 + \lambda_2)\|u'''(\zeta_i)\| + Cc_\varepsilon h^4(1 - 12\lambda_1)\|u^{(4)}(\zeta_i)\|,$$

we obtain

$$\begin{aligned}
|Lu(x_i) - L_R^h u_i| &\leq \frac{Ch^2}{h + c_\varepsilon} \|u''(x_i)\| + c_\varepsilon Ch^2 \|u^{(4)}(x_i)\| + \lambda_1 Ch \|u''(\zeta)\| \\
&\quad + \lambda_2 Ch^2 \|u'''(\zeta)\| + Ch^4(-2\lambda_1 + \lambda_2)\|u'''(\zeta_i)\| \\
&\quad + Cc_\varepsilon h^4(1 - 12\lambda_1)\|u^{(4)}(\zeta_i)\|.
\end{aligned}$$

Using the bounds for the derivatives of the solution in Lemma 3 gives

$$\begin{aligned}
|Lu(x_i) - L_R^h u_i| &\leq \frac{Ch^2}{h + c_\varepsilon} (1 + c_\varepsilon^{-2} e^{\left(\frac{-p^*(1-x_i)}{c_\varepsilon}\right)}) + Ch^2 [c_\varepsilon (1 + c_\varepsilon^{-4} e^{\left(\frac{-p^*(1-x_i)}{c_\varepsilon}\right)}) \\
&\quad + Ch\lambda_1 (1 + c_\varepsilon^{-2} e^{\left(\frac{-p^*(1-x_i)}{c_\varepsilon}\right)}) + Ch^2 \lambda_2 (1 + c_\varepsilon^{-3} e^{\left(\frac{-p^*(1-x_i)}{c_\varepsilon}\right)}) \\
&\quad + Ch^4(-2\lambda_1 + \lambda_2) (1 + c_\varepsilon^{-3} e^{\left(\frac{-p^*(1-x_i)}{c_\varepsilon}\right)}) \\
&\quad + Cc_\varepsilon h^4(1 - 12\lambda_1) (1 + c_\varepsilon^{-4} e^{\left(\frac{-p^*(1-x_i)}{c_\varepsilon}\right)})] \\
&\leq \frac{Ch^2}{h + c_\varepsilon} (1 + c_\varepsilon^{-2} e^{\left(\frac{-p^*(1-x_i)}{c_\varepsilon}\right)}) + Ch^2 (c_\varepsilon + c_\varepsilon^{-3} e^{\left(\frac{-p^*(1-x_i)}{c_\varepsilon}\right)}) \\
&\quad + Ch\lambda_1 (1 + c_\varepsilon^{-2} e^{\left(\frac{-p^*(1-x_i)}{c_\varepsilon}\right)}) + Ch^2 \lambda_2 (1 + c_\varepsilon^{-3} e^{\left(\frac{-p^*(1-x_i)}{c_\varepsilon}\right)}) \\
&\quad + Ch^4(-2\lambda_1 + \lambda_2) (1 + c_\varepsilon^{-3} e^{\left(\frac{-p^*(1-x_i)}{c_\varepsilon}\right)}) \\
&\quad + Ch^4(1 - 12\lambda_1) (c_\varepsilon + c_\varepsilon^{-3} e^{\left(\frac{-p^*(1-x_i)}{c_\varepsilon}\right)}) \\
&\leq Ch (1 + c_\varepsilon^{-3} e^{\left(\frac{-p^*(1-x_i)}{c_\varepsilon}\right)}), \quad \text{since } c_\varepsilon^{-3} \geq c_\varepsilon^{-2}.
\end{aligned}$$

□

Lemma 6. For a fixed number of mesh numbers N and for $c_\varepsilon \rightarrow 0$, it holds

$$\lim_{c_\varepsilon \rightarrow 0} \max_{1 \leq i \leq N-1} \frac{\exp\left(\frac{-\alpha(x_i)}{c_\varepsilon}\right)}{c_\varepsilon^m} = 0, \quad \lim_{c_\varepsilon \rightarrow 0} \max_{1 \leq i \leq N-1} \frac{\exp\left(\frac{-\alpha(1-x_i)}{c_\varepsilon}\right)}{c_\varepsilon^m} = 0, \tag{44}$$

for $m = 1, 2, 3, \dots$, where $x_i = ih$, $h = 1/N$, for all $i = 1, 2, \dots, N-1$.

Proof. See [39].

□

Theorem 2. Let $u(x_i)$ and u_i be the solution of (4)–(5) and (29), respectively. Then it satisfies the error bound

$$\sup_{0 < c_\varepsilon \ll 1} \|u(x_i) - u_i\| \leq Ch. \quad (45)$$

Proof. Using the result in Lemma 6 into Theorem 1 and using the result in Lemma 5, we obtain the required bound. \square

4 Numerical results and discussion

In this section, we consider examples to illustrate the theoretical analysis of the proposed scheme.

Example 1. Consider the problem from [28]

$$-\varepsilon u''(x) + (1 + e^{x^2})u'(x) + xe^x u(x - \delta) + x^2 u(x) + (1 - e^{-x})u(x + \eta) = -1$$

with interval conditions $u(x) = 1$, $-\delta \leq x \leq 0$, and $u(x) = -1$, $1 \leq x \leq 1 + \eta$.

Example 2. Consider the problem from [28]

$$-\varepsilon u''(x) + u'(x) - 2u(x - \delta) + 5u(x) - u(x + \eta) = 0$$

with interval conditions $u(x) = 1$, $-\delta \leq x \leq 0$ and $u(x) = -1$, $1 \leq x \leq 1 + \eta$. The exact solution is given as

$$u(x) = \frac{(1 + e^{m_2})e^{m_1 x} - (1 + e^{m_1})e^{m_2 x}}{e^{m_2} - e^{m_1}},$$

where

$$m_1 = \frac{-(-1 - 2\delta + \eta) + \sqrt{(-1 - 2\delta + \eta)^2 - 4(\varepsilon + \delta^2 + \eta^2/2)}}{2(\varepsilon + \delta^2 + \eta^2/2)},$$

$$m_2 = \frac{-(-1 - 2\delta + \eta) - \sqrt{(-1 - 2\delta + \eta)^2 - 4(\varepsilon + \delta^2 + \eta^2/2)}}{2(\varepsilon + \delta^2 + \eta^2/2)}.$$

Example 3. Consider the problem from [28]

$$-\varepsilon u''(x) - u'(x) + 2u(x - \delta) + 5u(x) - u(x + \eta) = 0$$

with interval conditions $u(x) = 1$, $-\delta \leq x \leq 0$ and $u(x) = 0$, $1 \leq x \leq 1 + \eta$. The exact solution is given as

$$u(x) = \frac{e^{m_1 x + m_2} - e^{m_1 + m_2 x}}{e^{m_2} - e^{m_1}},$$

where

$$m_1 = \frac{-(1+2\delta+\eta) + \sqrt{(1+2\delta+\eta)^2 - 4(\varepsilon - \delta^2 + \eta^2/2)}}{2(\varepsilon - \delta^2 + \eta^2/2)},$$

$$m_2 = \frac{-(1+2\delta+\eta) - \sqrt{(1+2\delta+\eta)^2 - 4(\varepsilon - \delta^2 + \eta^2/2)}}{2(\varepsilon - \delta^2 + \eta^2/2)}.$$

Example 4. Consider the problem

$$-\varepsilon u''(x) - (1 + \exp(-x^2))u'(x) - xu(x - \delta) - x^2 u(x) - (1.5 - \exp(-x))u(x + \eta) = 1$$

with interval conditions $u(x) = 1$, $-\delta \leq x \leq 0$ and $u(x) = 1$, $1 \leq x \leq 1 + \eta$.

Since, the exact solution of the variable coefficient problems is not known, we applied the double mesh technique to calculate the maximum absolute error.

Let U_i^N denote the computed solution of the problem on N number of mesh points and let U_i^{2N} denote the computed solution on a double number of mesh points $2N$ by including the mid-points $x_{i+1/2} = \frac{x_{i+1} + x_i}{2}$ into the mesh points. The maximum absolute error is given by

$$E_{\varepsilon, \delta, \eta}^N = \max_i |U_i^N - u(x_i)|, \text{ or } E_{\varepsilon, \delta, \eta}^N = \max_i |U_i^N - U_i^{2N}|,$$

and the ε -uniform error is calculated using $E^N = \max_{\varepsilon, \delta, \eta} |E_{\varepsilon, \delta, \eta}^N|$. The rate of convergence of the scheme is calculated using $r_{\varepsilon, \delta, \eta}^N = \log_2 (E_{\varepsilon, \delta, \eta}^N / E_{\varepsilon, \delta, \eta}^{2N})$, and the ε -uniform rate of convergence is calculated using $r^N = \log_2 (E^N / E^{2N})$.

Table 1: Example 1, maximum absolute error of the scheme for $\lambda_1 = 1/12$, $\lambda_2 = 5/12$.

$\varepsilon \downarrow$	$N = 2^5$	2^6	2^7	2^8	2^9	2^{10}
2^0	9.4299e-04	5.6749e-04	3.0865e-04	1.6056e-04	8.1847e-05	4.1316e-05
2^{-2}	1.4597e-03	6.2724e-04	2.9779e-04	1.4502e-04	7.1636e-05	3.5613e-05
2^{-4}	1.5547e-03	8.9197e-04	4.0302e-04	1.5441e-04	7.2554e-05	3.5885e-05
2^{-6}	2.6082e-03	1.0350e-03	4.1282e-04	2.3469e-04	1.0215e-04	3.9059e-05
2^{-8}	3.0394e-03	1.5279e-03	6.7558e-04	2.6400e-04	1.0473e-04	5.9369e-05
2^{-10}	3.0383e-03	1.5588e-03	7.8897e-04	3.8892e-04	1.7038e-04	6.6326e-05
2^{-12}	3.0377e-03	1.5585e-03	7.8910e-04	3.9701e-04	1.9905e-04	9.7666e-05
2^{-14}	3.0376e-03	1.5584e-03	7.8907e-04	3.9699e-04	1.9911e-04	9.9709e-05
2^{-16}	3.0376e-03	1.5584e-03	7.8906e-04	3.9699e-04	1.9911e-04	9.9708e-05
2^{-18}	3.0376e-03	1.5584e-03	7.8906e-04	3.9699e-04	1.9911e-04	9.9708e-05
2^{-20}	3.0376e-03	1.5584e-03	7.8906e-04	3.9699e-04	1.9911e-04	9.9708e-05
E^N	3.0376e-03	1.5584e-03	7.8906e-04	3.9699e-04	1.9911e-04	9.9708e-05
r^N	0.9629	0.9819	0.9910	0.9955	0.9978	-

Four examples with their solution exhibiting a boundary layer are considered. Examples 1 and 2 have the boundary layer on the right side of the domain and Examples 3 and 4 have it on the left side of the domain. For the detail, one can observe in Figure 1, the layer formation of the solutions for different values of ε . In Figure 2, the influence of the delay parameter on the solution profile is given by considering different values of the delay

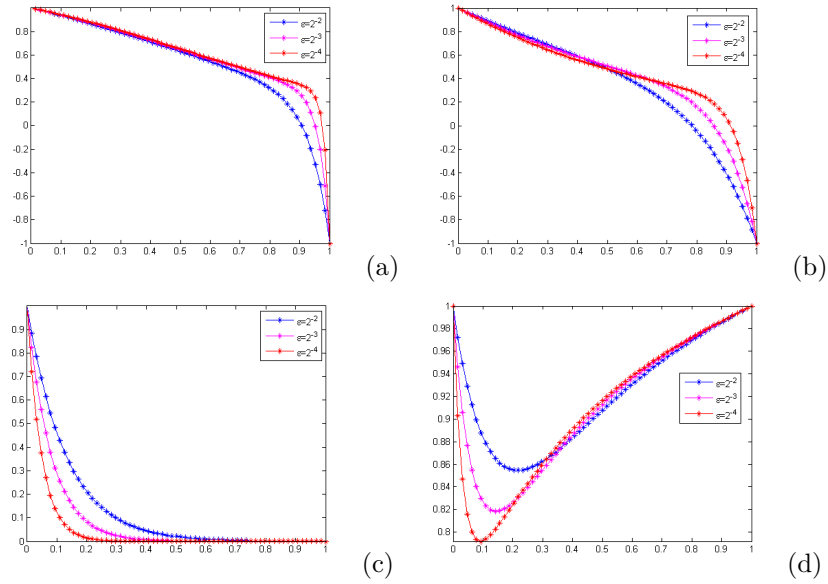


Figure 1: Boundary layer formation for different values of ε and $\delta = 0.6\varepsilon, \eta = 0.5\varepsilon$, (a) Example 1, (b) Example 2, (c) Example 3 and (d) Example 4.

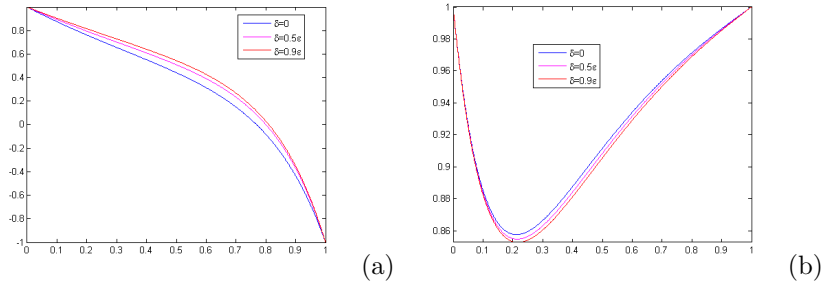


Figure 2: Solution profile for different values of delay parameter for $\varepsilon = 2^{-2}$, (a) Example 2, (b) Example 4.

parameter for $\varepsilon = 2^{-2}$. The maximum absolute error of the proposed scheme for $\lambda_1 = 1/12$ and $\lambda_2 = 5/12$ is given in Tables 1–4, for different values of the perturbation parameter ε . One observes that as $\varepsilon \rightarrow 0$ in each column, the maximum absolute error becomes stable and uniform. This indicates that the proposed scheme is uniformly convergent. In the last two rows of these tables, the uniform error and uniform rate of the convergence of the scheme are given. In Table 5, the rate of convergence of the scheme is given for different values of ε ranging from 2^{-12} to 2^{-20} . It is observed that the scheme gives a linear order uniform convergence. In Tables 6–8, the comparison of the proposed scheme with the result in [28] is given. As we observe, the uni-

Table 2: Example 2, maximum absolute error of the scheme for $\lambda_1 = 1/12, \lambda_2 = 5/12$.

$\varepsilon \downarrow$	$N = 2^5$	2^6	2^7	2^8	2^9	2^{10}
2^0	7.4098e-04	3.7006e-04	1.8494e-04	9.2446e-05	4.6218e-05	2.3108e-05
2^{-2}	2.1452e-03	1.0603e-03	5.2699e-04	2.6277e-04	1.3121e-04	6.5561e-05
2^{-4}	3.3473e-03	1.5615e-03	7.5382e-04	3.7142e-04	1.8438e-04	9.1872e-05
2^{-6}	4.4315e-03	2.1468e-03	9.3461e-04	4.3381e-04	2.0955e-04	1.0317e-04
2^{-8}	6.2230e-03	2.7096e-03	1.1361e-03	5.5377e-04	2.4090e-04	1.1169e-04
2^{-10}	6.2230e-03	3.2715e-03	1.6132e-03	6.9087e-04	2.8575e-04	1.3957e-04
2^{-12}	6.3987e-03	3.2747e-03	1.6565e-03	8.3275e-04	4.0708e-04	1.7359e-04
2^{-14}	6.4018e-03	3.2751e-03	1.6567e-03	8.3326e-04	4.1787e-04	2.0915e-04
2^{-16}	6.4025e-03	3.2752e-03	1.6568e-03	8.3329e-04	4.1788e-04	2.0915e-04
2^{-18}	6.4025e-03	3.2752e-03	1.6568e-03	8.3329e-04	4.1788e-04	2.0915e-04
2^{-20}	6.4025e-03	3.2752e-03	1.6568e-03	8.3329e-04	4.1788e-04	2.0915e-04
E^N	6.4025e-03	3.2752e-03	1.6568e-03	8.3329e-04	4.1788e-04	2.0915e-04
r^N	0.9671	0.9832	0.9915	0.9957	0.9986	-

Table 3: Example 3, maximum absolute error of the scheme for $\lambda_1 = 1/12, \lambda_2 = 5/12$.

$\varepsilon \downarrow$	$N = 2^5$	2^6	2^7	2^8	2^9	2^{10}
2^0	9.7173e-04	4.7967e-04	2.3842e-04	1.1886e-04	5.9346e-05	2.9652e-05
2^{-2}	2.2582e-03	1.0918e-03	5.3781e-04	2.6712e-04	1.3311e-04	6.6445e-05
2^{-4}	4.1684e-03	1.9446e-03	9.3525e-04	4.6062e-04	2.2884e-04	1.1409e-04
2^{-6}	9.0197e-03	3.4155e-03	1.3580e-03	3.1052e-04	2.0955e-04	1.5427e-04
2^{-8}	1.3302e-02	6.2951e-03	2.4966e-03	9.1533e-04	3.7245e-04	1.7575e-04
2^{-10}	1.3600e-02	7.2675e-03	3.6950e-03	1.6616e-03	6.4185e-04	2.3339e-04
2^{-12}	1.3615e-02	7.2783e-03	3.7679e-03	1.9170e-03	9.4973e-04	4.2126e-04
2^{-14}	1.3618e-02	7.2804e-03	3.7690e-03	1.9182e-03	9.6771e-04	4.8587e-04
2^{-16}	1.3619e-02	7.2809e-03	3.7690e-03	1.9183e-03	9.6778e-04	4.8607e-04
2^{-18}	1.3619e-02	7.2809e-03	3.7690e-03	1.9183e-03	9.6778e-04	4.8607e-04
2^{-20}	1.3619e-02	7.2809e-03	3.7690e-03	1.9183e-03	9.6778e-04	4.8607e-04
E^N	1.3619e-02	7.2809e-03	3.7690e-03	1.9183e-03	9.6778e-04	4.8607e-04
r^N	0.9034	0.9499	0.9744	0.9871	0.9935	-

form error and uniform rate of convergence of the proposed scheme is better than that of in [28].

5 Conclusion

This article dealt with the numerical treatment of SPDDEs having shifts on the reaction terms. The solution of the considered problem exhibited the boundary layer on the left or right side of the domain as $\varepsilon \rightarrow 0$. The terms involving the shift were approximated using the Taylor series approximation. The exponentially fitted tension spline method was used for treating the resulting singularly perturbed boundary value problem. The first derivative terms were approximated using left-shifted, central, and right-shifted finite

Table 4: Example 4, maximum absolute error of the scheme for $\lambda_1 = 1/12, \lambda_2 = 5/12$.

$\varepsilon \downarrow$	$N = 2^5$	2^6	2^7	2^8	2^9	2^{10}
2^0	1.9784e-04	9.7513e-05	4.8442e-05	2.4139e-05	1.2049e-05	6.0194e-06
2^{-2}	3.7410e-04	2.0860e-04	1.0982e-04	5.6313e-05	2.8509e-05	1.4342e-05
2^{-4}	1.4674e-04	1.6633e-04	1.2390e-04	7.2396e-05	3.8825e-05	2.0070e-05
2^{-6}	1.4706e-03	3.3509e-04	4.4397e-05	4.6273e-05	3.5377e-05	2.0784e-05
2^{-8}	2.0794e-03	1.0391e-03	3.9571e-04	8.8140e-05	1.1611e-05	1.2172e-05
2^{-10}	2.0806e-03	1.0780e-03	5.4814e-04	2.6668e-04	1.0074e-04	2.2328e-05
2^{-12}	2.0806e-03	1.0780e-03	5.4844e-04	2.7658e-04	1.3881e-04	6.7101e-05
2^{-14}	2.0806e-03	1.0780e-03	5.4844e-04	2.7658e-04	1.3888e-04	6.9589e-05
2^{-16}	2.0806e-03	1.0780e-03	5.4844e-04	2.7658e-04	1.3888e-04	6.9589e-05
2^{-18}	2.0806e-03	1.0780e-03	5.4844e-04	2.7658e-04	1.3888e-04	6.9589e-05
2^{-20}	2.0806e-03	1.0780e-03	5.4844e-04	2.7658e-04	1.3888e-04	6.9589e-05
E^N	2.0806e-03	1.0780e-03	5.4844e-04	2.7658e-04	1.3888e-04	6.9589e-05
r^N	0.9486	0.9750	0.9876	0.9939	0.9969	-

Table 5: Example 1, $(r_{\varepsilon, \delta, \eta}^N)$ of the scheme for $\lambda_1 = 1/12, \lambda_2 = 5/12$.

$\varepsilon \downarrow$	$N = 2^5$	2^6	2^7	2^8	2^9
Example 1					
2^{-12}	0.9628	0.9819	0.9910	0.9960	1.0272
2^{-14}	0.9629	0.9819	0.9910	0.9955	0.9978
2^{-16}	0.9629	0.9819	0.9910	0.9955	0.9978
2^{-18}	0.9629	0.9819	0.9910	0.9955	0.9978
2^{-20}	0.9629	0.9819	0.9910	0.9955	0.9978
Example 4					
2^{-12}	0.9486	0.9750	0.9876	0.9946	1.0487
2^{-14}	0.9486	0.9750	0.9876	0.9939	0.9969
2^{-16}	0.9486	0.9750	0.9876	0.9939	0.9969
2^{-18}	0.9486	0.9750	0.9876	0.9939	0.9969
2^{-20}	0.9486	0.9750	0.9876	0.9939	0.9969

Table 6: Comparison of uniform error and uniform rate of convergence of Example 1.

$\varepsilon \downarrow$	$N = 2^5$	2^6	2^7	2^8	2^9	2^{10}
Propose Scheme						
E^N	3.0376e-03	1.5584e-03	7.8906e-04	3.9699e-04	1.9911e-04	9.9708e-05
r^N	0.9629	0.9819	0.9910	0.9955	0.9978	-
Result in [28]						
E^N	8.9743e-02	4.6893e-02	2.4148e-02	1.5602e-02	7.5110e-03	3.5319e-03
r^N	0.9364	0.9575	0.6302	1.0547	1.0886	-

difference approximation. The stability of the scheme was investigated using the comparison principle and solution bound. The uniform convergence of the scheme was proved, and it gave the first-order uniform convergent. The performance of the scheme was compared with some published articles and it gave an accurate result.

Table 7: Comparison of uniform error and uniform rate of convergence of Example 2.

$\varepsilon \downarrow$	$N = 2^5$	2^6	2^7	2^8	2^9	2^{10}
	Propose	Scheme				
E^N	6.4025e-03	3.2752e-03	1.6568e-03	8.3329e-04	4.1788e-04	2.0915e-04
r^N	0.9671	0.9832	0.9915	0.9957	0.9986	-
	Result	in [28]				
E^N	1.0273e-01	6.1537e-02	3.8643e-02	2.2077e-02	1.2395e-02	7.0772e-03
r^N	0.7393	0.6712	0.8074	0.8328	0.8085	-

Table 8: Comparison of uniform error and uniform rate of convergence of Example 3.

$\varepsilon \downarrow$	$N = 2^5$	2^6	2^7	2^8	2^9	2^{10}
	Propose	Scheme				
E^N	1.3619e-02	7.2809e-03	3.7690e-03	1.9183e-03	9.6778e-04	4.8607e-04
r^N	0.9034	0.9499	0.9744	0.9871	0.9935	-
	Result	in [28]				
E^N	9.6126e-02	5.7165e-02	3.3247e-02	1.8984e-02	1.0685e-02	5.9444e-03
r^N	0.7498	0.7819	0.8084	0.8293	0.8459	-

Acknowledgements

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

References

1. Adilaxmi, M., Bhargavi, D. and Reddy, Y. *An initial value technique using exponentially fitted non standard finite difference method for singularly perturbed differential-difference equations*, Appl. Appl. Math. 14(1) (2019), 245–269.
2. Adilaxmi, M., Bhargavi, D. and Phaneendra, K. *Numerical integration of singularly perturbed differential difference problem using nonpolynomial interpolating function*, J. Inf. Math. Sci. 11(2) (2019), 195–208.
3. Adivi Sri Venkata, R.K. and Palli, M.M.K. *A numerical approach for solving singularly perturbed convection delay problems via exponentially fitted spline method*, Calcolo, 54 (2017) 943–961.
4. Baker C.T.H., Bocharov G.A. and Rihan F.A. *A report on the use of delay differential equations in numerical modeling in the biosciences*, Citeseer, 1999.
5. Bellen, A. and Zennaro, M. *Numerical methods for delay differential equations*, Clarendon Press, 2003.

6. Clavero, C., Gracia, J.L. and Jorge, J.C. *Higher order numerical methods for one dimensional parabolic singularly perturbed problems with regular layers*, Numerical Methods Partial Differential Eq. 21(1) (2005), 149–169.
7. Kadalbajoo, M.K. and Jha, A. *Analysis of fitted spline in compression for convection diffusion problems with two small parameters*, Neural, Parallel Sci. Comput. 19 (2011), 307–322.
8. Kadalbajoo, M.K. and Patidar, K.C. *Tension spline for the solution of self-adjoint singular perturbation problems*, Int.l J. Comput. Math. 79(7) (2002), 849–865.
9. Kadalbajoo, M.K., Patidar, K.C. and Sharma, K.K. *ε -uniformly convergent fitted methods for the numerical solution of the problems arising from singularly perturbed general DDEs*, Appl. Math. Comput. 182(1) (2006), 119–139.
10. Kadalbajoo, M.K. and Ramesh, V.P. *Numerical methods on Shishkin mesh for singularly perturbed delay differential equations with a grid adaptation strategy*, Appl. Math. Comput. 188 (2007), 1816–1831.
11. Kadalbajoo, M.K. and Sharma, K.K. *Numerical analysis of boundary-value problems for singularly-perturbed differential-difference equations with small shifts of mixed type*, J. Optim. Theory Appl. 115(1) (2002), 145–163.
12. Kadalbajoo, M.K. and Sharma, K.K. *Numerical treatment of a mathematical model arising from a model of neuronal variability*, J. Math. Anal. Appl. 307(2) (2005), 606–627.
13. Kadalbajoo, M.K. and Sharma, K.K. *An ε -uniform convergent method for a general boundary-value problem for singularly perturbed differential-difference equations: Small shifts of mixed type with layer behavior*, J. Comput. Methods Sci. Eng. 6(1-4) (2006), 39–55.
14. Kanth, A.S.V.R. and Kumar, P.M.M. *Numerical method for a class of nonlinear singularly perturbed delay differential equations using parametric cubic spline*, Int. J. Nonlinear Sci. Numer. Simul. 19(3-4) (2018), 357–365.
15. Kanth, A.S.V.R. and Kumar, P.M.M. *Computational results and analysis for a class of linear and nonlinear singularly perturbed convection delay problems on Shishkin mesh*, Hacet. J. Math. Stat. 49(1) (2020), 221–235.
16. Kanth, A.S.V.R. and Murali, M.K. *A numerical technique for solving nonlinear singularly perturbed delay differential equations*, Math. Model. Anal. 23(1) (2018), 64–78.
17. Khan, I. and Tariq, A. *Tension spline method for second-order singularly perturbed boundary-value problems*, Int. J. Comput. Math. 82(12) (2005), 1547–1553.

18. Kellogg, R.B. and Tsan, A. *Analysis of some difference approximations for a singular perturbation problem without turning points*, Math. Comput. 32(144) (1978), 1025–1039.
19. Kuang, Y. *Delay differential equations: with applications in population dynamics*, Academic Press, 1993.
20. Kumar, V. and Sharma, K.K. *An optimized B-spline method for solving singularly perturbed differential difference equations with delay as well as advance*, Neural Parallel Sci. Comput. 16(3) (2008), 371–386.
21. Lange, C.G. and Miura, R.M. *Singular perturbation analysis of boundary value problems for differential-difference equations*, SIAM J. Appl. Math. 42 (1982), 502–531.
22. Lange, C.G. and Miura, R.M. *Singular perturbation analysis of boundary value problems for differential-difference equations III. Turning point problems*, SIAM J. Appl. Math., 45(5) (1985), 708–734.
23. Lange, C.G. and Miura, R.M. *Singular perturbation analysis of boundary value problems for differential-difference equations. V. Small shifts with layer behavior*, SIAM J. Appl. Math. 54 (1994), 249–272.
24. Lange, C.G. and Miura, R.M. *Singular perturbation analysis of boundary value problems for differential-difference equations. VI. Small shifts with rapid oscillations*, SIAM J. Appl. Math. 54 (1994), 273–283.
25. Mahaffy, P.R., A’Hearn, M.F., Atreya, S.K., Bar-Nun, A., Bruston, P., Cabane, M., Carignan, G.R., Coll, P., Crifo, J.F., Ehrenfreund, P., Harpold, D. and Gorevan, S. *The Champollion cometary molecular analysis experiment*, Adv. Space Res., 23(2) (1999), 349–359.
26. Melesse, W.G., Tiruneh, A.A. and Derese G.A. *Solving linear second-order singularly perturbed differential difference equations via initial value method*, Int. J. Differ. Equ. (2019) 1-16.
27. Miller, J.J.H., O’Riordan, E. and Shishkin, G.I. *Fitted numerical methods for singular perturbation problems: error estimates in the maximum norm for linear problems in one and two dimensions*, World Scientific, 2012.
28. Mohapatra, J. and Natesan, S. *Uniformly convergent numerical method for singularly perturbed differential-difference equation using grid equidistribution*, Int. J. Numer. Meth. Biomed. Eng. 27(9) (2011), 1427–1445.
29. O’Malley, R.E. *Introduction to singular perturbations*, Academic Press, New York, 1974.
30. Palli, M.M. and Kanth, R.A. *Numerical simulation for a class of singularly perturbed convection delay problems*, Khayyam J. Math. 7(1) (2021), 52–64.

31. Ranjan, R. and Prasad, H.S. *A novel approach for the numerical approximation to the solution of singularly perturbed differential-difference equations with small shifts*, J. Appl. Math. Comput. 65(1-2) (2021), 403–427.
32. Roos, H.G., Stynes, M. and Tobiska, L.R. *Numerical methods for singularly perturbed differential equations*, Springer Science & Business Media, 2008.
33. Shishkin, GI. and Shishkina, LP. *Difference methods for singularly perturbed problems*, CRC, 2008.
34. Sirisha, L., Phaneendra, K. and Reddy Y.N. *Mixed finite difference method for singularly perturbed differential difference equations with mixed shifts via domain decomposition*, Ain Shams Eng. J. 9 (2018), 647–654.
35. Swamy, D.K., Phaneendra, K. and Reddy, Y.N. *Solution of singularly perturbed differential difference equations with mixed shifts using Galerkin method with exponential fitting*, Chin. J. Math. 2016, 1–10.
36. Swamy, D.K., Phaneendra, K. and Reddy, Y.N. *A fitted nonstandard finite difference method for singularly perturbed differential difference equations with mixed shifts*, J. de Afrikaana 3(4)(2016), 1–20.
37. Swamy, D.K., Phaneendra, K. and Reddy, Y.N. *Accurate numerical method for singularly perturbed differential difference equations with mixed shifts*, Khayyam J. Math. 4(2)(2018), 110–122.
38. Tian, H. *The exponential asymptotic stability of singularly perturbed delay differential equations with a bounded lag*, J. Math. Anal. Appl., 270(1)(2002), 143–149.
39. Turuna, D.A., Woldaregay, M.M. and Duressa, G.F. *Uniformly Convergent Numerical Method for Singularly Perturbed Convection-Diffusion Problems*, Kyungpook Math. J. 60(3) (2020).
40. Woldaregay, M.M. and Duressa, G.F. *Higher-Order Uniformly Convergent Numerical Scheme for Singularly Perturbed Differential Difference Equations with Mixed Small Shifts*, Int. J. Differ. Equ. 2020, (2020), 1–15.



New class of hybrid explicit methods for numerical solution of optimal control problems

M. Ebadi*, I. Malih Maleki and A. Ebadian

Abstract

Forward-backward sweep method (FBSM) is an indirect numerical method used for solving optimal control problems, in which the differential equation arising from this method is solved by the Pontryagin's maximum principle. In this paper, a set of hybrid methods based on explicit 6th-order Runge-Kutta method is presented for the FBSM solution of optimal control problems. Order of truncation error, stability region, and numerical results of the new hybrid methods were compared with those of the 6th-order Runge-Kutta method. Numerical results show that new hybrid methods are more accurate than the 6th-order Runge-Kutta method and that their stability regions are also wider than that of the 6th-order Runge-Kutta method.

AMS subject classifications (2020): 65K10; 65L20.

Keywords: FBSM; OCP; Stability analysis; Hybrid methods.

1 Introduction

Numerical methods used for solving optimal control problems (OCPs) are generally divided into two categories, direct and indirect methods. Indirect methods solve an OCP numerically based on the Pontryagin's maximum

*Corresponding author

Received 25 December 2020; revised 4 May 2021; accepted 20 May 2021

Moosa Ebadi

Department of Mathematics, University of Farhangian, Tehran, Iran. e-mail: Moosa.ebadi@yahoo.com

Isfand Malih Maleki

Department of Mathematics, Payam-e-Nour University, Tehran, Iran. e-mail: Esfand.malih@yahoo.com

Ali Ebadian

Department of Mathematics, Urmia university, Urmia, Iran. e-mail: A.ebadian@urmia.ac.ir

principle. The Pontryagin's maximum principle was proposed in 1954 by Pontryagin, a Russian mathematician. In the indirect method, an OCP is converted into a two-point boundary-value problem (TPBVP). The forward-backward sweep method (FBSM) is one of indirect numerical methods, which was first proposed in 2007 in a book entitled as "Optimal Control Applied to Biological Models" written by Lenhart and Workman [13]. In 2009, Silveira et al. [23] suggested six methods for classifying skin lesions in medical images and concluded that the FBSM performs better than the other methods. In 2012, McAsey, Moua, and Han [16] proved the convergence of the FBSM. In 2015, Moualeu et al. [17] used the FBSM to treat and control a type of tuberculosis with unknown cases in Comeroon. In 2015, Rose in his thesis [20] reported that the FBSM is more accurate than the direct shooting method and optimization method of MATLAB. In 2015, Sana et al. [22] used trapezoidal and Euler methods, instead of Runge-Kutta method through the FBSM to solve an OCP, compared them with the 4th-order Runge-Kutta method and concluded that trapezoidal and Runge-Kutta methods have the same performance in solving the OCP. Indeed, the performance of these two methods was improved by increasing step length compared to the Euler method; see [21]. In 2017, Lhous et al. [15] proposed a discrete mathematical model and optimal control to reduce the divorce rate. They used the Pontryagin's maximum principle and FBSM. They informed the people of the community about advantages of marriage and disadvantages of divorce, and in this way, they were able to reduce the divorce rate. In 2018, Kheiri and Jafari [11] proposed a general formulation for a fractional optimal control problem (FOCP), in which state and co-state equations are given in terms of left fractional derivatives. They used an improved FBSM, by the Adams-type method to solve the FOCP. In 2019, Duran, Candelo, and Ortiz [3] used a modified FBSM for reconfiguring unbalanced distribution systems. In 2019, Kongjeen et al. [12] proposed a modified FBSM for analyzing electrical charge of microgrids. In 2020, Ameen, Hidan, and Mostefaoui [1] proposed a mathematical model for studying the relationship between fish consumption and the prevalence of chronic heart disease (CHD). They used an improved FBSM based on a predictor-corrector method and concluded that eating fish reduces the risk of CHD and its mortality. In 2020, Bhih et al. [2] proposed a new model of rumor on social media and determined three optimal controls theoretically minimizing the number of spreader users, fake pages, and related costs. They used the FBSM to solve their optimization system in a duplicate process. In 2021, Ebadi et al. [9, 8] presented hybrid methods to numerically solve the OCP by the FBSM. They concluded that their proposed methods are more accurate than the FBSM in the presence of the explicit Runge-Kutta methods.

In this work, we present new hybrid methods for the FBSM solution of OCP. The paper is organized as follows: In Section 2, new methods and their order of truncation errors are presented. Stability analysis of the methods discussed in Section 3. We review the OCP, FBSM, and its convergence in

Sections 4 – 6. The results and final conclusions are presented in Sections 7 and 8, respectively.

2 Hybrid methods and order of truncation errors

Hybrid methods used for solving stiff differential equations are mostly efficient and have better response than other numerical methods. Consider the following initial value problem (IVP):

$$x' = f(t, x), \quad x \in \mathbb{R}^n, \quad x(t_0) = x_0, \quad t_0 \leq t \leq t_f, \quad (1)$$

where $f : [t_0, t_f] \times \mathbb{R}^n \rightarrow \mathbb{R}^n$. There are several explicit and implicit methods for solving such problems, but hybrid methods are more accurate than Runge–Kutta and backward differential formulas methods and have a wide range of stability; see [7, 5, 6, 4, 10]. Let us consider IVP in the form of (1). Linear k -step methods of the following form have $2k + 1$ arbitrary parameter:

$$x_{n+1} = \alpha_1 x_n + \alpha_2 x_{n-1} + \cdots + \alpha_k x_{n-k+1} + h\{\beta_0 f_{n+1} + \beta_1 f_n + \cdots + \beta_k f_{n-k+1}\}, \quad (2)$$

where $f_{n+1} = f(t_n + h, x_{n+1})$, $f_{n-m} = f(t_n - mh, x_{n-m})$ and $f_n = f(t_n, x_n)$ for $m = 1, 2, \dots, k - 1$. For increasing order of k -step methods in the form of (2), a linear combination of the slopes is used at several points between t_n and t_{n+1} , where $t_{n+1} = t_n + h$ in which h is the step length on $[t_0, t_f]$. Then, the modified form of (2) with m slopes is given by

$$x_{n+1} = \sum_{j=1}^k \alpha_j x_{n-j+1} + h \sum_{j=0}^k \beta_j f_{n-j+1} + h \sum_{j=1}^m \mu_j f_{n+v_j}, \quad (3)$$

where α_j , β_j , and μ_j are $2k + m + 1$ arbitrary parameters; see [10]. Methods of form (3) with m off-step points are called hybrid methods, where

$$0 < v_j < 1, \quad v_j \in \mathbb{R}, \quad j = 1, 2, \dots, m,$$

and herein, we set $k = 1$ and $m = 4$. Hence, we write (3) as

$$x_{n+1} = \alpha_1 x_n + h\{\beta_0 f_{n+1} + \beta_1 f_n\} + h\{\mu_1 f_{n+v_1} + \mu_2 f_{n+v_2} + \mu_3 f_{n+v_3} + \mu_4 f_{n+v_4}\}, \quad (4)$$

where $\alpha_1, \alpha_2, \beta_0, \beta_1$, and v_i s are arbitrary parameters and v_i is not equal to 0 or 1 for $i = 1, 2, 3, 4$. Expanding each term in (4), in Taylor's series about t_n , we can obtain a family of the 6th-order methods if the equations

Table 1: Values of v_i s in cases No. 1, 2, 3, 4, and 5 of new proposed methods

v_i	case 1	case 2	case 3	case 4	case 5
v_1	2.020e+00	2.020e+00	2.020e+00	2.020e+00	2.020e+00
v_2	4.000e-01	2.000e-01	4.000e-01	4.000e-01	1.000e-01
v_3	-2.000e-01	-3.000e-01	-2.000e-01	-2.000e-01	-2.000e-01
v_4	-1.840e+00	-1.540e+00	-1.440e+00	-1.140e+00	-1.840e+00

Table 2: Values of v_i s in cases No. 6, 7, 8, 9, and 10 of new proposed methods

v_i	case 6	case 7	case 8	case 9	case 10
v_1	2.020e+00	2.090e+00	2.020e+00	2.020e+00	2.020e+00
v_2	4.000e-01	4.000e-01	3.000e-01	3.000e-01	4.000e-01
v_3	-2.000e-01	-5.000e-01	-3.000e-01	-2.000e-01	-2.000e-01
v_4	-1.840e+00	-1.840e+00	-1.840e+00	-9.540e-02	-5.240e-01

$$\begin{aligned}
\alpha_1 &= 1, \\
\beta_0 + \beta_1 + c_1 + c_2 + c_3 + c_4 &= 1, \\
\beta_0 + c_1 v_1 + c_2 v_2 + c_3 v_3 + c_4 v_4 &= \frac{1}{2}, \\
\frac{1}{2}(\beta_0 + c_1 v_1^2 + c_2 v_2^2 + c_3 v_3^2 + c_4 v_4^2) &= \frac{1}{6}, \\
\frac{1}{6}(\beta_0 + c_1 v_1^3 + c_2 v_2^3 + c_3 v_3^3 + c_4 v_4^3) &= \frac{1}{24}, \\
\frac{1}{24}(\beta_0 + c_1 v_1^4 + c_2 v_2^4 + c_3 v_3^4 + c_4 v_4^4) &= \frac{1}{120}, \\
\frac{1}{120}(\beta_0 + c_1 v_1^5 + c_2 v_2^5 + c_3 v_3^5 + c_4 v_4^5) &= \frac{1}{720}, \\
\frac{1}{720}(\beta_0 + c_1 v_1^6 + c_2 v_2^6 + c_3 v_3^6 + c_4 v_4^6) &= \frac{1}{5040},
\end{aligned}$$

are satisfied, where the principal term of the truncation error is

$$\frac{1}{7!} c_7 h^7 x^{(7)}(t_n) + o(h^8), \quad c_7 = 1 - 7(\beta_0 + c_1 v_1^7 + c_2 v_2^7 + c_3 v_3^7 + c_4 v_4^7).$$

In this paper, the coefficients of (4) are proposed and obtained by searching and using coefficients that are more stable than the explicit Runge–Kutta methods. Values of v_i are determined with a wide stability region by searching for the interval of $[-2.5, 2.5]$. The v_i values are presented in ten cases as shown in Tables 1 and 2. They are called as the new proposed methods in this paper:

$$x_{n+1} = x_n + h\{\beta_1 f_n + \mu_1 f_{n+v_1} + \mu_2 f_{n+v_2} + \mu_3 f_{n+v_3} + \mu_4 f_{n+v_4} + \beta_0 f_{n+1}\}, \quad (5)$$

where $f_{n+1} = f(t_n + h, x_{n+1})$, $f_{n+v_i} = f(t_n + v_i h, x_{n+v_i})$, and $f_n = f(t_n, x_n)$ for $i = 1, 2, 3, 4$. Note that x_{n+1} , x_{n+v_i} , and x_n are numerical approximations according to the exact values of the solution $x(t)$ at $t_{n+1} = t_n + h$, $t_{n+v_i} = t_n + v_i h$. For converting (5) into explicit methods at each step, the values of x_{n+1} and x_{n+v_i} are predicted and used on right-hand side of the proposed methods using the 6th-order explicit Runge–Kutta (RK6) method, respectively, as follows:

$$\begin{aligned}
x_{n+1} &= x_n + \frac{h}{180}(9k_1 + 64k_3 + 49k_5 + 49k_6 + 9k_7), \\
k_1 &= hf(x_n, y_n), \\
k_2 &= hf(x_n + h, y_n + k_1), \\
k_3 &= hf(x_n + \frac{1}{2}h, y_n + \frac{1}{8}(3k_1 + k_2)), \\
k_4 &= hf(x_n + \frac{2}{3}h, y_n + \frac{1}{27}(8k_1 + 2k_2 + 8k_3)), \\
k_5 &= hf(x_n + \frac{h}{14}(7 - \sqrt[2]{21}), y_n + k_{5h}), \\
k_{5h} &= \frac{1}{392}(3(3\sqrt[2]{21} - 7)k_1 - 8(7 - \sqrt[2]{21})k_2 + 48(7 - \sqrt[2]{21})k_3) \\
&\quad + \frac{1}{392}(-3(21 - \sqrt[2]{21})k_4), \\
k_6 &= hf(x_n + \frac{h}{14}(7 + \sqrt[2]{21}), y_n + \frac{1}{1960}(-5(231 + 5\sqrt[2]{21})k_1 \\
&\quad - 40(7 + \sqrt[2]{21})k_2 + k_{6h}), \\
k_{6h} &= -320(\sqrt[2]{21})k_3 + 3(21 + 121\sqrt[2]{21})k_4 + 392(6 + \sqrt[2]{21})k_5, \\
k_7 &= hf(x_n + h, y_n + \frac{1}{180}(15(22 + 7\sqrt[2]{21})k_1 + 120k_2 \\
&\quad + 40(7\sqrt[2]{21} - 5)k_3 + k_{7h}), \\
k_{7h} &= -63(3\sqrt[2]{21} - 2)k_4 - 14(49 + 9\sqrt[2]{21})k_5 + 70(7 - \sqrt[2]{21})k_6.
\end{aligned} \tag{6}$$

$$\bar{x}_{n+1} = x_n + \frac{h}{180}(9k_1 + 64k_3 + 49k_5 + 49k_6 + 9k_7), \tag{7}$$

$$\bar{x}_{n+v_i} = x_n + \frac{v_i h}{180}(9k_1 + 64k_{3i} + 49k_{5i} + 49k_{6i} + 9k_{7i}), \quad i = 1, 2, 3, 4.$$

$$x_{n+1} = x_n + h\{\beta_1 f_n + \mu_1 \bar{f}_{n+v_1} + \mu_2 \bar{f}_{n+v_2} + \mu_3 \bar{f}_{n+v_3} + \mu_4 \bar{f}_{n+v_4} + \beta_0 \bar{f}_{n+1}\}, \tag{8}$$

where $k_{3i}, k_{5i}, k_{6i}, k_{7i}$, $i = 1, 2, 3, 4$, can be obtained by using the method 6 in which h is replaced by $v_i h$ and

$$f_{n+1} = f(t_n + h, x_{n+1}), \quad f_{n+v_i} = f(t_n + v_i h, x_{n+v_i}), \quad f_n = f(t_n, x_n).$$

Now suppose that the order of (8) is 6 similar to (6). Thus, the difference between exact and numerical solutions would be as follows:

$$x(t_{n+m}) - x_{n+m} = C_7 h^7 x^{(p_1)}(t_n) + O(h^8). \tag{9}$$

The difference operator associated with the 6th-order (8) can be written as

$$x(t_{n+1}) - x_{n+1} = C h^7 x^{(7)}(t_n) + O(h^8),$$

where C is the error constant of (8). Therefore, we have the following theorem.

Theorem 1. Given that (8) is of order p , then, p is equal to 6.

Proof. Suppose that $i = 1, 2, 3, 4$ and that x_n is exact. According to (7) and (8), one can write

$$\begin{aligned}
x(t_{n+1}) - x_{n+1} = & h \sum_{i=1}^4 \mu_i [f(t_{n+v_i}, x(t_{n+v_i})) - f(t_{n+v_i}, \bar{x}_{n+v_i})] \\
& + h\beta_0 [f(t_{n+1}, x(t_{n+1})) - f(t_{n+1}, \bar{x}_{n+1})] \\
& + Ch^p x^{(p)}(t_n) + O(h^{p+1}).
\end{aligned}$$

Considering the properties of the IVPs in (1), some values such as η_{v_i} and η_1 belong to intervals of $(\bar{x}_{n+v_i}, x(t_{n+v_i}))$ and $(\bar{x}_{n+1}, x(t_{n+1}))$, respectively. Thus, we can write

$$\begin{aligned}
f(t_{n+v_i}, x(t_{n+v_i})) - f(t_{n+v_i}, \bar{x}_{n+v_i}) &= \frac{\partial f}{\partial x}(t_{n+v_i}, \eta_{n+v_i})(x(t_{n+v_i}) - \bar{x}_{n+v_i}), \\
f(t_{n+1}, x(t_{n+1})) - f(t_{n+1}, \bar{x}_{n+1}) &= \frac{\partial f}{\partial x}(t_{n+1}, \eta_{n+1})(x(t_{n+1}) - \bar{x}_{n+1}).
\end{aligned}$$

Therefore, using (9), we have

$$\begin{aligned}
x(t_{n+1}) - x_{n+1} = & h \sum_{i=1}^4 \mu_i \left[\frac{\partial f}{\partial x}(t_{n+v_i}, \eta_{n+v_i})(x(t_{n+v_i}) - \bar{x}_{n+v_i}) \right] \\
& + h\beta_0 \left[\frac{\partial f}{\partial x}(t_{n+1}, \eta_{n+1})(x(t_{n+1}) - \bar{x}_{n+1}) \right] \\
& + Ch^p x^{(p)}(t_n) + O(h^{p+1}).
\end{aligned}$$

Applying (9) to this, we have

$$\begin{aligned}
x(t_{n+1}) - x_{n+1} = & h \sum_{i=1}^4 \mu_i \left[\frac{\partial f}{\partial x}(t_{n+v_i}, \eta_{n+v_i}) C_{v_i} h^6 x^{(6)}(t_n) + O(h^{6+1}) \right] \\
& + h\beta_0 \left[\frac{\partial f}{\partial x}(t_{n+1}, \eta_{n+1}) C_1 h^6 x^{(6)}(t_n) + O(h^{6+1}) \right] \\
& + Ch^p x^{(p)}(t_n) + O(h^{p+1}) \\
= & h^6 \left\{ \sum_{i=1}^4 \mu_i \left[\frac{\partial f}{\partial y}(t_{n+v_i}, \eta_{n+v_i}) C_{v_i} h^{6-p+1} x^{(p_1)}(t_n) \right] \right\} \\
& + h^6 \left\{ \beta_0 \left[\frac{\partial f}{\partial x}(t_{n+1}, \eta_{n+1}) C_1 h^{6-p+1} x^{(p_1)}(t_n) \right] \right. \\
& \left. + C x^{(7)}(t_n) \right\} + O(h^8).
\end{aligned}$$

Thus, it can be concluded that the order of (8) is 6. \square

3 Stability analysis of the new methods

In this section, stability analysis is done on new methods. Dahlquist test problem is considered to investigate stability region of the methods presented

in this study. Applying the Dahlquist test problem to (5) and inserting $p = 6$, the following equations can be obtained:

$$\bar{x}_{n+m} = \left(1 + m\bar{h} + \frac{(m\bar{h})^2}{2!} + \frac{(m\bar{h})^3}{3!} + \frac{(m\bar{h})^4}{4!} + \frac{(m\bar{h})^5}{5!} + \frac{(m\bar{h})^6}{6!}\right) x_n, \quad (10)$$

$$m = 1, v_1, v_2, v_3, v_4,$$

where $\bar{h} = h\lambda$ and $\mu_j \in \mathbb{R}$. Now, let us consider the new hybrid method presented in this work:

$$x_{n+1} = x_n + h\{\beta_1 f_n + \beta_0 \bar{f}_{n+1} + \mu_1 \bar{f}_{n+v_1} + \mu_2 \bar{f}_{n+v_2} + \mu_3 \bar{f}_{n+v_3} + \mu_4 \bar{f}_{n+v_4}\}. \quad (11)$$

Substituting (10) into (11), the following equation is obtained:

$$\begin{aligned} x_{n+1} = & x_n + \bar{h} \left\{ \beta_1 x_n + \beta_0 x_n \left(1 + \bar{h} + \frac{(\bar{h})^2}{2!} + \frac{(\bar{h})^3}{3!} + \frac{(\bar{h})^4}{4!} + \frac{(\bar{h})^5}{5!} + \frac{(\bar{h})^6}{6!} \right) \right\} \\ & + \bar{h} \left\{ x_n \sum_{i=1}^4 \mu_i \left(1 + (v_i \bar{h}) + \frac{(v_i \bar{h})^2}{2!} + \frac{(v_i \bar{h})^3}{3!} + \frac{(v_i \bar{h})^4}{4!} \right. \right. \\ & \quad \left. \left. + \frac{(v_i \bar{h})^5}{5!} + \frac{(v_i \bar{h})^6}{6!} \right) \right\}, \end{aligned}$$

Inserting $x_n = r^n$ into (11) and dividing it by r^n , we can obtain

$$r^{n+1} = r^n \{1 + a_1 \bar{h} + a_2 \bar{h}^2 + a_3 \bar{h}^3 + a_4 \bar{h}^4 + a_5 \bar{h}^5 + a_6 \bar{h}^6 + a_7 \bar{h}^7\}.$$

$$\Rightarrow r = 1 + a_1 \bar{h} + a_2 \bar{h}^2 + a_3 \bar{h}^3 + a_4 \bar{h}^4 + a_5 \bar{h}^5 + a_6 \bar{h}^6 + a_7 \bar{h}^7,$$

$$a_1 = \beta_1 + \beta_0 + \sum_{i=1}^4 \mu_i, \quad a_2 = \beta_0 + \sum_{i=1}^4 (v_i \mu_i),$$

$$a_3 = \frac{1}{2}(\beta_0 + \sum_{i=1}^4 (v_i^2 \mu_i)), \quad a_4 = \frac{1}{6}(\beta_0 + \sum_{i=1}^4 (v_i^3 \mu_i)),$$

$$a_5 = \frac{1}{24}(\beta_0 + \sum_{i=1}^4 (v_i^4 \mu_i)), \quad a_6 = \frac{1}{120}(\beta_0 + \sum_{i=1}^4 (v_i^5 \mu_i)),$$

$$a_7 = \frac{1}{720}(\beta_0 + \sum_{i=1}^4 (v_i^6 \mu_i)),$$

which is a stability polynomial of (11). Figure 1 compares the stability region of methods 1 and 2 with that of the 6th-order Runge–Kutta (RK6) method (note that method i means that the new method related to the case i , $i = 1, 2, 3, 4$). As can be clearly seen, the stability region of the proposed

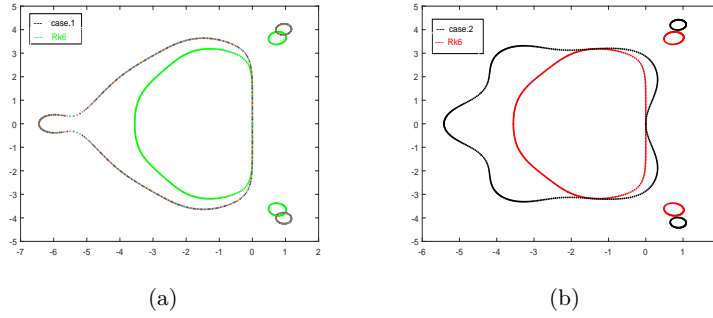


Figure 1: (a) Stability region of the method 1 and RK6 method. (b) Stability region of the method 2 and RK6 method.

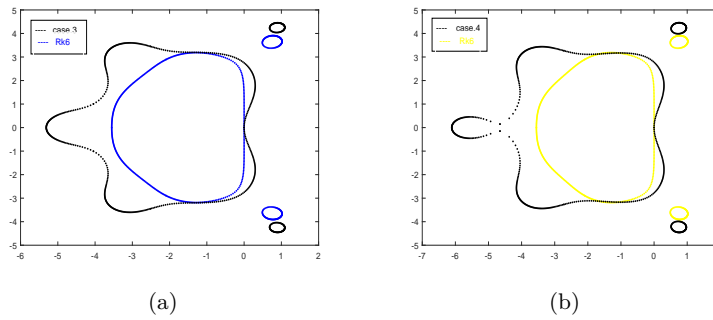


Figure 2: (a) Stability region of the method 3 and RK6 method. (b) Stability region of the method 4 and RK6 method.

methods 1 and 2 is much wider than that of the RK6 method. Stability region of the proposed methods 3 and 4 is presented in Figure 2 and is compared with that of the RK6 method. According to Figure 2, methods 3 and 4 presented in this paper have a wider range of stability than the RK6 method, showing the efficiency of the proposed methods. Methods 5 and 6 are compared with RK6 method in terms of stability region in Figure 3. As demonstrated in Figure 3, the proposed methods 5 and 6 have a wider stability region than the RK6 method, showing that the efficiency of the proposed methods is higher than the RK6 method.

Figure 4 compares the stability region of methods 7 and 8 with that of the RK6 method. As can be seen, methods 7 and 8 have a wider range of stability than the RK6 method. Figure 5 also shows the superiority of the methods proposed in this paper over the RK6 method. As depicted in Figure 5, the width of stability region of methods 9 and 10 is higher compared to the RK6 method.

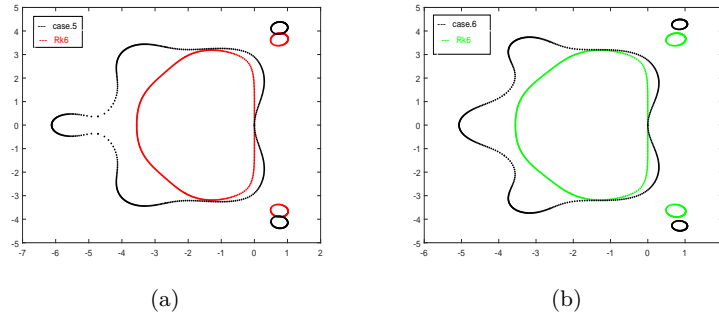


Figure 3: (a) Stability region of the method 5 and RK6 method. (b) Stability region of the method 6 and RK6 method.

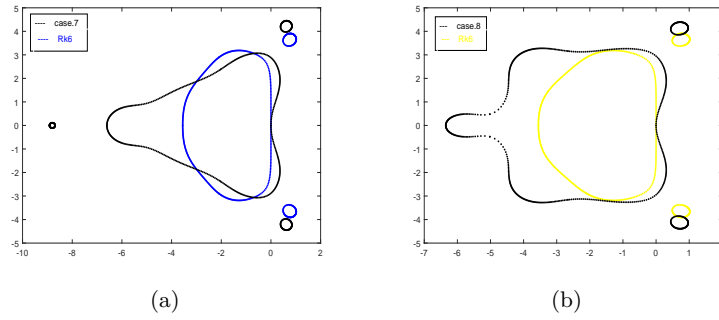


Figure 4: (a) Stability region of the method 7 and RK6 method. (b) Stability region of the method 8 and RK6 method.

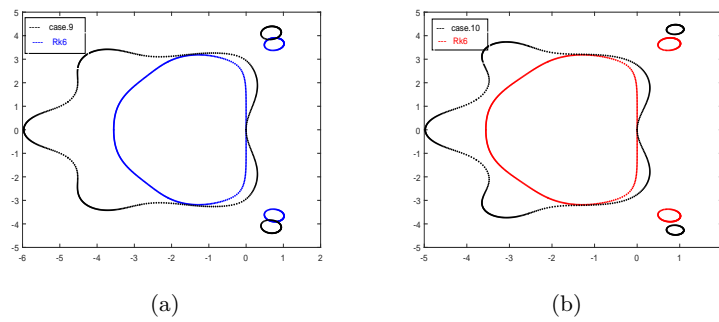


Figure 5: (a) Stability region of the method 9 and RK6 method. (b) Stability region of the method 10 and RK6 method.

4 Optimal control problems

An OCP includes a cost function $J(x, u)$, a set of state variables, $x \in X$, and a set of control variables, $u \in U$. An OCP is solved to find a piecewise continuous control $u(t)$, $t_0 \leq t \leq t_f$, and the associated continuous state variable $x(t)$, in order to minimize the given objective function. For more explanation, we need a few definitions.

Definition 1 (Lagrange and Bolza problems). The basic problem in Lagrange form is

$$J(x, u) = \int_{t_0}^{t_f} g(t, x(t), u(t)) dt. \quad (12)$$

Adding another term to functional (12), the Bolza problem is obtained:

$$J(x, u) = h(t_f, x(t_f)) + \int_{t_0}^{t_f} g(t, x(t), u(t)) dt.$$

An OCP is

$$\max_u J = \int_{t_0}^{t_f} g(t, x(t), u(t)) dt, \quad (13)$$

$$\begin{aligned} x'(t) &= f(t, x(t), u(t)), \\ x(t_0) &= x_0. \end{aligned}$$

Note that $\min\{J\} = -\max\{-J\}$; see [19].

Definition 2 (Hamiltonian). Consider the OCP (13). The function $H(t, x, u, \lambda)$ is called as the Hamiltonian function and is equal to

$$H(t, x, u, \lambda) = g(t, x, u) + \lambda f(t, x, u)$$

where λ is an adjoint variable.

Theorem 2 (Pontryagin Maximum Principle). Consider the OCP (13). Suppose that $g(t, x, u)$ and $f(t, x, u)$ are both continuously differentiable functions in their three arguments and concave in x and u . If u^* is a control with associated state x^* and λ is a piecewise differentiable function such that u^* , x^* , and λ together are satisfied

$$\begin{aligned} g_u + \lambda f_u &= 0 \Leftrightarrow \frac{\partial H}{\partial u} = 0, \\ \lambda' &= -(g_x + \lambda f_x) \Leftrightarrow \lambda' = -\frac{\partial H}{\partial x}, \\ \lambda(t_f) &= 0, \\ \lambda(t) &\geq 0, \end{aligned}$$

on $t_0 \leq t \leq t_f$, then

$$J(x^*, u^*) \geq J(x, u),$$

for any admissible pair (x, u) .

Proof. We refer readers to [13]. □

Theorem 3. [13]. Let the set of controls for problem (13), be Lebesgue integrable functions and let $t_0 \leq t \leq t_f$ in \mathbb{R} . Suppose that $f(t, x, u)$ is concave in u and there exist constants $c_1, c_2, c_3 > 0, c_4$ and $\beta > 1$, such that

$$\begin{aligned} f(t, x, u) &= \alpha(t, x) + \beta(t, x)u, \\ |f(t, x, u)| &\leq c_1(1 + |x| + |u|), \\ |f(t, x_1, u) - f(t, x, u)| &\leq c_2|x_1 - x|(1 + |u|), \\ g(t, x, u) &\leq c_3|u|^\beta - c_4, \end{aligned}$$

for all t with $t_0 \leq t \leq t_f$, $x_1, x_2, u \in \mathbb{R}$.

Then there exists an optimal pair (x^*, u^*) maximizing J , with finite $J(x^*, u^*)$.

5 Forward-backward sweep method

For numerically solving the OCP (13) using the indirect method, an algorithm is introduced according to the literature [13]. For solving such problems numerically first, an algorithm that generates an approximation to an optimal piecewise continuous control u^* , must divide the time interval of $[t_0, t_f]$ into pieces with specific points of interest $t_0 = b_1, b_2, \dots, b_N, b_{N+1} = t_f$; and these points will usually be equally spaced. Approximation will be a vector $\vec{u} = (u_1, u_2, \dots, u_{N+1})$, where $u_i \approx u(b_i)$. Any solution to the above OCP must also be satisfied:

$$\begin{aligned} x'(t) &= f(t, x(t), u(t)), \quad x(t_0) = x_0, \\ \lambda' &= -\frac{\partial H}{\partial x}, \quad \lambda(t_f) = 0, \\ \frac{\partial H}{\partial u} &= 0 \quad \text{at} \quad u^*. \end{aligned}$$

The third equation, namely, optimality condition, can usually be manipulated to find a representation of u^* in terms of t, x , and λ . Then, the first two equations form a TPBVP. The generalized problem can be solved using indirect methods, which are numerical techniques used for solving. The FBSM is one of these methods. A rough outline of the algorithm is given below.

Here, $\vec{x} = (x_1, x_2, \dots, x_{N+1})$ and $\vec{\lambda} = (\lambda_1, \lambda_2, \dots, \lambda_{N+1})$ are the vector approximations for the state and adjoint.

1. Make an initial guess for \vec{u} over the interval.
2. Using the initial condition $x_1 = x(t_0) = a$ and the value for \vec{u} , solve \vec{x} forward in time according to its differential equation in the optimality system.
3. Using the transversality condition $\lambda_{N+1} = \lambda(t_f) = 0$ and the values for \vec{u} and \vec{x} , solve $\vec{\lambda}$ backward in time according to its differential equation in optimality system.
4. Update \vec{u} by entering the new \vec{x} and $\vec{\lambda}$ values into the characterization of the optimal control.
5. Check convergence. If values of variables in this iteration and the last iteration are negligibly close, then the current values are considered as output solutions. If values are not close, then return to Step 2.

6 Convergence of FBSM

For notational simplicity, we express the problem as finding $(x(t), \lambda(t), u(t))$ such that

$$\begin{aligned} x'(t) &= f(t, x(t), u(t)), & x(t_0) &= x_0, \\ \lambda'(t) &= k_1(t, x(t), u(t)) + \lambda(t)k_2(t, x(t), u(t)), & \lambda(t_f) &= 0, \\ u(t) &= k_3(t, x(t), u(t)). \end{aligned}$$

Here, $x_0 \in \mathbb{R}^n$ and $t_0 < t_f$ are the given real numbers. For a convergence analysis of the FBSM, we will make the following assumptions:

(T) The functions of f, k_1, k_2 , and k_3 are Lipschitz continuous with respect to their second and third arguments, with Lipschitz constants of $L_f, L_{k_1}, L_{k_2}, L_{k_3}$. Moreover, $\Lambda = \|\lambda\|_\infty$ and $H = \|k_2\|_\infty < \infty$.

Theorem 4. Under the assumptions (T), if

$$c_0 \equiv L_{k_3} \{[\exp(L_f(t_f - t_0)) - 1]\} + L_{k_3} \{(L_{k_1} + \Lambda L_{k_2}) \frac{1}{H} [\exp(H(t_f - t_0)) - 1][\exp(L_f(t_f - t_0)) + 1]\} < 1,$$

then the FBSM is convergent, that is, as $n \rightarrow \infty$,

$$\max_{t_0 \leq t \leq t_f} |x(t) - x^{(n)}(t)| + \max_{t_0 \leq t \leq t_f} |\lambda(t) - \lambda^{(n)}(t)| + \max_{t_0 \leq t \leq t_f} |u(t) - u^{(n)}(t)| \rightarrow 0.$$

Proof. We refer the reader to [16]. □

7 Numerical results

In this section, examples of various types of OCPs are solved using the proposed methods, and their numerical results are compared with those of the FBSM using the *RK* method of order 6 (FBSM–RK6). One of the most important OCPs is the linear regulator problem, which is generally defined as follows.

Example 1. Let E , $Q(t)$, and $R(t)$ be symmetric and nonnegative definite matrices of appropriate dimensions. The so-called linear regulator problem (with a linear state-space description) involves a cost functional of the form

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

For example, we consider an OCP as follows [16]:

$$\begin{aligned} \min_u \quad & \frac{1}{2} \int_0^1 [x(t)^2 + u(t)^2] dt \\ \text{s.t.} \quad & x'(t) = -x(t) + u(t), \quad x(0) = 1. \end{aligned}$$

The Pontryagin's maximum principle can be used to construct an analytic solution

$$\begin{aligned} H(t, x, u, \lambda) &= \frac{1}{2}(x(t)^2 + u(t)^2) + \lambda(-x(t) + u(t)), \\ \frac{\partial H}{\partial u} &= 0 \quad \text{at} \quad u^* \Rightarrow u^* + \lambda = 0 \Rightarrow u^* = -\lambda, \\ \lambda' &= -\frac{\partial H}{\partial x} = -x + \lambda, \quad \lambda(1) = 0. \end{aligned}$$

Together with the state equation, the result will be the following linear differential algebraic system:

$$\begin{aligned} x'(t) &= -x(t) + u(t), \quad x(0) = 1, \\ \lambda'(t) &= \lambda(t) - x(t), \quad \lambda(1) = 0, \quad u(t) = -\lambda(t). \end{aligned}$$

The solution is

$$\begin{aligned} x^*(t) &= \frac{\sqrt{2} \cosh(\sqrt{2}(t-1)) - \sinh(\sqrt{2}(t-1))}{\sqrt{2} \cosh(\sqrt{2}) + \sinh(\sqrt{2})}, \\ u^*(t) &= \frac{\sinh(\sqrt{2}(t-1))}{\sqrt{2} \cosh(\sqrt{2}) + \sinh(\sqrt{2})}. \end{aligned}$$

Optimal value of the objective functional is $J = 0.1929092981$. The final values of the state and optimal are $x(1) = 0.2819695346$ and 0, respectively,

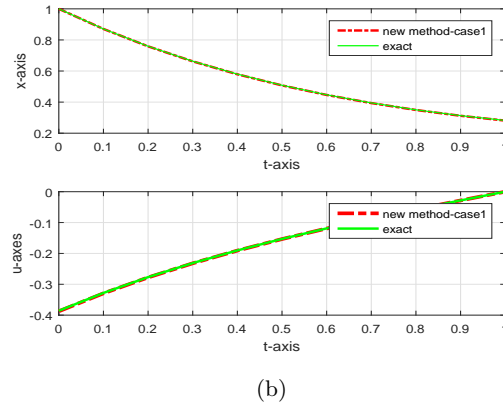
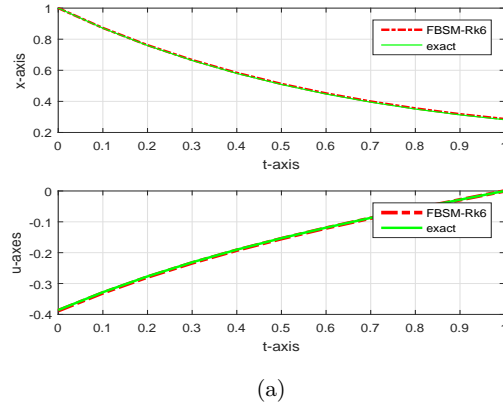


Figure 6: (a) Optimal state and control values of Example 1 FBSM–RK6. (b) Optimal state and control values of Example 1 (new proposed method)

and the initial value of the co-state is $\lambda(0) = 0.3858185962$. Numerical results of the problem are shown in Figure 6 with $h = \frac{1}{10}$ and in Tables 3 and 4.

Numerical results presented in Tables 3 and 4 indicate that each of new ten suggested methods calculates the amount of control variable values much more accurately than the FBSM–RK6 method. Figure 6 also indicates that the new suggested method is exactly based on figure of analytical answer. For avoiding overstatement in this paper, one of the diagrams was selected and drawn. The rest of figures are similar to each other. Approximate performance index is calculated for the proposed method, and the results are presented in Tables 5 and 6. According to the results, the precision of performance index of new proposed methods is more than that of the FBSM–RK6 method by two digits. The Pontryagin's Theorem is also used to solve linear-quadratic problems.

Table 3: Error of control values in Example 1 for FBSM–RK6 and new proposed Methods

t	h	FBSM_RK6	case 1	case 2	case 3	case 4	case 5
0.90	$\frac{1}{10}$	3.0520e-4	7.3292e-5	2.5423e-5	7.2638e-5	7.2172e-5	1.4836e-5
0.90	$\frac{1}{50}$	5.3457e-5	2.1141e-5	1.5067e-5	2.1006e-5	2.0916e-5	4.1655e-6
0.90	$\frac{1}{100}$	2.6278e-5	1.0962e-5	1.4606e-6	1.0895e-5	1.0850e-5	2.5154e-6
0.90	$\frac{1}{200}$	1.3036e-5	5.5876e-7	8.4673e-7	5.5537e-6	5.5314e-6	1.3740e-6

Table 4: Error of control values in Example 1 for FBSM–RK6 and new proposed methods

t	h	FBSM_RK6	case 6	case 7	case 8	case 9	case 10
0.90	$\frac{1}{10}$	3.0520e-4	6.7372e-5	1.1476e-4	3.9201e-5	2.9350e-5	7.1123e-5
0.90	$\frac{1}{50}$	5.3457e-5	1.9984e-5	1.5321e-5	5.6249e-7	1.3363e-6	2.0701e-5
0.90	$\frac{1}{100}$	2.6278e-5	1.0386e-5	7.1974e-6	1.6043e-7	1.1054e-6	1.0742e-5
0.90	$\frac{1}{200}$	1.3036e-5	5.2998e-6	3.4747e-6	1.9879e-7	6.7022e-7	5.4774e-6

Table 5: Errors of the performance index approximation in Example 1

h	FBSM_RK6	case 1	case 2	case 3	case 4	case 5
$\frac{1}{50}$	4.0877e-4	2.8316e-5	4.7743e-5	2.7735e-5	2.7293e-5	4.3302e-5
$\frac{1}{100}$	1.9310e-4	6.8092e-6	3.0969e-5	6.5235e-6	6.3012e-6	2.8836e-5
$\frac{1}{200}$	9.2656e-5	4.877e-7	1.8338e-5	3.4612e-7	2.3451e-7	1.7294e-5
$\frac{1}{1000}$	1.8014e-5	2.6262e-7	4.0178e-6	2.9075e-7	3.1314e-7	3.8123e-6

Table 6: Errors of the performance index approximation in Example 1

h	FBSM_RK6	case 6	case 7	case 8	case 9	case 10
$\frac{1}{50}$	4.0877e-4	2.3141e-5	1.0728e-4	6.1400e-5	5.4438e-5	2.6399e-5
$\frac{1}{100}$	1.9310e-4	4.2415e-6	6.0399e-5	3.7814e-5	3.4359e-5	5.8643e-6
$\frac{1}{200}$	9.2656e-5	7.9113e-7	3.2969e-5	2.1765e-5	2.0043e-5	1.8678e-8
$\frac{1}{1000}$	1.8014e-5	5.1760e-7	6.9304e-6	4.7036e-6	4.3604e-6	3.5590e-7

Table 7: Control values errors in Example 2 by using FBSM–RK6 and new proposed methods

t	h	FBSM–RK6	case 1	case 2	case 3	case 4	case 5
0.70	$\frac{1}{10}$	8.2317e-3	1.3560e-4	3.7467e-4	1.3073e-4	1.3148e-4	3.1883e-4
0.70	$\frac{1}{50}$	8.1160e-4	3.7614e-5	6.5252e-5	3.6716e-5	3.6772e-5	5.6347e-5
0.70	$\frac{1}{100}$	4.0487e-4	1.9678e-5	3.1791e-5	1.9236e-5	1.9258e-5	2.7488e-5
0.70	$\frac{1}{200}$	1.0203e-4	1.0221e-5	1.5522e-5	1.0001e-5	1.0010e-5	1.3409e-5

Example 2. Consider the following OCP [18]:

$$\min_u \int_0^1 \frac{5}{8}x(t)^2 + \frac{1}{2}x(t)u(t) + \frac{1}{2}u(t)^2 dt$$

$$st. \ x'(t) = \frac{1}{2}x(t) + u(t), \ x(0) = 1.$$

For solving the above example, using the FBSM and proposed methods, we should apply the Pontryagin's Theorem as follows:

$$H(t, x, u, \lambda) = \frac{5}{8}x(t)^2 + \frac{1}{2}x(t)u(t) + \frac{1}{2}u(t)^2 + \lambda(\frac{1}{2}x(t) + u(t)),$$

$$\frac{\partial H}{\partial u} = 0 \quad at \quad u^* \Rightarrow u^* = -\lambda - \frac{1}{2}x,$$

$$\lambda' = -\frac{\partial H}{\partial x} = -\frac{10}{8}x - \frac{1}{2}u - \frac{1}{2}\lambda, \quad \lambda(1) = 0.$$

Analytical solutions are as follows [18]:

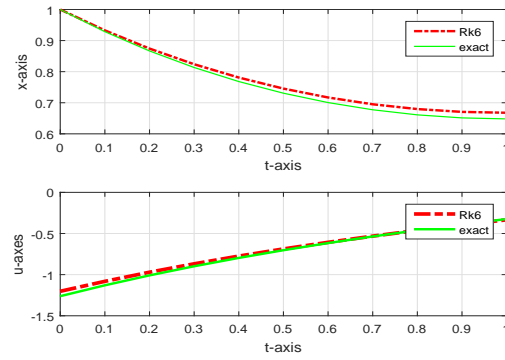
$$u^*(t) = -\frac{(\tanh(1-t) + .5) \cosh(1-t)}{\cosh(1)},$$

$$x^*(t) = \frac{\cosh(1-t)}{\cosh(1)}.$$

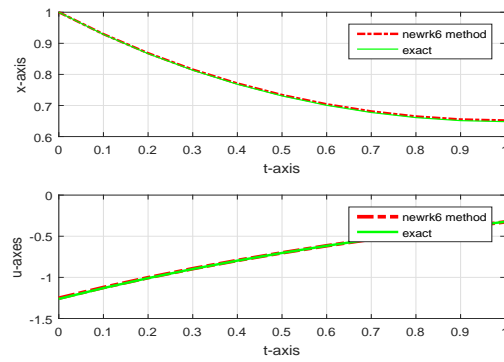
The state variable at the end point is $x(1) = 6.4805427366388e - 1$. Variable control endpoint is $u(1) = -3.24027136831e - 1$. Optimal value of the objective function is $J^* = 0.3807970779$. The proposed methods of Example 2 were determined as follows in MATLAB environment:

Table 8: Control values errors in Example 2 by using FBSM–RK6 and new proposed methods

t	h	FBSM–RK6	case 6	case 7	case 8	case 9	case 10
0.70	$\frac{1}{10}$	8.2317e-3	9.8758e-5	8.1761e-4	4.4379e-4	3.9543e-4	1.2035e-4
0.70	$\frac{1}{50}$	8.1160e-4	3.0223e-5	5.6542e-5	8.1127e-5	8.5984e-5	3.4750e-5
0.70	$\frac{1}{100}$	4.0487e-4	1.5985e-5	7.3752e-5	3.9860e-5	3.5044e-5	1.8260e-5
0.70	$\frac{1}{200}$	2.0203e-4	8.3745e-6	3.6430e-5	1.9590e-5	1.7183e-5	9.5155e-6



(a)



(b)

Figure 7: (a) Optimal state and control values of Example 2 by using FBSM–RK6.
 (b) Optimal state and control values of Example 2 by using new proposed method

Table 9: Errors of the performance index approximation in Example 2

h	FBSM–RK6	case 1	case 2	case 3	case 4	case 5
$\frac{1}{50}$	5.5920e-4	7.6989e-5	2.1442e-5	7.6139e-5	7.6159e-5	1.2794e-5
$\frac{1}{100}$	2.9187e-4	2.9588e-5	1.9712e-5	2.9168e-5	2.9173e-5	1.5522e-5
$\frac{1}{200}$	1.5038e-4	1.1235e-5	1.3437e-5	1.1026e-5	1.1027e-5	1.1376e-5
$\frac{1}{1000}$	3.0605e-5	1.8661e-6	3.0719e-6	1.8246e-6	1.8246e-6	2.6652e-6

Numerical results presented in Tables 7 and 8 indicate that each of the new ten suggested methods calculates the amount of control variable values much more accurately than the FBSM–RK6 method. Figure 7 also indicates that the new proposed method is exactly based on figure of analytical answer. For simplicity of reporting the results, one of diagrams was selected and drawn. The rest of figures are similar to each other. Numerical results presented in Tables 9 and 10 indicate that the estimated performance index of the new methods is more precise than those of the FBSM–RK6 methods.

Example 3. Consider the following OCP for a fixed T [14]:

$$\begin{aligned} \min_u \int_0^T (\int_0^t x(\eta) d\eta + (u(t))^2) dt \\ s.t. \quad x'(t) = -x(t) + u(t), \quad x(0) = a. \end{aligned}$$

For converting the problem into the standard form, we can add another state and obtain two-dimensional system as follows:

$$\begin{aligned} x_1(t) &= x(t), \\ x_2(t) &= \int_0^t x(\eta) d\eta, \end{aligned}$$

We redefine $x(t) := [x_1(t), x_2(t)]^T$. Thus, we have an OCP of the form:

Table 10: Errors of the performance index approximation in Example 2

h	FBSM–RK6	case 6	case 7	case 8	case 9	case 10
$\frac{1}{50}$	5.5920e-4	6.9979e-5	1.027e-4	3.6537e-5	2.7287e-5	7.4275e-5
$\frac{1}{100}$	2.9187e-4	2.6081e-5	6.0153e-5	2.7391e-5	2.2764e-5	2.8242e-5
$\frac{1}{200}$	1.5038e-4	9.4814e-6	3.3603e-5	1.73105e-5	1.4995e-5	1.0565e-5
$\frac{1}{1000}$	3.0605e-5	1.5153e-6	7.0964e-6	3.8519e-6	3.3889e-6	1.7326e-6

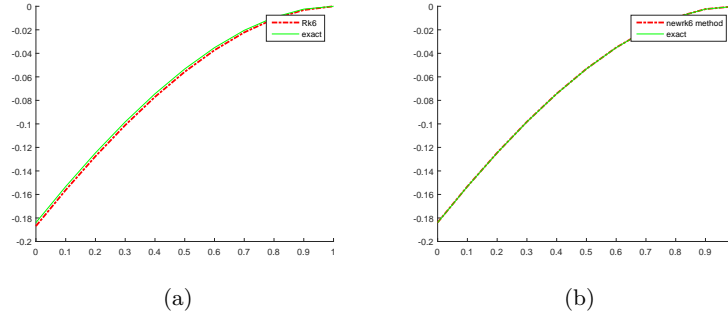


Figure 8: (a) Optimal state and control values of Example 3 using FBSM–RK6. (b) Optimal state and control values of Example 3 using new proposed method

$$\begin{aligned} \min_u \int_0^T (x_2(t) + u(t)^2) dt \\ \text{s.t., } x_1'(t) = -x(t) + u(t), \\ x_2'(t) = x_1(t), \\ x(0) = [a, 0]^T. \end{aligned}$$

Analytical solution of the problem is

$$\begin{aligned} H &= (x_2 + u^2) + \lambda_1(-x_1 + u) + \lambda_2 x_1, \\ \frac{\partial H}{\partial u} &= 2u + \lambda_1 = 0 \text{ at } u^* \Rightarrow u^* = -\frac{1}{2}\lambda_1, \lambda_1(T) = \lambda_2(T) = 0, \\ \Rightarrow \lambda_1' &= -\frac{\partial H}{\partial x_1} = \lambda_1 - \lambda_2, \lambda_2' = -\frac{\partial H}{\partial x_2} = -1, \\ \Rightarrow \lambda_1(t) &= -(t - T) - 1 + e^{(t-T)}, \\ u^*(t) &= -\frac{1}{2}\lambda_1(t) = \frac{1}{2}(1 + t - T - e^{(t-T)}). \end{aligned}$$

Numerical results for Example 3 are obtained and shown in Figure 8 and Tables 11 and 12.

Table 11: Control values errors in Example 3 using FBSM–RK6 and new proposed method

t	h	FBSM–RK6	case 1	case 2	case 3	case 4	case 5
0.80	$\frac{1}{10}$	1.3027e-3	4.3609e-5	1.9320e-4	4.5033e-5	4.5546e-5	1.7794e-4
0.80	$\frac{1}{50}$	2.4376e-4	1.6351e-6	2.9366e-5	1.8584e-6	1.9768e-6	2.7439e-5
0.80	$\frac{1}{100}$	1.2086e-4	4.1195e-7	1.4147e-5	2.2016e-7	5.8037e-7	1.3244e-5
0.80	$\frac{1}{200}$	6.0174e-5	1.0924e-7	6.9448e-6	1.6251e-7	1.9287e-7	6.5116e-6

Table 12: Control values errors in Example 3 by using FBSM–RK6 and new proposed methods

t	h	FBSM–RK6	case 6	case 7	case 8	case 9	case 10
0.80	$\frac{1}{10}$	1.3027e-3	5.4085e-5	3.2411e-4	2.1473e-4	2.0072e-4	4.8041e-5
0.80	$\frac{1}{50}$	2.4376e-4	3.5454e-6	5.1626e-5	3.4084e-5	3.1524e-5	2.3610e-6
0.80	$\frac{1}{100}$	1.2086e-4	1.3566e-6	2.5050e-5	1.6529e-5	1.5263e-5	7.6572e-7
0.80	$\frac{1}{200}$	6.0174e-5	5.7899e-7	1.2340e-5	8.1415e-6	7.5122e-6	2.8398e-7

Numerical results presented in Tables 11 and 12 indicate that each of the new ten suggested methods calculates the amount of control variable values much more accurately than the FBSM–RK6 method. Figure 8 also indicates that the figure of the new proposed method is quite matched on real answer and is much better than the FBSM–RK6 method. For simplicity of reporting the results, one of diagrams was selected and drawn. The rest of figures are similar to each other.

8 Conclusion

A new class of the 6th-order explicit hybrid methods was presented for which the 6th-order Runge–Kutta method is used as a predictor scheme to gain whole method of the same order, and the order of truncation errors was investigated for the explicit hybrid Runge–Kutta methods. The stability of the methods was discussed, and the results revealed that the stability regions of the proposed methods are wider compared to the 6th-order explicit Runge–Kutta method. Finally, three examples of OCPs were solved using MATLAB, FBSM scheme, and the presented methods and numerical results related to given examples were presented in Tables 3–12. According to the findings, it can be concluded that the new explicit hybrid methods have a good performance in accuracy and performance index approximation compared to the RK6 method.

References

1. Ameen, I., Hidan, M., Mostefaoui, Z., and Ali, H.M. *An efficient algorithms for solving the fractional optimal control of SIRV epidemic model with a combination of vaccination and treatment*, Chaos Solitons Fractals, 137 (2020), 109892, 12 pp.

2. Bhih, A.E., Ghazzali, R., Rhila, S.B., Rachik, M., and Laaroussi, A.E.A. *A discrete mathematical modeling and optimal control of the rumor propagation in online social network*, Discrete Dyn. Nat. Soc. 2020, Art. ID 4386476, 12 pp.
3. Duran, M.Q., Candelo, J.E., and Ortiz, J.S. *A modified backward/forward sweep-based method for reconfiguration of unbalanced distribution networks*, Int. J. Electr. Comput. Eng. (IJECE), 1(9) (2019), 85–101.
4. Ebadi, M. *Hybrid BDF methods for the numerical solution of ordinary differential equation*, Numer. Algorithms, 55(2010), 1–17.
5. Ebadi, M. *A class of multi-step methods based on a super-future points technique for solving IVPs*, Comput. Math. Appl. 61(2011), 3288–3297.
6. Ebadi, M. *Class 2+1 hybrid BDF-like methods for the numerical solution of ordinary differential equation*, Calcolo, 48(2011), 273–291.
7. Ebadi, M. *New class of hybrid BDF methods for the computation of numerical solution of IVPs*, Numer. Algorithms, 79(2018), 179–193.
8. Ebadi, M., Haghighi, A.R., Malih maleki, I., and Ebadian, A. *FBSM solution of optimal control problems using hybrid Runge–Kutta based methods*, J. Math. Ext. 15(4) (2021), In press.
9. Ebadi, M., Malih maleki, I., Haghighi, A.R., and Ebadian, A. *An explicit single-step method for numerical solution of optimal control problems*, Int. J. Ind. Math. 13(1) (2021), 71–89.
10. Jain, M.K. *Numerical solution of differential equations*, 2nd Edition, New Age International publishers, 2002.
11. Kheiri, H. and Jafari, M. *Optimal control of a fractional-order model for the HIV/AIDS epidemic*, Int. J. Biomath. 11(7) (2018), 1850086, 23 pp.
12. Kongjeen, Y., Bhummikittipich, K., Mithulananthan, N., Amiri, I. S., and Yupapin, P. *A modified backward and forward sweep method for micro-grid load flow analysis under different electric vehicle load mathematical models*, Electr. Pow. Syst. Res., 168(2019), 46–54.
13. Lenhart, S. and Workman, J.T. *Optimal control applied to biological models*, Chapman & Hall/CRC, Boca Raton, 2007.
14. Lewis, F.L., Vrabie, D.L., and Syrmos, W.L. *Optimal Control*, 2ed, John Wiley, sons Inc, 1995.
15. Lhous, M., Rachik, M., Laarabi, H., and Abdelhak, A. *Discrete mathematical modeling and optimal control of the marital status: the monogamous marriage case*, Adv. Difference Equ. 2017, Paper No. 339, 16 pp.

16. McAsey, M., Moua, L., and Han, W. *Convergence of the forward-backward sweep method in optimal control*, Comput. Optim. Appl., 53(2012), 207–226.
17. Moualeu, D.P., Weiser, M., Ehrig, R., and Deuffhard, P. *Optimal control for tuberculosis model with undetected cases in Cameroon*, Commun. Nonlinear Sci. Numer. Simul., 20(2015), 986–1003.
18. Rafiei, Z., Kafash, B., and Karbassi, S.M. *A computational method for solving optimal control problems and their applications*, Control and Optimization in Applied Mathematics, 2(1), (2017), 1–13.
19. Rodrigues, H.S., Teresa, M., Monteiro, T., and Torres, D.F.M. *Optimal control and numerical software: an overview*, Systems Theory: Perspectives, Applications and Developments , Nova Science publishers, 2014.
20. Rose, G.R. *Numerical methods for solving optimal control problems*, University of Tennessee, Knoxville, A Thesis for the master of science Degree, 2015.
21. Saleem, R., Habib, M., and Manaf, A. *Review of forward-backward sweep method for unbounded control problem with payoff term*, Sci. Int., 27(1) (2014), 69–72.
22. Sana, M., Saleem, R., Manaf, A., and Habib, M. *Varying forward backward sweep method using Runge-Kutta, Euler and Trapezoidal scheme as applied to optimal control problems*, Sci. Int. , 27(2015), 839–843.
23. Silveira, M., Nascimento, J.C., Marques, J.S., Marçal, A.R.S. *Comparison of segmentation methods for melanoma diagnosis in Dermoscopy images*, IEEE Journal of Selected Topics in Signal Processing, 3(1) (2009), 35–45.



The strict complementarity in linear fractional optimization

M. Mehdiloo*, K. Tone and M.B. Ahmadi

Abstract

As an important duality result in linear optimization, the Goldman–Tucker theorem establishes strict complementarity between a pair of primal and dual linear programs. Our study extends this result into the framework of linear fractional optimization. Associated with a linear fractional program, a dual program can be defined as the dual of the equivalent linear program obtained from applying the Charnes–Cooper transformation to the given program. Based on this definition, we propose new criteria for primal and dual optimality by showing that the primal and dual optimal sets can be equivalently modeled as the optimal sets of a pair of primal and dual linear programs. Then, we define the concept of strict complementarity and establish the existence of at least one, called *strict complementary*, pair of primal and dual optimal solutions such that in every pair of complementary variables, exactly one variable is positive and the other is zero. We geometrically interpret the strict complementarity in terms of the relative interiors of two sets that represent the primal and dual optimal sets in higher dimensions. Finally, using this interpretation, we develop two approaches for finding a strict complementary solution in linear fractional optimization. We illustrate our results with two numerical examples.

AMS subject classifications (2020): 90C32; 90C46; 49N15.

Keywords: Linear fractional optimization; Charnes–Cooper transformation; Duality; Strict complementarity.

*Corresponding author

Received 29 August 2020; revised 27 May 2021; accepted 29 May 2021

Mahmood Mehdiloo

Department of Mathematics and Applications, University of Mohaghegh Ardabili, Ardabil, Iran. e-mail: m.mehdiloozad@gmail.com, m.mehdiloo@uma.ac.ir

Kaoru Tone

National Graduate Institute for Policy Studies, Tokyo, Japan. e-mail: tone@grips.ac.jp

Mohammad Bagher Ahmadi

Department of Mathematics, College of Sciences, Shiraz University, Shiraz, Iran. e-mail: mbahmadi@shirazu.ac.ir

1 Introduction

A mathematical optimization problem is specified as a (primal) linear fractional program (LFP) when a linear fractional function (i.e., ratio of two affine functions) is optimized subject to a set of linear constraints on the given variables.¹ The linear fractional optimization frequently appears in a wide variety of real-world applications, including information theory, numerical analysis, game theory, cutting stock problems, shipping schedules, macroeconomic planning model, and so on. More details can be found in [1, 11, 24, 25] and references therein. It is also applied in the measurement of efficiency by data envelopment analysis; see, for example, [10, 15, 22, 26, 27, 28, 29, 30, 31] among others. Therefore, considerable research interest has been devoted to this branch of optimization.

The literature on the duality of linear fractional optimization associate various duals to the primal LFP. Chadha [7] suggested a dual in the form of a linear program (LP) and proved some duality statements directly. However, the constant scalars in the numerator and denominator of the primal objective function are assumed to be absent in their work. Chadha and Chadha [8] extended Chadha's results to the general case, where the constant scalars are taken into account. An interesting note regarding their impressive work is that their results can be deduced in an alternative way from the duality of linear optimization. In fact, an indirect approach for constructing a dual program is to transform the primal LFP into an equivalent problem that its dual can be constructed in the classical way; see [25]. By using this approach, it can be verified (as in Section 2) that the dual program proposed in [8] is nothing else than the dual of the equivalent LP resulting from applying the well-known transformation of Charnes and Cooper [9] to the primal LFP.

Though demonstrating the common complementary slackness condition between the primal LFP and its dual, Chadha and Chadha [8] did not investigate the strict complementarity between them. Furthermore, to the best of our knowledge, no other research exists on such investigation. Motivated by these, we extend an important duality result proved by Goldman and Tucker [12] from linear optimization to linear fractional optimization. The so-called Goldman–Tucker theorem establishes the strict complementarity between a pair of primal and dual LPs. It states that at least one, so-called strict complementary, pair of primal and dual optimal solutions exists such that the sum of each pair of complementary variables is positive. That is, in every pair of complementary variables, exactly one variable is positive and the other is zero; see, for example, [20] for more details on the theory and applications of strict complementarity in linear optimization.

¹ If the given objective function is optimized with no restrictions on the values of its variables, then the optimization problem is called *unconstrained*. Useful information on approaches developed for solving unconstrained optimization problems can be found in [3, 16, 21], among others.

As a complementary to the work of Chadha and Chadha [8], this paper shows that the primal and dual optimal sets can be equivalently modeled as the optimal sets of a pair of primal and dual LPs. Using this fact, we propose new criteria for primal and dual optimality in terms of the belongingness of these LPs' objective vectors to the binding polyhedral cones at primal and dual feasible solutions. Then we define the strict complementary slackness condition for an LFP and demonstrate the existence of a strict complementary solution. We also show that any strict complementary solution induces unique optimal partitions for the sets of indices of nonnegative decision variables.

To deal with the problem of finding a strict complementary solution, we equivalently represent the primal and dual optimal sets by two nonnegative polyhedral sets in higher dimensions, which are described only by equality *defining constraints*.² Then we geometrically interpret the strict complementarity by proving that any pair of relative interior points of these polyhedral sets is a strict complementary solution, and vice versa. Based on this interpretation, we turn the problem under consideration to the equivalent problem of identifying a maximal element of a nonnegative polyhedral set. Exploiting the recent work of Mehdiloozad et al. [20], who have addressed the latter problem, we develop two linear optimization approaches for finding a strict complementary solution.

The remainder of this paper is organized as follows. Section 2 provides the necessary background needed for the rest of the paper. Section 3 proposes new criteria for primal and dual optimality and illustrates them with a numerical example. Section 4 establishes the strict complementarity for LFPs. Section 5 proposes an LP for finding a maximal element of a nonnegative polyhedral set and, thereby, develops two approaches for finding a strict complementary solution. Section 6 illustrates these approaches by a numerical example. Section 7 contains concluding remarks and suggestions for future research. Appendix A provides the GAMS (General Algebraic Modeling System) code of our proposed approaches.

2 Background

2.1 Notation

Let \mathbb{R}^d denote the d -dimensional Euclidean space, and let \mathbb{R}_+^d denote its nonnegative orthant. We denote sets by uppercase calligraphic letters, vectors by boldface lowercase letters, and matrices by boldface uppercase letters. We

² A polyhedral set is said to be nonnegative if it is a subset of the nonnegative orthant of Euclidean space. By the “defining constraints” of such a polyhedral set, we refer to the constraints imposed other than the nonnegativity conditions.

denote the cardinality of a set \mathcal{S} by $\text{Card}(\mathcal{S})$. By convention, all vectors are column vectors. The superscript \top denotes the transpose of a vector or matrix.

Vectors $\mathbf{0}$ and $\mathbf{1}$ are vectors all components of that are equal to 0 and 1, respectively. The dimensions of these vectors are clear from the context in which they are used. For simplicity, the notation $(\mathbf{a}; \mathbf{b}) \in \mathbb{R}^{d+d'}$ is used to show the column vector obtained by adding vector $\mathbf{b} \in \mathbb{R}^{d'}$ below the vector $\mathbf{a} \in \mathbb{R}^d$. For vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$, the inequality $\mathbf{a} \geq \mathbf{b}$ (resp., $\mathbf{a} > \mathbf{b}$) means that $a_i \geq b_i$ (resp., $a_i > b_i$) for all $i = 1, \dots, d$.

Matrix $\mathbf{0}$ is the matrix all components of that are equal to 0, and matrix \mathbf{I} is the identity matrix. The dimensions of these matrices are clear from the context in which they are used. We denote the i th ($i = 1, \dots, d$) row and the j th ($j = 1, \dots, d'$) column of a $d \times d'$ matrix \mathbf{A} by \mathbf{a}^i and \mathbf{a}_j , respectively. In particular, we use the notation \mathbf{e}_j to denote the j th column of the identity matrix of size $d \times d$, that is, $\mathbf{e}_j = (0, \dots, \underset{j\text{th}}{1}, \dots, 0)^\top \in \mathbb{R}^d$ for $j = 1, \dots, d$.

Recall from [23] that the relative interior of a subset \mathcal{X} of \mathbb{R}^d , denoted by $\text{ri}(\mathcal{X})$, is defined as the interior we get when \mathcal{X} is regarded as a subset of its affine hull, denoted by $\text{aff}(\mathcal{X})$. Formally,

$$\text{ri}(\mathcal{X}) = \{\mathbf{x}^o \in \mathcal{X} : \mathcal{N}_\varepsilon(\mathbf{x}^o) \cap \text{aff}(\mathcal{X}) \subseteq \mathcal{X} \text{ for some } \varepsilon > 0\},$$

where $\mathcal{N}_\varepsilon(\mathbf{x}^o) = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x} - \mathbf{x}^o\| < \varepsilon\}$.

Recall also from [20] that any convex (and, in particular, polyhedral) subset of \mathbb{R}_+^d is called a nonnegative convex (polyhedral) set. Additionally, any element of a nonnegative convex set is said to be maximal, if the number of its positive components is maximum. We denote the support of a nonnegative vector $\mathbf{a} \in \mathbb{R}_+^d$ by $\text{supp}(\mathbf{a})$, that is, $\text{supp}(\mathbf{a}) = \{i \in \{1, \dots, d\} : a_i > 0\}$. We also denote by $\text{me}(\mathcal{X})$ the set of all maximal elements of \mathcal{X} , that is, $\text{me}(\mathcal{X}) = \underset{\mathbf{x} \in \mathcal{X}}{\text{argmax}} \text{Card}(\text{supp}(\mathbf{x}))$.

2.2 Linear fractional program

A function of variables is said to be *linear fractional* if both its numerator and denominator are affine functions of the given variables. A mathematical optimization problem that optimizes a linear fractional objective function subject to a set of linear constraints is called as a linear fractional program (LFP). Formally, the general form³ of the primal LFP is defined as

³ The *standard* form of the primal LFP results from (1) by replacing the inequality sign “ \leq ” in (1b) by the equality sign; see [1, Section 1.3].

$$\max f(\mathbf{x}) = \frac{\mathbf{c}^\top \mathbf{x} + \alpha}{\mathbf{d}^\top \mathbf{x} + \beta} \quad (1a)$$

subject to

$$\mathbf{A}\mathbf{x} \leq \mathbf{b}, \quad (1b)$$

$$\mathbf{x} \geq \mathbf{0}, \quad (1c)$$

where $\mathbf{x} \in \mathbb{R}^n$ is the vector of decision variables, $\mathbf{c} \in \mathbb{R}^n$ and $\mathbf{d} \in \mathbb{R}^n$ are, respectively, numerator and denominator vectors of objective function, $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}$ are objective scalars, \mathbf{A} is an $m \times n$ constraint matrix, and $\mathbf{b} \in \mathbb{R}^m$ is the right-hand side vector.

Let \mathcal{S} denote the feasible set of program (1), which is clearly a polyhedral set in \mathbb{R}^n . To ensure that the function f is well-defined on \mathcal{S} , it is assumed that its denominator maintains a constant sign on \mathcal{S} . Without loss of generality, we assume that $\mathbf{d}^\top \mathbf{x} + \beta > 0$ for all $\mathbf{x} \in \mathcal{S}$. Then the objective function f is both quasi-convex and quasi-concave over \mathcal{S} and, therefore, every local maximum is a global maximum (see, e.g., [3]). To guarantee the occurrence of finite optimality for program (1), we also assume that \mathcal{S} is regular (i.e., nonempty and bounded).

An effective approach for solving program (1) is to transform it into an equivalent LP by the well-known Charnes–Cooper transformation [9]. In fact, if we define $t = \frac{1}{\mathbf{d}^\top \mathbf{x} + \beta}$ and $\bar{\mathbf{x}} = t\mathbf{x}$, then multiplying both sides of (1b) by t converts program (1) to the following LP:

$$\max \mathbf{c}^\top \bar{\mathbf{x}} + \alpha t \quad (2a)$$

subject to

$$\mathbf{A}\bar{\mathbf{x}} - \mathbf{b}t \leq \mathbf{0}, \quad (2b)$$

$$\mathbf{d}^\top \bar{\mathbf{x}} + \beta t = 1, \quad (2c)$$

$$\bar{\mathbf{x}} \geq \mathbf{0}, t \geq 0. \quad (2d)$$

Let $\bar{\mathcal{S}}$ be the feasible set of program (2). Because $t > 0$ for all $(\bar{\mathbf{x}}; t) \in \bar{\mathcal{S}}$,⁴ the following implication between the feasible solutions of programs (1) and (2) is established:

$$(\bar{\mathbf{x}}; t) \in \bar{\mathcal{S}} \quad \Rightarrow \quad \frac{1}{t}\bar{\mathbf{x}} \in \mathcal{S}, t > 0. \quad (3)$$

Especially, if $(\bar{\mathbf{x}}^*; t^*)$ is an optimal solution to program (2), then $\frac{1}{t^*}\bar{\mathbf{x}}^*$ is an optimal solution to program (1) (see, e.g., [1, p. 57]).

⁴ Indeed, if $t = 0$ for some $(\bar{\mathbf{x}}; t) \in \bar{\mathcal{S}}$, then it follows from (2b)–(2c) that $\mathbf{A}\bar{\mathbf{x}} \leq \mathbf{0}$ and $\bar{\mathbf{x}} \neq \mathbf{0}$. This means that the vector $\bar{\mathbf{x}}$ is a recession direction of the feasible set \mathcal{S} , thereby contradicting the regularity assumption. Therefore, $t > 0$ for all $(\bar{\mathbf{x}}; t) \in \bar{\mathcal{S}}$.

2.3 Dual of linear fractional program

To state the dual to LFP (1), let the vector $\mathbf{y} \in \mathbb{R}^m$ be dual to (2b) and the scalar z be dual to (2c). Then, by the duality of linear optimization, the dual to LP (2) is stated as follows:⁵

$$\min \quad g(\mathbf{y}; z) = z \quad (4a)$$

subject to

$$\mathbf{A}^\top \mathbf{y} + \mathbf{d}z \geq \mathbf{c}, \quad (4b)$$

$$-\mathbf{b}^\top \mathbf{y} + \beta z = \alpha, \quad (4c)$$

$$\mathbf{y} \geq \mathbf{0}, z \text{ sign free.} \quad (4d)$$

Observe that program (4) is nothing else than the LP introduced in [8] as the dual of program (1). We denote by \mathcal{D} the feasible set of this program.

Throughout this paper, LP (4) is defined to be the dual of LFP (1). The next three theorems demonstrate the duality relationships between programs (1) and (4).

Theorem 1 (Weak duality). [8] For any $\mathbf{x} \in \mathcal{S}$ and any $(\mathbf{y}; z) \in \mathcal{D}$, we have $f(\mathbf{x}) \leq g(\mathbf{y}; z)$.

Theorem 2 (Optimality criterion). [8] If the feasible solutions $\mathbf{x} \in \mathcal{S}$ and $(\mathbf{y}; z) \in \mathcal{D}$ satisfy $f(\mathbf{x}) = g(\mathbf{y}; z)$, then they are optimal solutions to programs (1) and (4), respectively.

Theorem 3 (Strong duality). [8] If \mathbf{x}^* is an optimal solution to program (1), then there exists some optimal solution $(\mathbf{y}^*; z^*)$ to program (4) such that $f(\mathbf{x}^*) = g(\mathbf{y}^*; z^*)$.

The following result gives a necessary and sufficient optimality condition, called *complementary slackness condition* (CSC), in terms of the complementarity of the primal and dual feasible solutions.

Theorem 4. [8] Feasible solutions $\mathbf{x}^* \in \mathcal{S}$ and $(\mathbf{y}^*; z^*) \in \mathcal{D}$ are optimal if and only if they fulfill the following conditions:

$$\mathbf{v}^{*\top} \mathbf{x}^* = \mathbf{u}^{*\top} \mathbf{y}^* = 0, \quad (5)$$

where $\mathbf{u}^* = \mathbf{b} - \mathbf{A}\mathbf{x}^*$ and $\mathbf{v}^* = \mathbf{A}^\top \mathbf{y}^* + \mathbf{d}z^* - \mathbf{c}$.

From Theorem 4, the pairs (x_j, v_j) , $j = 1, \dots, n$, and (u_i, y_i) , $i = 1, \dots, m$, are called complementary variables.

⁵ Note that the inequality constraint $-\mathbf{b}^\top \mathbf{y} + \beta z \geq \alpha$ has been replaced with its equality form in program (4), because the optimal value of its corresponding dual variable in program (2), t , is always positive.

3 Our criteria for primal and dual optimality

In this section, we propose new criteria for the optimality of LFP (1) and its dual (4), and present their geometrical interpretations.

Denote by f^* the optimal objective value of program (1), and let $\mathbf{d}^* = f^* \mathbf{d}$ and $\beta^* = f^* \beta$. Then the set of all optimal solutions of program (1) can be defined by conditions (1b)–(1c) and the additional equality requiring that the objective function of program (1) to be equal to f^* . Equivalently, this set is stated as follows:

$$\mathcal{X}^* = \{\mathbf{x} \in \mathbb{R}^n : (\mathbf{c} - \mathbf{d}^*)^\top \mathbf{x} = -\alpha + \beta^*, \mathbf{A}\mathbf{x} \leq \mathbf{b}, \mathbf{x} \geq \mathbf{0}\}.$$

Similarly, an equivalent statement of the optimal set of program (4) is

$$\{(\mathbf{y}; z^*) \in \mathbb{R}^{m+1} : \mathbf{b}^\top \mathbf{y} = -\alpha + \beta^*, \mathbf{A}^\top \mathbf{y} \geq \mathbf{c} - \mathbf{d}^*, \mathbf{y} \geq \mathbf{0}\},$$

where z^* denotes the optimal objective value of program (4) and is equal to f^* . Observe that the last components of all optimal solutions of program (4) are equal to z^* . Therefore, without losing anything, we can remove the last dimension of the optimal set of program (4) by projecting it onto the space of \mathbf{y} -variables. This results the following set:

$$\mathcal{Y}^* = \{\mathbf{y} \in \mathbb{R}^m : \mathbf{b}^\top \mathbf{y} = -\alpha + \beta^*, \mathbf{A}^\top \mathbf{y} \geq \mathbf{c} - \mathbf{d}^*, \mathbf{y} \geq \mathbf{0}\},$$

which will be loosely referred to as the optimal set of program (4).

It is clear that the nonempty optimal sets \mathcal{X}^* and \mathcal{Y}^* are polyhedral subsets of \mathbb{R}_+^n and \mathbb{R}_+^m , respectively. The next result shows that these sets are interestingly the optimal sets of the LP

$$\begin{aligned} & \max \quad (\mathbf{c} - \mathbf{d}^*)^\top \mathbf{x} \\ & \text{subject to} \\ & \quad (1b) - (1c), \end{aligned} \tag{6}$$

and its dual

$$\min \quad \mathbf{b}^\top \mathbf{y} \tag{7a}$$

subject to

$$\mathbf{A}^\top \mathbf{y} \geq \mathbf{c} - \mathbf{d}^*, \tag{7b}$$

$$\mathbf{y} \geq \mathbf{0}. \tag{7c}$$

Theorem 5. Let \mathcal{F}_P^* and \mathcal{F}_D^* be the optimal sets of programs (6) and (7), respectively. Then, $\mathcal{F}_P^* = \mathcal{X}^*$ and $\mathcal{F}_D^* = \mathcal{Y}^*$.

Proof. Let $\hat{\mathbf{x}} \in \mathcal{X}^*$ and $\hat{\mathbf{y}} \in \mathcal{Y}^*$. Then $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ are, respectively, feasible solutions to LPs (6) and (7) such that $(\mathbf{c} - \mathbf{d}^*)^\top \hat{\mathbf{x}} = \mathbf{b}^\top \hat{\mathbf{y}}$. By the optimality

criterion theorem of linear optimization, it follows that $\hat{\mathbf{x}} \in \mathcal{F}_P^*$ and $\hat{\mathbf{y}} \in \mathcal{F}_D^*$. Therefore, $\mathcal{X}^* \subseteq \mathcal{F}_P^*$ and $\mathcal{Y}^* \subseteq \mathcal{F}_D^*$.

Conversely, let $\hat{\mathbf{x}} \in \mathcal{F}_P^*$ and $\hat{\mathbf{y}} \in \mathcal{F}_D^*$. By the strong duality theorem of linear optimization, we have $(\mathbf{c} - \mathbf{d}^*)^\top \hat{\mathbf{x}} = \mathbf{b}^\top \hat{\mathbf{y}}$. Because $\mathcal{X}^* \neq \emptyset$ and $\mathcal{X}^* \subseteq \mathcal{F}_P^*$, the weak duality theorem of linear optimization implies that $-\alpha + \beta^* \leq \mathbf{b}^\top \hat{\mathbf{y}}$. Similarly, it follows from $\mathcal{Y}^* \neq \emptyset$ and $\mathcal{Y}^* \subseteq \mathcal{F}_D^*$ that $(\mathbf{c} - \mathbf{d}^*)^\top \hat{\mathbf{x}} \leq -\alpha + \beta^*$. Consequently, we have $(\mathbf{c} - \mathbf{d}^*)^\top \hat{\mathbf{x}} = \mathbf{b}^\top \hat{\mathbf{y}} = -\alpha + \beta^*$. Therefore, $\hat{\mathbf{x}} \in \mathcal{X}^*$ and $\hat{\mathbf{y}} \in \mathcal{Y}^*$, which, respectively, imply $\mathcal{F}_P^* \subseteq \mathcal{X}^*$ and $\mathcal{F}_D^* \subseteq \mathcal{Y}^*$. \square

Associated with any feasible solution of an LP, the *binding cone* is defined as the convex cone generated by the gradients of all constraints that are binding (active) at that solution. Recall from linear optimization that a feasible solution to an LP is optimal if and only if its corresponding binding cone includes the gradient of the objective function. Based on this fact, we apply Theorem 5 to provide necessary and sufficient geometrical conditions for feasible solutions of LFP (1) and its dual (4) to be optimal.

Let \mathbf{x}^* be a feasible solution to LFP (1), and let \mathcal{G}_P^* denote the union of the gradients of all binding constraints at \mathbf{x}^* , that is,

$$\mathcal{G}_P^* = \left\{ (\mathbf{a}^i)^\top \in \mathbb{R}^n : \mathbf{a}^i \mathbf{x}^* = b_i \right\} \cup \left\{ -\mathbf{e}_j \in \mathbb{R}^n : x_j^* = 0 \right\}.$$

Furthermore, denote by \mathcal{B}_P^* the binding polyhedral cone generated by \mathcal{G}_P^* , that is, $\mathcal{B}_P^* = \text{cone}(\mathcal{G}_P^*)$, where the operator “cone” denotes the conical hull. Because the feasible regions of programs (1) and (6) are equal, \mathbf{x}^* is a feasible solution of LP (6). Therefore, the next corollary follows immediately from Theorem 5.

Corollary 1. Let $\mathbf{x}^* \in \mathcal{S}$. Then, $\mathbf{x}^* \in \mathcal{X}^*$ if and only if $\mathbf{c} - \mathbf{d}^* \in \mathcal{B}_P^*$.

Similarly, let $(\mathbf{y}^*; z^*)$ be a feasible solution to program (4). Additionally, assume that $\mathcal{B}_D^* = \text{cone}(\mathcal{G}_D^*)$, where

$$\mathcal{G}_D^* = \left\{ \mathbf{a}_j \in \mathbb{R}^m : \mathbf{a}_j^\top \mathbf{y}^* = c_j - d_j^* \right\} \cup \left\{ \mathbf{e}_i \in \mathbb{R}^m : y_i^* = 0 \right\}.$$

Then we obtain the following corollary as a consequence of Theorem 5.

Corollary 2. Let $(\mathbf{y}^*; z^*) \in \mathcal{D}$. Then, $\mathbf{y}^* \in \mathcal{Y}^*$ if and only if $\mathbf{b} \in \mathcal{B}_D^*$.

We now present a numerical example verifying Corollaries 1 and 2.

Example 1. Consider the following LFP:

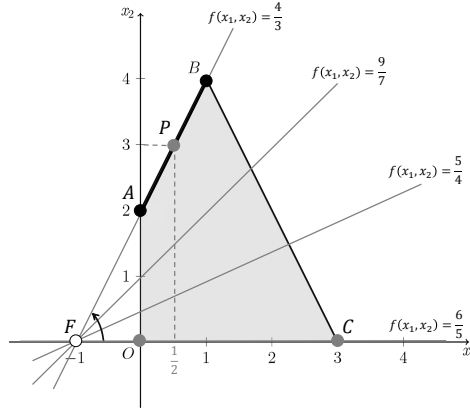


Figure 1: Multiple optimal solutions of program (8)

$$\max f(x_1, x_2) = \frac{6x_1 + 3x_2 + 6}{5x_1 + 2x_2 + 5} \quad (8a)$$

subject to

$$2x_1 + x_2 \leq 6, \quad (8b)$$

$$-2x_1 + x_2 \leq 2, \quad (8c)$$

$$x_1, x_2 \geq 0. \quad (8d)$$

A graphical approach for finding optimal solution(s) of two-dimensional LFPs is to rotate the level-line around its focus point in positive direction (i.e., counterclockwise).⁶ Figure 1 illustrates an application of this approach to program (8). The feasible region of program (8) in two dimensions x_1 and x_2 is the bounded polyhedral set $OABC$ (shaded in gray), and the focus point is $F = (-1, 0)$. Therefore, the optimal objective value of program (8) is $f^* = \frac{4}{3}$. Additionally, the set of all optimal solutions to program (8) is the segment AB , which is stated below as all convex combinations of the two extreme points A and B of the feasible region:

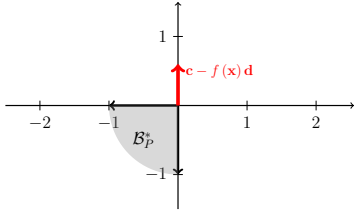
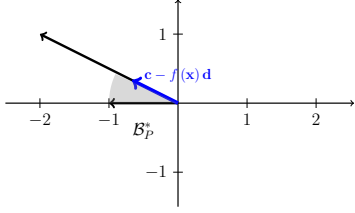
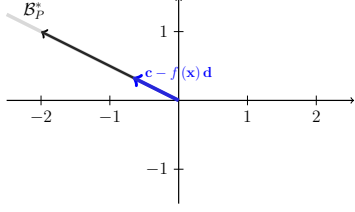
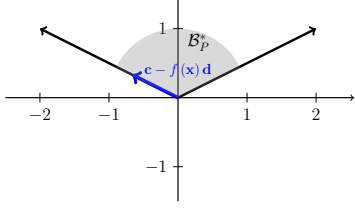
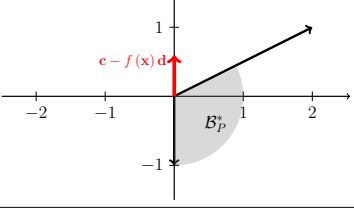
$$\mathcal{X}^* = \{\mathbf{x}^\lambda \in \mathbb{R}^2: (x_1^\lambda, x_2^\lambda) = \lambda(0, 2) + (1 - \lambda)(1, 4), \lambda \in [0, 1]\}.$$

Table 1 presents the geometrical investigation of the proposed condition of primal optimality in Corollary 1 at four extreme points O , A , B , and C , and one nonextreme point P of the feasible region of program (8). As expected, the condition holds at the optimal points A , B , and P , but not at the nonoptimal points O and C .

The dual to program (8) is the following LP:

⁶ More details on graphical solution of LFPs involving only two variables can be found in [1, Chapter 3].

Table 1: The proposed criterion of primal optimality for program (8)

	\mathbf{x}	$f(\mathbf{x})$	$\mathbf{c} - f(\mathbf{x})\mathbf{d}$	\mathcal{G}_P^*	Belongingness of $\mathbf{c} - f(\mathbf{x})\mathbf{d}$ to \mathcal{B}_P^*
O	$\begin{pmatrix} 0 \\ 0 \end{pmatrix}$	$\frac{6}{5}$	$\begin{pmatrix} 0 \\ \frac{3}{5} \end{pmatrix}$	$\begin{pmatrix} -1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ -1 \end{pmatrix}$	
A	$\begin{pmatrix} 0 \\ 2 \end{pmatrix}$	$\frac{4}{3}$	$\begin{pmatrix} -\frac{2}{3} \\ \frac{1}{3} \end{pmatrix}$	$\begin{pmatrix} -2 \\ 1 \end{pmatrix}, \begin{pmatrix} -1 \\ 0 \end{pmatrix}$	
P	$\begin{pmatrix} \frac{1}{2} \\ 3 \end{pmatrix}$	$\frac{4}{3}$	$\begin{pmatrix} -\frac{2}{3} \\ \frac{1}{3} \end{pmatrix}$	$\begin{pmatrix} -2 \\ 1 \end{pmatrix}$	
B	$\begin{pmatrix} 1 \\ 4 \end{pmatrix}$	$\frac{4}{3}$	$\begin{pmatrix} -\frac{2}{3} \\ \frac{1}{3} \end{pmatrix}$	$\begin{pmatrix} 2 \\ 1 \end{pmatrix}, \begin{pmatrix} -2 \\ 1 \end{pmatrix}$	
C	$\begin{pmatrix} 3 \\ 0 \end{pmatrix}$	$\frac{6}{5}$	$\begin{pmatrix} 0 \\ \frac{3}{5} \end{pmatrix}$	$\begin{pmatrix} 0 \\ -1 \end{pmatrix}, \begin{pmatrix} 2 \\ 1 \end{pmatrix}$	

$$\min z \quad (9a)$$

subject to

$$2y_1 - 2y_2 + 5z \geq 6, \quad (9b)$$

$$y_1 + y_2 + 2z \geq 3, \quad (9c)$$

$$-6y_1 - 2y_2 + 5z = 6, \quad (9d)$$

$$y_1, y_2 \geq 0, z \text{ sign free.} \quad (9e)$$

By Theorem 3, the optimal objective value of the dual program (9) is equal to that of the primal LFP (8), so $z^* = \frac{4}{3}$. By Theorem 4, $y_1^* = 0$ for any optimal solution (y_1^*, y_2^*) of (9) because the point A is an optimal solution to program (8) for which the inequality constraint (8b) is strict. Additionally, both constraints (9b) and (9c) must be binding at optimality because point B with both positive components is an optimal solution to program (8). Taking these into account, it follows from the constraints of program (9) that $y_2^* = \frac{1}{3}$ for any optimal solution. Therefore, $(y_1^*, y_2^*, z^*) = (0, \frac{1}{3}, \frac{4}{3})$ is the unique optimal solution of LP (9). The stated facts are observable from Figure 2, which draws the feasible region of program (9) in three dimensions y_1, y_2 , and z as the section $LMNK$ of the two-dimensional hyperplane $\mathcal{H} = \{(\mathbf{y}; z) \in \mathbb{R}^3: -6y_1 - 2y_2 + 5z = 6\}$.

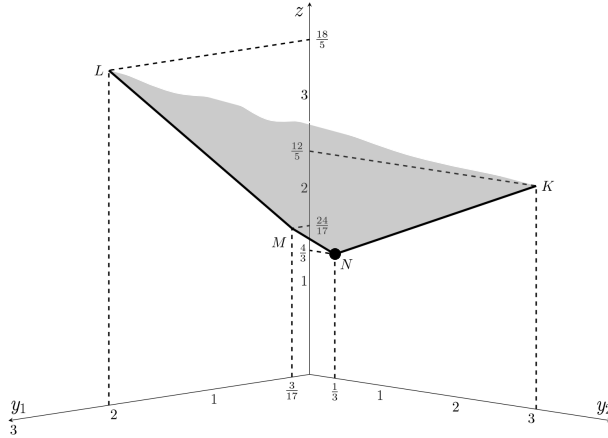


Figure 2: Unique optimal solution of program (9)

Observe that projecting the unique optimal solution of program (4) onto the space of \mathbf{y} -variables follows that $\mathcal{Y}^* = \{(0, \frac{1}{3})^\top\}$. As an illustration to Theorem 5, Figure 3 shows the singleton set \mathcal{Y}^* to be the optimal extreme point of the following LP:

$$\min 6y_1 + 2y_2 \quad (10a)$$

subject to

$$2y_1 - 2y_2 \geq \frac{-2}{3}, \quad (10b)$$

$$y_1 + y_2 \geq \frac{1}{3}, \quad (10c)$$

$$y_1, y_2 \geq 0. \quad (10d)$$

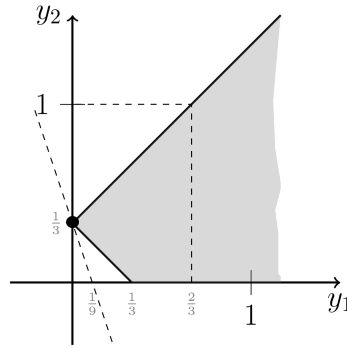
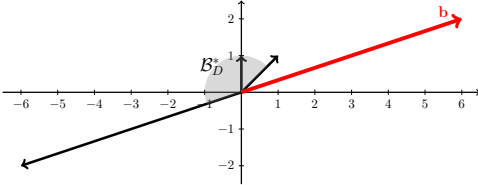
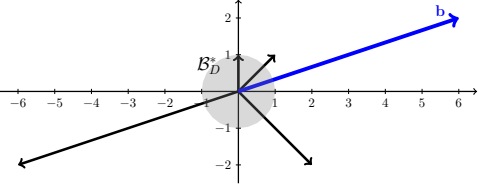


Figure 3: Representing \mathcal{Y}^* as the unique optimal solution of program (10)

Table 2 geometrically investigates the proposed condition of dual optimality in Corollary 2 at the two extreme points M and N of the feasible region of program (9). While the condition is met by the optimal point N , it is not true at the nonoptimal point M . This verifies that the projection of the optimal point N onto the space of \mathbf{y} -variables is in \mathcal{Y}^* .

Remark 1. Corollary 1 suggests a geometrical approach for finding optimal solution(s) of program (1). It states that the optimality of any feasible point in \mathcal{S} is equivalent to satisfying the condition given in Corollary 1. Therefore, taking into account the fact that the finite optimum must occur at some extreme points of \mathcal{S} , optimal solution(s) of program (1) can be found by examining the proposed condition only at extreme points of \mathcal{S} . (A similar approach for finding optimal solution(s) of program (4) can be devised based on Corollary 2.) It is important to note that this graphical approach does not require the enumeration of all extreme points of the feasible region.

Table 2: The proposed criterion of primal optimality for program (8)

	\mathbf{y}	z	\mathcal{G}_D^*	Belongingness of \mathbf{b} to \mathcal{B}_D^*
M	$\begin{pmatrix} 3 \\ 17 \\ 0 \end{pmatrix}$	$\frac{24}{17}$	$\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} -6 \\ -2 \end{pmatrix}$	
N	$\begin{pmatrix} 0 \\ \frac{1}{3} \end{pmatrix}$	$\frac{4}{3}$	$\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} -6 \\ -2 \end{pmatrix}, \begin{pmatrix} 2 \\ -2 \end{pmatrix}$	

4 Strict complementarity

4.1 Strict complementary solution

By Theorem 4, the CSC requires only that the product of each pair of complementary variables is zero at optimality. Therefore, not only one of the complementary variables must be zero at optimality, but also both are allowed to take simultaneously zero optimal values. It means that the CSC does not imply the positivity of pairwise sum of the complementary variables. If such positivity holds for a pair of optimal solutions for the primal LFP and its dual, then in every pair of complementary variables, exactly one variable is positive and the other is zero. Calling this property as *strict complementarity*, we present the following definition.

Definition 1. Feasible solutions $\mathbf{x}^{*s} \in \mathcal{S}$ and $(\mathbf{y}^{*s}; z^{*s}) \in \mathcal{D}$ satisfy the *strict complementary slackness condition* (SCSC), if they fulfill the following conditions in addition to the conditions given in (5):

$$\mathbf{x}^{*s} + \mathbf{v}^{*s} > \mathbf{0}, \quad \mathbf{u}^{*s} + \mathbf{y}^{*s} > \mathbf{0}. \quad (11)$$

It is clear that feasible solutions to programs (1) and (4) that satisfy the SCSC are optimal. We refer to such a pair of solutions as a *strict complementary solution* and denote it by $(\mathbf{x}^{*s}, \mathbf{y}^{*s})$. By the next result, we prove the existence of such a strict complementary solution.

Theorem 6. The following statements are true:

- (i) Any strict complementary solution to LPs (6) and (7) is a strict complementary solution to programs (1) and (4).
- (ii) There exists at least one strict complementary solution to LFP (1) and its dual (4).

Proof. Part (i) Let $(\mathbf{x}^{*s}, \mathbf{y}^{*s})$ be a strict complementary solution to LPs (6) and (7). Then, by the definition of strict complementary in linear optimization, \mathbf{x}^{*s} and \mathbf{y}^{*s} are, respectively, optimal solutions to these LPs, such that

$$\mathbf{v}^{*s\top} \mathbf{x}^{*s} = \mathbf{u}^{*s\top} \mathbf{y}^{*s} = 0, \quad \mathbf{x}^{*s} + \mathbf{v}^{*s} > \mathbf{0}, \quad \mathbf{u}^{*s} + \mathbf{y}^{*s} > \mathbf{0}, \quad (12)$$

where \mathbf{u}^{*s} and \mathbf{v}^{*s} are, respectively, the slack vectors added to the inequality constraints in \mathcal{X}^* and \mathcal{Y}^* .

By Theorem 5, \mathbf{x}^{*s} and $(\mathbf{y}^{*s}; z^{*s})$ are optimal solutions to programs (1) and (4), respectively. Furthermore, it follows from (12) that these solutions meet the SCSC in the sense of Definition 1. Therefore, $(\mathbf{x}^{*s}, \mathbf{y}^{*s})$ is a strict complementary solution to programs (1) and (4).

Part (ii) Because a finite optimum occurs for LPs (6) and (7), the Goldman–Tucker theorem implies the existence of a strict complementary solution to these LPs, which is a strict complementary solution to programs (1) and (4) by part (i) of the theorem. \square

4.2 Optimal partitions

Let $(\mathbf{x}^{*s}, \mathbf{y}^{*s})$ be a strict complementary solution to programs (1) and (4). Then, the supports of vectors \mathbf{x}^{*s} and \mathbf{v}^{*s} are disjoint and their union is equal to the index set $\{1, \dots, n\}$. Similarly, the supports of vectors \mathbf{y}^{*s} and \mathbf{u}^{*s} form a partition for the index set $\{1, \dots, m\}$. Formally, we can write

$$\begin{aligned} \text{supp}(\mathbf{x}^{*s}) \cap \text{supp}(\mathbf{v}^{*s}) &= \emptyset, & \text{supp}(\mathbf{x}^{*s}) \cup \text{supp}(\mathbf{v}^{*s}) &= \{1, \dots, n\}; \\ \text{supp}(\mathbf{u}^{*s}) \cap \text{supp}(\mathbf{y}^{*s}) &= \emptyset, & \text{supp}(\mathbf{u}^{*s}) \cup \text{supp}(\mathbf{y}^{*s}) &= \{1, \dots, m\}. \end{aligned} \quad (13)$$

We call the above partitions as the *optimal partitions* induced for programs (1) and (4). By the next result, we show that these partitions, being independent from the given strict complementary solution, are unique.

Theorem 7. The optimal partitions induced for programs (1) and (4) as in (13) are the same across all strict complementary solutions and are, therefore, unique.

Proof. By contradiction, let $(\mathbf{x}^1, \mathbf{y}^1)$ and $(\mathbf{x}^2, \mathbf{y}^2)$ be two distinct strict complementary solutions such that $x_j^1 > 0$ and $x_j^2 = 0$ for some $\hat{j} \in \{1, \dots, n\}$. Then, $v_j^1 = 0$ and $v_j^2 > 0$. Because both optimal sets of programs (1) and (4)

are convex, $\frac{1}{2}(\mathbf{x}^1, \mathbf{y}^1) + \frac{1}{2}(\mathbf{x}^2, \mathbf{y}^2)$ must be an optimal solution such that $\frac{1}{2}(\mathbf{u}^1 + \mathbf{u}^2)$ and $\frac{1}{2}(\mathbf{v}^1 + \mathbf{v}^2)$ are, respectively, its primal and dual slack vectors. For this solution, we have the contradiction (with the CSC) that $\frac{1}{2}(x_j^1 + x_j^2) > 0$ and $\frac{1}{2}(v_j^1 + v_j^2) > 0$. Therefore, the optimal partition of $\{1, \dots, n\}$ is unique. The uniqueness of the optimal partition of $\{1, \dots, m\}$ follows from a similar argument. \square

It is worth noting that the optimal partitions can be useful in situations, where knowing the positivity of a variable in some optimal solution of an LFP is concerned with. For example, while the slack-based measure (SBM) model of Tone [26] is used for the measurement of efficiency in the field of data envelopment analysis, the global reference set (peer group) of an inefficient decision making unit can be identified by the optimal partition of the index set of intensity vector.⁷

4.3 Geometrical interpretation

We begin this section by recalling the following definition from [4].

Definition 2. Let $\mathcal{S} \subset \mathbb{R}^d$. A subset \mathcal{S}^+ of $\mathbb{R}^{d+d'}$ is a *representing set* for \mathcal{S} , if its projection onto the space of \mathbf{x} -variables is exactly \mathcal{S} , that is, $\mathbf{x} \in \mathcal{S}$ if and only if there exists some $\mathbf{s} \in \mathbb{R}^{d'}$ such that $(\mathbf{x}; \mathbf{s}) \in \mathcal{S}^+$:

$$\mathcal{S} = \left\{ \mathbf{x} \in \mathbb{R}^d : (\mathbf{x}; \mathbf{s}) \in \mathcal{S}^+ \text{ for some } \mathbf{s} \in \mathbb{R}^{d'} \right\}.$$

By adding slack vectors to the inequality constraints of \mathcal{X}^* and \mathcal{Y}^* , we define the following nonnegative polyhedral sets:

$$\begin{aligned} \mathcal{X}^{*+} &= \left\{ (\mathbf{x}; \mathbf{u}) \in \mathbb{R}^{n+m} : (\mathbf{c} - \mathbf{d}^*)^\top \mathbf{x} = -\alpha + \beta^*, \mathbf{A}\mathbf{x} + \mathbf{u} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}, \mathbf{u} \geq \mathbf{0} \right\}, \\ \mathcal{Y}^{*+} &= \left\{ (\mathbf{y}; \mathbf{v}) \in \mathbb{R}^{m+n} : \mathbf{b}^\top \mathbf{y} = -\alpha + \beta^*, \mathbf{A}^\top \mathbf{y} - \mathbf{v} = \mathbf{c} - \mathbf{d}^*, \mathbf{y} \geq \mathbf{0}, \mathbf{v} \geq \mathbf{0} \right\}. \end{aligned}$$

By Definition 2, \mathcal{X}^{*+} and \mathcal{Y}^{*+} are polyhedral representing sets for \mathcal{X}^* and \mathcal{Y}^* , respectively. The next result shows that projecting the relative interiors of these representing sets gives the strict complementary solutions of programs (1) and (4).

Theorem 8. It follows that $(\mathbf{x}^{*s}, \mathbf{y}^{*s})$ is a strict complementary solution to LFP (1) and its dual (4) if and only if $(\mathbf{x}^{*s}; \mathbf{u}^{*s}) \in \text{ri}(\mathcal{X}^{*+})$ and $(\mathbf{y}^{*s}; \mathbf{v}^{*s}) \in \text{ri}(\mathcal{Y}^{*+})$.

⁷ For more details on the concept of global reference set and its identification, the reader may refer to [17, 18, 19].

Proof. Let $(\mathbf{x}^{*s}; \mathbf{u}^{*s}) \in \text{ri}(\mathcal{X}^{*+})$ and let $(\mathbf{y}^{*s}; \mathbf{v}^{*s}) \in \text{ri}(\mathcal{Y}^{*+})$. By [20, Theorem 4.1], the relative interior of a nonnegative polyhedral set with equality defining constraints consists of its maximal elements. It follows that $(\mathbf{x}^{*s}; \mathbf{u}^{*s}) \in \text{me}(\mathcal{X}^{*+})$ and $(\mathbf{y}^{*s}; \mathbf{v}^{*s}) \in \text{me}(\mathcal{Y}^{*+})$. By the Goldman–Tucker theorem, $(\mathbf{x}^{*s}, \mathbf{y}^{*s})$ is thus a strict complementary solution to LPs (6) and (7). By part (i) of Theorem 6, it follows that $(\mathbf{x}^{*s}, \mathbf{y}^{*s})$ is a strict complementary solution to programs (1) and (4).

Conversely, let $(\mathbf{x}^{*s}, \mathbf{y}^{*s})$ be a strict complementary solution to programs (1) and (4). Then, similar to the proof of Theorem 6, it can be proved that $(\mathbf{x}^{*s}, \mathbf{y}^{*s})$ is a strict complementary solution to LPs (2) and (4). \square

To illustrate the concept of strict complementarity, we return back to Example 1. Adding the nonnegative slack variables u_1 and u_2 to (8b) and (8c) obtains the following representing set for \mathcal{X}^* :

$$\mathcal{X}^{*+} = \{(\mathbf{x}^\lambda; \mathbf{u}^\lambda) \in \mathbb{R}_+^4 : (x_1^\lambda, x_2^\lambda, u_1^\lambda, u_2^\lambda) = (1 - \lambda, 4 - 2\lambda, 4\lambda, 0), \lambda \in [0, 1]\}.$$

Similarly, adding the nonnegative slack variables v_1 and v_2 to (10b) and (10c) results the following representing set for \mathcal{Y}^* :

$$\mathcal{Y}^{*+} = \{(\mathbf{y}; \mathbf{v}) \in \mathbb{R}_+^4 : (y_1, y_2, v_1, v_2) = \left(0, \frac{1}{3}, 0, 0\right)\}.$$

Consider the midpoint $P = (\frac{1}{2}, 3)$ of the line segment AB in Figure 1. This point is associated with the vector $(\mathbf{x}^{\frac{1}{2}}; \mathbf{u}^{\frac{1}{2}}) = \left((\frac{1}{2}, 3)^\top; (2, 0)^\top\right)$, which is a maximal element and, therefore, a relative interior point of the set \mathcal{X}^{*+} . Furthermore, consider the point $(0, \frac{1}{3})$ in Figure 3 that is associated with the single element $(\mathbf{y}; \mathbf{v}) = \left((0, \frac{1}{3})^\top; \mathbf{0}\right)$ of the set \mathcal{Y}^{*+} . Clearly, $(\mathbf{x}^{\frac{1}{2}}, \mathbf{y})$ is a strict complementary solution to programs (8) and (9).

Note that $(\mathbf{x}^\lambda, \mathbf{y})$ is a strict complementary solution for all $\lambda \in (0, 1)$. However, this is not true for $\lambda = 0, 1$. This is because in either of these two cases, pairwise sum of the complementary variables is not positive.

5 Finding a strict complementary solution

5.1 Finding a maximal element of a nonnegative polyhedral set

Consider the following nonempty polyhedral set in \mathbb{R}^d :

$$\mathcal{P} = \{\mathbf{x} \in \mathbb{R}^d : \mathbf{P}\mathbf{x} + \mathbf{Q}\mathbf{y} + \mathbf{R}\mathbf{z} = \mathbf{t}, \mathbf{x}, \mathbf{y} \geq \mathbf{0}, \mathbf{z} \text{ sign free}\},$$

where $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{y} \in \mathbb{R}^e$, and $\mathbf{z} \in \mathbb{R}^f$ are the vectors of variables, \mathbf{P} , \mathbf{Q} , and \mathbf{R} are, respectively, matrices of coefficients of orders $c \times d$, $c \times e$, and $c \times f$, and $\mathbf{t} \in \mathbb{R}^c$ is a constant vector.

Mehdiloozad et al. [20] developed a general convex optimization program for finding a maximal element of a nonnegative convex set. As a consequence of their Theorem 3.2, the following result develops an LP for finding a maximal element of \mathcal{P} .

Theorem 9. Let $(\mathbf{x}^{1*}, \mathbf{x}^{2*}, \mathbf{y}^*, \mathbf{z}^*, w^*)$ be an optimal solution to the following LP:

$$\begin{aligned} & \max \quad \mathbf{1}^\top \mathbf{x}^1 \\ & \text{subject to} \\ & \mathbf{P}(\mathbf{x}^1 + \mathbf{x}^2) + \mathbf{Q}\mathbf{y} + \mathbf{R}\mathbf{z} = \mathbf{t}w, \\ & \mathbf{1} \geq \mathbf{x}^1 \geq \mathbf{0}, \mathbf{x}^2, \mathbf{y} \geq \mathbf{0}, \mathbf{z} \text{ sign free}, w \geq 1. \end{aligned} \tag{15}$$

Then $\frac{1}{w^*}(\mathbf{x}^{1*} + \mathbf{x}^{2*}) \in \text{me}(\mathcal{P})$.

Proof. By [20, Definition 2.5], the characteristic cone of the nonnegative polyhedral set \mathcal{P} is $\mathcal{C}_{\mathcal{P}} = \left\{ x_{d+1} \begin{pmatrix} \mathbf{x} \\ 1 \end{pmatrix} : \mathbf{x} \in \mathcal{P}, x_{d+1} > 0 \right\}$. To find a maximal element of \mathcal{P} , this cone is incorporated into the convex program proposed in [20, Theorem 3.2]. This leads to the following LP:

$$\begin{aligned} & \max \quad \mathbf{1}^\top \mathbf{x}^1 + w^1 \\ & \text{subject to} \\ & \mathbf{P}(\mathbf{x}^1 + \mathbf{x}^2) + \mathbf{Q}\mathbf{y} + \mathbf{R}\mathbf{z} = \mathbf{t}(w^1 + w^2), \\ & \mathbf{1} \geq \mathbf{x}^1 \geq \mathbf{0}, \mathbf{x}^2, \mathbf{y} \geq \mathbf{0}, \mathbf{z} \text{ sign free}, 1 \geq w^1 \geq 0, w^2 \geq 0. \end{aligned} \tag{16}$$

Because the maximization linear program (16) is feasible and its objective function is upper bounded by $d + 1$, it has a finite optimal solution, namely, $(\mathbf{x}^{1*}, \mathbf{x}^{2*}, \mathbf{y}^*, \mathbf{z}^*, w^{1*}, w^{2*})$.

By the assumption, we have $\mathcal{P} \neq \emptyset$. Hence, it follows from [20, Theorem 3.2] that $w^{1*} = 1$. It is clear that program (15) is derived from program (16) by replacing w^1 with its optimal value and using the variable substitution $w = w^1 + 1$. This implies that any optimal solution of program (15) gives an optimal solution to program (16). Namely, if we define $\mathbf{x}^{1'} = \mathbf{x}^{1*}$, $\mathbf{x}^{2'} = \mathbf{x}^{2*}$, $\mathbf{y}' = \mathbf{y}^*$, $\mathbf{z}' = \mathbf{z}^*$, $w^{1'} = 1$, and $w^{2'} = w^* - 1$, then $(\mathbf{x}^{1'}, \mathbf{x}^{2'}, \mathbf{y}', \mathbf{z}', w^{1'}, w^{2'})$ is an optimal solution to program (16). Therefore, the statement of the theorem follows from [20, Theorem 3.2]. \square

5.2 Our proposed approaches

Though Theorem 6 demonstrates the existence of a strict complementary solution for programs (1) and (4), it does not specify how to identify such a solution. To deal with this issue, we develop two approaches in this section by applying Theorem 9. The GAMS code of these approaches is provided in Appendix A.

5.2.1 First approach

From Theorem 8, any pair of relative interior points of \mathcal{X}^{*+} and \mathcal{Y}^{*+} determines a strict complementary solution to programs (1) and (4). By Theorem 9, we develop the following two LPs to find such relative interior points:

$$\begin{aligned} & \max \quad \mathbf{1}^\top \mathbf{x}^1 + \mathbf{1}^\top \mathbf{u}^1 \\ & \text{subject to} \\ & \begin{bmatrix} (\mathbf{c} - \mathbf{d}^*)^\top & \mathbf{0}^\top \\ \mathbf{A} & \mathbf{I} \end{bmatrix} \begin{pmatrix} \mathbf{x}^1 + \mathbf{x}^2 \\ \mathbf{u}^1 + \mathbf{u}^2 \end{pmatrix} = \begin{pmatrix} -\alpha + \beta^* \\ \mathbf{b} \end{pmatrix} w_P, \\ & \mathbf{1} \geq \begin{pmatrix} \mathbf{x}^1 \\ \mathbf{u}^1 \end{pmatrix} \geq \mathbf{0}, \quad \begin{pmatrix} \mathbf{x}^2 \\ \mathbf{u}^2 \end{pmatrix} \geq \mathbf{0}, \quad w_P \geq 1. \end{aligned} \quad (17)$$

$$\begin{aligned} & \max \quad \mathbf{1}^\top \mathbf{y}^1 + \mathbf{1}^\top \mathbf{v}^1 \\ & \text{subject to} \\ & \begin{bmatrix} \mathbf{b}^\top & \mathbf{0}^\top \\ \mathbf{A}^\top & -\mathbf{I} \end{bmatrix} \begin{pmatrix} \mathbf{y}^1 + \mathbf{y}^2 \\ \mathbf{v}^1 + \mathbf{v}^2 \end{pmatrix} = \begin{pmatrix} -\alpha + \beta^* \\ \mathbf{c} - \mathbf{d}^* \end{pmatrix} w_D, \\ & \mathbf{1} \geq \begin{pmatrix} \mathbf{y}^1 \\ \mathbf{v}^1 \end{pmatrix} \geq \mathbf{0}, \quad \begin{pmatrix} \mathbf{y}^2 \\ \mathbf{v}^2 \end{pmatrix} \geq \mathbf{0}, \quad w_D \geq 1. \end{aligned} \quad (18)$$

Let $(\mathbf{x}^{1*}, \mathbf{x}^{2*}, \mathbf{u}^{1*}, \mathbf{u}^{2*}, w_P^*)$ and $(\mathbf{y}^{1*}, \mathbf{y}^{2*}, \mathbf{v}^{1*}, \mathbf{v}^{2*}, w_D^*)$ be optimal solutions to programs (17) and (18), respectively. Then, it follows from Theorem 9 that

$$(\mathbf{x}^{\text{ri}}; \mathbf{u}^{\text{ri}}) = \frac{1}{w_P^*} (\mathbf{x}^{1*} + \mathbf{x}^{2*}; \mathbf{u}^{1*} + \mathbf{u}^{2*}) \in \text{me}(\mathcal{X}^{*+}), \quad (19a)$$

$$(\mathbf{y}^{\text{ri}}; \mathbf{v}^{\text{ri}}) = \frac{1}{w_D^*} (\mathbf{y}^{1*} + \mathbf{y}^{2*}; \mathbf{v}^{1*} + \mathbf{v}^{2*}) \in \text{me}(\mathcal{Y}^{*+}). \quad (19b)$$

Therefore, taking into account [20, Theorem 4.1] and Theorem 8, it follows from (19) that $(\mathbf{x}^{\text{ri}}, \mathbf{y}^{\text{ri}})$ is a strict complementary solution to programs (1) and (4).

5.2.2 Second approach

Our first approach of identifying a strict complementary solution requires the knowledge of the optimal objective value of program (1). In this section, we propose an alternative approach that is exempt from this requirement. We exploit the fact that the optimality of feasible solutions to a pair of primal and dual LPs follows from the equality of their corresponding objective function values. Specifically, we consider the following set:

$$\begin{aligned} \mathcal{W}^* = \{(\bar{\mathbf{x}}; \mathbf{y}) \in \mathbb{R}^{n+m} : & \mathbf{c}^\top \bar{\mathbf{x}} + \alpha t = z, \\ & \mathbf{A}\bar{\mathbf{x}} \leq \mathbf{b}t, \mathbf{d}^\top \bar{\mathbf{x}} + \beta t = 1, \\ & -\mathbf{b}^\top \mathbf{y} + \beta z = \alpha, \mathbf{A}^\top \mathbf{y} + \mathbf{d}z \geq \mathbf{c}, \\ & \bar{\mathbf{x}} \geq \mathbf{0}, t \geq 0, \mathbf{y} \geq \mathbf{0}, z \text{ sign free}\}. \end{aligned} \quad (20)$$

By the projection lemma,⁸ it follows that \mathcal{W}^* is a nonnegative polyhedral set in \mathbb{R}_+^{n+m} . By adding slack vectors $\bar{\mathbf{u}}$ and \mathbf{v} to the inequality constraints of this set, we obtain the following set:

$$\begin{aligned} \mathcal{W}^{*+} = \{(\bar{\mathbf{x}}; \mathbf{y}; \bar{\mathbf{u}}; \mathbf{v}) \in \mathbb{R}^{2(n+m)} : & \mathbf{c}^\top \mathbf{x} + \alpha t - z = 0, \\ & \mathbf{A}\bar{\mathbf{x}} + \bar{\mathbf{u}} - \mathbf{b}t = \mathbf{0}, \mathbf{d}^\top \mathbf{x} + \beta t = 1, \\ & -\mathbf{b}^\top \mathbf{y} + \beta z = \alpha, \mathbf{A}^\top \mathbf{y} + \mathbf{d}z - \mathbf{v} = \mathbf{c}, \\ & \mathbf{x}, \mathbf{v} \geq \mathbf{0}, \mathbf{y}, \bar{\mathbf{u}} \geq \mathbf{0}, t \geq 0, z \text{ sign free}\}. \end{aligned} \quad (21)$$

By Definition 2, \mathcal{W}^{*+} is a polyhedral representing set for \mathcal{W}^* . Let $(\bar{\mathbf{x}}; \mathbf{y}; \bar{\mathbf{u}}; \mathbf{v}) \in \mathcal{W}^{*+}$. Then $(\bar{\mathbf{x}}; \mathbf{y}; \bar{\mathbf{u}}; \mathbf{v})$ satisfies (21) with some scalars t and z . The set \mathcal{W}^* is defined by conditions (2b)–(2d) and (4b)–(4d) and the additional equality requiring that the objective function of LP (2) to be equal to the objective function of its dual (4). By the optimality criterion theorem of linear optimization, it follows that $(\bar{\mathbf{x}}, t)$ and (\mathbf{y}, z) are optimal solutions to LPs (2) and (4), respectively. Additionally, $\bar{\mathbf{u}}$ and \mathbf{v} are their corresponding slack vectors added to inequalities (2b) and (4b), respectively. Consequently, we have $\frac{1}{t}(\bar{\mathbf{x}}; \bar{\mathbf{u}}) \in \mathcal{X}^{*+}$ and $(\mathbf{y}; \mathbf{v}) \in \mathcal{Y}^{*+}$. Based on this, the next result shows that any maximal element of \mathcal{W}^{*+} determines two relative interior points of \mathcal{X}^{*+} and \mathcal{Y}^{*+} , and therefore a strict complementary solution to programs (1) and (4).

Theorem 10. Let $(\bar{\mathbf{x}}^{\text{me}}; \mathbf{y}^{\text{me}}; \bar{\mathbf{u}}^{\text{me}}; \mathbf{v}^{\text{me}}) \in \text{me}(\mathcal{W}^{*+})$ satisfy (21) with some scalars t^{me} and z^{me} . Then $(\frac{1}{t^{\text{me}}} \bar{\mathbf{x}}^{\text{me}}, \mathbf{y}^{\text{me}})$ is a strict complementary solution to programs (1) and (4).

⁸ The projection lemma states that the projection of a polyhedral set onto the space of any subset of its characterizing variables is a polyhedral set; see, for example, [6, Corollary 2.4].

Proof. Let $(\bar{\mathbf{x}}^{\text{me}}; \mathbf{y}^{\text{me}}; \bar{\mathbf{u}}^{\text{me}}; \mathbf{v}^{\text{me}})$ be a maximal element of \mathcal{W}^{*+} that satisfies (21) with some scalars t^{me} and z^{me} . Then $\frac{1}{t^{\text{me}}}(\bar{\mathbf{x}}^{\text{me}}; \bar{\mathbf{u}}^{\text{me}}) \in \text{me}(\mathcal{X}^{*+})$ and $(\mathbf{y}^{\text{me}}; \mathbf{v}^{\text{me}}) \in \text{me}(\mathcal{Y}^{*+})$. Therefore, by [20, Theorem 4.1], the statement follows from Theorem 8. \square

Based on Theorem 9, we develop the following LP to find a maximal element of \mathcal{W}^{*+} :

$$\begin{aligned} & \max \quad \mathbf{1}^\top \bar{\mathbf{x}}^1 + \mathbf{1}^\top \bar{\mathbf{u}}^1 + \mathbf{1}^\top \mathbf{y}^1 + \mathbf{1}^\top \mathbf{v}^1 \\ & \text{subject to} \\ & \begin{bmatrix} \mathbf{c}^\top & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{A} & \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{d}^\top & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & -\mathbf{b}^\top & \mathbf{0} \\ \mathbf{A}^\top & \mathbf{0} & \mathbf{0} & -\mathbf{I} \end{bmatrix} \begin{pmatrix} \bar{\mathbf{x}}^1 + \bar{\mathbf{x}}^2 \\ \bar{\mathbf{u}}^1 + \bar{\mathbf{u}}^2 \\ \mathbf{y}^1 + \mathbf{y}^2 \\ \mathbf{v}^1 + \mathbf{v}^2 \end{pmatrix} + \begin{bmatrix} \alpha \\ -\mathbf{b} \\ \beta \\ 0 \\ 0 \end{bmatrix} t + \begin{bmatrix} -1 \\ \mathbf{0} \\ 0 \\ \beta \\ \mathbf{d} \end{bmatrix} z = \begin{pmatrix} 0 \\ \mathbf{0} \\ 1 \\ \alpha \\ \mathbf{c} \end{pmatrix} w, \quad (22) \\ & \mathbf{1} \geq \begin{pmatrix} \bar{\mathbf{x}}^1 \\ \bar{\mathbf{u}}^1 \\ \mathbf{y}^1 \\ \mathbf{v}^1 \end{pmatrix} \geq \mathbf{0}, \quad \begin{pmatrix} \bar{\mathbf{x}}^2 \\ \bar{\mathbf{u}}^2 \\ \mathbf{y}^2 \\ \mathbf{v}^2 \end{pmatrix} \geq \mathbf{0}, \quad t \geq 0, \quad z \text{ sign free}, \quad w \geq 1. \end{aligned}$$

Let $(\bar{\mathbf{x}}^{1*}, \bar{\mathbf{x}}^{2*}, \bar{\mathbf{u}}^{1*}, \bar{\mathbf{u}}^{2*}, \mathbf{y}^{1*}, \mathbf{y}^{2*}, \mathbf{v}^{1*}, \mathbf{v}^{2*}, t^*, z^*, w^*)$ be an optimal solution to program (22), and define

$$(\bar{\mathbf{x}}^{\text{me}}; \mathbf{y}^{\text{me}}; \bar{\mathbf{u}}^{\text{me}}; \mathbf{v}^{\text{me}}) = \frac{1}{w^*} (\bar{\mathbf{x}}^{1*} + \bar{\mathbf{x}}^{2*}; \mathbf{y}^{1*} + \mathbf{y}^{2*}; \bar{\mathbf{u}}^{1*} + \bar{\mathbf{u}}^{2*}; \mathbf{v}^{1*} + \mathbf{v}^{2*}).$$

By Theorem 9, we have $(\bar{\mathbf{x}}^{\text{me}}; \mathbf{y}^{\text{me}}; \bar{\mathbf{u}}^{\text{me}}; \mathbf{v}^{\text{me}}) \in \text{me}(\mathcal{W}^{*+})$. If $t^{\text{me}} = \frac{t^*}{w^*}$, then it follows by Theorem 10 that $(\frac{1}{t^{\text{me}}} \bar{\mathbf{x}}^{\text{me}}, \mathbf{y}^{\text{me}}) = (\frac{1}{t^*} (\bar{\mathbf{x}}^{1*} + \bar{\mathbf{x}}^{2*}), \frac{1}{w^*} (\mathbf{y}^{1*} + \mathbf{y}^{2*}))$ is a strict complementary solution to programs (1) and (4).

6 Numerical example

In this section, we illustrate our proposed approaches of finding strict complementary solutions with a numerical example, taken from [1, 14].

Example 2. Consider the following LFP:

$$\begin{aligned}
& \max \quad \frac{x_1 + 2x_2 + 3.5x_3 + x_4 + 1}{2x_1 + 2x_2 + 3.5x_3 + 3x_4 + 4} \\
& \text{subject to} \\
& 2x_1 + x_2 + 3x_3 + 3x_4 \leq 10, \\
& x_1 + 2x_2 + x_3 + x_4 \leq 14, \\
& x_1, x_2, x_3, x_4 \geq 0.
\end{aligned} \tag{23}$$

The dual of program (23) is the following LP:

$$\begin{aligned}
& \min \quad z \\
& \text{subject to} \\
& 2y_1 + y_2 + 2z \geq 1, \\
& y_1 + 2y_2 + 2z \geq 2, \\
& 3y_1 + y_2 + 3.5z \geq 3.5, \\
& 3y_1 + y_2 + 3z \geq 1, \\
& -10y_1 - 14y_2 + 4z = 1, \\
& y_1, y_2 \geq 0, z \text{ sign free.}
\end{aligned} \tag{24}$$

Let us add primal slack variables u_1 and u_2 to program (23), and dual slack variables v_1, v_2, v_3 , and v_4 to the inequality constraints of program (24), to turn their inequality constraints to equalities. We use our proposed approaches to find a strict complementary solution to the above pair of programs. Running the modified version of the GAMS code provided in Appendix A results that the joint optimal objective value of programs (23) and (24) is equal to 0.857. Furthermore, the strict complementary solution obtained from both of our proposed approaches is

$$\begin{aligned}
x_1^* &= 0, & x_2^* &= 1.071, & x_3^* &= 1.2, & x_4^* &= 0, & u_1^* &= 0, & u_2^* &= 0; \\
v_1^* &= 1.071 & v_2^* &= 0 & v_3^* &= 0, & v_4^* &= 2.071, & y_1^* &= 0.143 & y_2^* &= 0.071.
\end{aligned}$$

7 Concluding remarks

An indirect approach for establishing duality results in linear fractional optimization is based on applying the well-known transformation of Charnes and Cooper [9]. This approach converts a primal LFP into an equivalent LP and then defines the dual of the obtained LP as the dual of the primal LFP. An advantage of using this approach is that it allows exploiting the duality results of linear optimization for establishing duality statements in linear fractional optimization.

In this paper, we show that the dual program derived from the above approach is the same dual program suggested in [8]. Based on this version of

duality, we provide new criteria for primal and dual optimality as our first contribution. We equivalently represent the primal and dual optimal sets as the optimal sets of a pair of primal and dual LPs. By this representation, it follows that a primal (resp., dual) feasible solution is optimal if and only if its binding polyhedral cone contains the objective vector of the corresponding primal (resp., dual) LP. This condition not only is (theoretically) necessary and sufficient for the optimality of any general LFP, but also is a new geometrical tool for solving two- and three-dimensional LFPs.

As our second contribution, we introduce the concept of strict complementarity into the framework of linear fractional optimization. We prove the existence of a strict complementary solution and show that all such solutions induce unique optimal partitions for the sets of indices of nonnegative variables. To geometrically interpret the strict complementarity, we equivalently represent primal and dual optimal sets by two nonnegative polyhedral sets that are described only by equality constraints. Then we prove that each pair of relative interior points of these representing sets is a strict complementary solution, and vice versa. By this result, we deal with the problem of identifying a strict complementary solution. Specifically, we turn this problem to the equivalent problem of identifying a maximal element of a nonnegative polyhedral set. Then, by applying the technique of finding a maximal element of a nonnegative polyhedral set, we develop two linear optimization approaches with different strategies for finding a strict complementary solution in linear fractional optimization.

Our first approach identifies a strict complementary solution by solving two LPs and requires knowing the optimal objective value of the given LFP. In contrast, our second approach involves solving a single (but larger) LP and does not need the per-knowledge of the optimal objective value. As our proposed approaches are linear optimization based, they allow for applying the ordinary simplex algorithm of linear optimization to identify a strict complementary solution in linear fractional optimization. Nonetheless, regarding the preference on using the proposed approaches, note that each of the LPs developed in our first approach has less number of constraints than the LP of our second approach. Therefore, the use of our first approach is particularly recommended in situations, where only primal or dual part of a strict complementary solution needs to be found. For example, while the SBM model of Tone [26] is used for the measurement of efficiency in the field of data envelopment analysis, the global reference set of an inefficient decision making unit can be identified by either the primal or dual part of a strict complementary solution. However, solving the primal SBM model is recommended because the number of its constraints are mostly less than that of its dual.

The approaches developed in our paper open up a number of further research avenues. First, it should be interesting to extend the results proposed in linear optimization literature on the use of strict complementarity for post-optimality analysis [5, 13] to linear fractional optimization. Second,

an interesting method for finding a strict complementary solution in linear optimization is to apply the so-called Balinski–Tucker tableau [2]. From our contribution, it is found that this method can be used to generate (indirectly) a strict complementary solution in linear fractional optimization. Therefore, it is worth exploring modification of the Balinski–Tucker method so that a strict complementary solution is directly obtained.

Appendix A

The following computer program written in GAMS identifies a strict complementary solution for the primal LFP (8) and its dual (9). Making this program applicable for any LFP in the general form of (1) just requires modifying “Sets”, “Table $A(i,j)$ ”, “Parameters”, “Alpha,” and “Beta” in Lines 1–23.

```

1  Sets
2      i          row number of matrix A  /i1*i2/
3      j          column number of matrix A /j1*j2/;
4
5  Table A(i,j)
6      j1      j2
7  i1      2      1
8  i2     -2      1;
9
10 Parameters
11     b/i1      6
12     i2        2/
13     c/j1      6
14     j2        3/
15     d/j1      5
16     j2        2/;
17
18 Scalars
19     Alpha
20     Beta;
21
22 Alpha=6;
23 Beta =5;
24
25 File ProgSC / Results.txt /;
26 Put ProgSC;
27
28 *****
29 *Stage 1: Solving program (3)
30
31 Free Variables
32     Theta;
33
34 Positive Variables
35     xbar(j)
36     t ;
37
38 Scalar
39     ThetaStar;
40
41 Equations
42     Obj

```

```

43      Con1
44      Con2;
45
46      Obj..      Theta =E= Sum(j, c(j)*xbar(j)) + Alpha*t;
47      Con1(i)..      Sum(j, a(i,j)*xbar(j)) =L= b(i)*t;
48      Con2..      Sum(j, d(j)*xbar(j)) + Beta*t =E= 1;
49
50 Model MainLFP      / Obj, Con1, Con2 /;
51
52 Put /'Finding the Optimal Obj. Value (ThetaStar)';
53 Put /'-----'/;
54 Option LP=CONOPT;
55 Solve MainLFP using LP Maximizing Theta;
56      Put 'Obj = ';>6; Put Theta.L:<10:3;
57      ThetaStar=Theta.L;
58 Put /'-----'/;
59
60 *End of Stage 1
61 *****
62
63 *****
64 *First approach: Solving program (18)
65
66 Positive Variables
67      xbar1(j)
68      xbar2(j)
69      ubar1(i)
70      ubar2(i)
71      y1(i)
72      y2(i)
73      v1(j)
74      v2(j)
75      w1
76      w2
77      p;
78
79 Free variable
80      q;
81
82 xbar2.up(j) = 1;
83 ubar2.up(i) = 1;
84 y2.up(i) = 1;
85 v2.up(j) = 1;
86 w2.up = 1;
87
88 Parameters
89      XbarStar
90      tStar
91      UbarStar
92      YStar(i)
93      zStar
94      VStar(j);
95
96 Equations
97      ObjP
98      ConP1
99      ConP2
100     ConP3
101     ObjD
102     ConD1
103     ConD2
104     ConD3;
105
106 ObjP..      Theta =E= Sum(j, xbar2(j)) + Sum(i, ubar2(i)) + w2;
107 ConP1(i)..      Sum(j, a(i,j)*(xbar1(j)+xbar2(j))) - b(i)*p
108               + ubar1(i)+ubar2(i) =E= 0;
109 ConP2..      Sum(j, d(j) *(xbar1(j)+xbar2(j))) + Beta*p

```

```

- w1+w2          =E= 0;
109  ConP3..      Sum(j, c(j) *(xbar1(j)+xbar2(j))) + Alpha*p
- (w1+w2)*ThetaStar =E= 0;
110
111  ObjD..      Theta =E= Sum(i, y2(i)) + Sum(j, v2(j)) + w2;
112  ConD1(j)..  Sum(i, a(i,j)*(y1(i)+y2(i))) + d(j)*q - v1(j)
- v2(j) - c(j)*(w1+w2) =E= 0;
113  ConD2..      -Sum(i, b(i) *(y1(i)+y2(i))) + Beta*q
- Alpha*(w1+w2) =E= 0;
114  ConD3..      q - ThetaStar*(w1+w2) =E= 0;
115
116  Models Primal_SCSC / ObjP , ConP1, ConP2, ConP3 /
117         Dual_SCSC / ObjD , ConD1, ConD2, ConD3 / ;
118
119  Solve Primal_SCSC using LP Maximizing Theta;
120  XbarStar(j) = (xbar1.L(j)+xbar2.L(j))/(w1.L+w2.L);
121  tStar      = p.L/(w1.L+w2.L);
122  UbarStar(i) = (ubar1.L(i)+ubar2.L(i))/(w1.L+w2.L);
123
124  Solve Dual_SCSC using LP Maximizing Theta;
125  YStar(i) = (y1.L(i)+y2.L(i))/(w1.L+w2.L);
126  zStar    = q.L/(w1.L+w2.L);
127  VStar(j) = (v1.L(j)+v2.L(j))/(w1.L+w2.L);
128
129  Put / 'Finding a SC Solution via Approach I';
130  Put /-----'/;
131  Put '          Primal          Dual          '/;
132  Put '-----'/;
133  Loop(j,
134      Put 'x':>5; Put ord(j):<>3:0; Put '= ':3; Put (XbarStar(j)/tStar):<10:3;
135      Put 'v':>5; Put ord(j):<>3:0; Put '= ':3; Put Vstar(j):<10:3;
136      Put /;
137  );
138  Put /;
139  Loop(i,
140      Put 'u':>5; Put ord(i):<>3:0; Put '= ':3; Put (UbarStar(i)/tStar):<10:3;
141      Put 'y':>5; Put ord(i):<>3:0; Put '= ':3; Put Ystar(i):<10:3;
142      Put /;
143  );
144  Put '-----' / / ;
145
146  *End of First approach
147  *****
148
149  *****
150  *Second approach: Solving program (22)
151
152  Equations
153      ObjPD
154      ConPD ;
155
156  ObjPD..      Theta =E= Sum(j, xbar2(j)) + Sum(i, ubar2(i)) + Sum(i, y2(i))
+ Sum(j, v2(j)) + w2;
157  ConPD..      Sum(j, c(j)*(xbar1(j)+xbar2(j))) + Alpha*p - q =E= 0;
158
159  Model PD_SCSC / ObjPD, ConP1, ConP2, ConD1, ConD2, ConPD/ ;
160
161  Solve PD_SCSC using LP Maximizing Theta;
162  XbarStar(j) = (xbar1.L(j)+xbar2.L(j))/(w1.L+w2.L);
163  UbarStar(i) = (ubar1.L(i)+ubar2.L(i))/(w1.L+w2.L);
164  Ystar(i) = (y1.L(i)+y2.L(i))/(w1.L+w2.L);
165  Vstar(j) = (v1.L(j)+v2.L(j))/(w1.L+w2.L);
166
167  Put / 'Finding a SC Solution via Approach II';
168  Put /-----'/;
169  Put '          Primal          Dual          '/;
170  Put '-----'/;

```

```

171 Loop(j,
172     Put 'x':>5; Put ord(j):<>3:0; Put '= ':3; Put (XbarStar(j)/tStar):<10:3;
173     Put 'v':>5; Put ord(j):<>3:0; Put '= ':3; Put Vstar(j):<10:3;
174     Put /;
175 );
176 Put /;
177 Loop(i,
178     Put 'u':>5; Put ord(i):<>3:0; Put '= ':3; Put (UbarStar(i)/tStar):<10:3;
179     Put 'y':>5; Put ord(i):<>3:0; Put '= ':3; Put Ystar(i):<10:3;
180     Put /;
181 );
182 Put '-----'/ / /;
183
184 *End of Second approach
185 *****

```

References

1. Bajalinov, E.B. *Linear-fractional programming theory, methods, applications and software*, Kluwer Academic publishers, Boston, 2003.
2. Balinski, M.L. and Tucker, A.W. *Duality theory of linear programs: A constructive approach with applications*, SIAM Rev. 11(3) (1969), 347–377.
3. Bazaraa, M.S., Sherali, H.D. and Shetty, C.M. *Nonlinear programming: Theory and algorithms*. John Wiley & Sons, Hoboken, NJ, 2006.
4. Ben-Tal, A., El Ghaoui, L. and Nemirovski A. *Robust optimization*, Princeton University Press, Princeton, NJ, 2009.
5. Berkelaar, A.B., Roos, K. and Terlaky, T. The optimal set and the optimal partition approach. *Advances in Sensitivity Analysis and Parametric Programming*, edited by T., Gal, and H.J. Greenberg, International Series in Operations Research & Management Science, Vol. 6. Springer, Boston, MA (1997) 6.1–6.45.
6. Bertsimas, D. and Tsitsiklis, J.N. *Introduction to linear optimization*, Athena Scientific, Massachusetts, 1997.
7. Chadha, S.S. *A dual fractional program*, Z. Angew. Math. Mech. 51 (1971), 560–561.
8. Chadha, S.S. and Chadha, V. *Linear fractional programming and duality*, Ann. Oper. Res. 15 (2007), 119–125.
9. Charnes, A. and Cooper, W.W. *Programming with linear fractional functional*, Naval Res. Logist. Q. 15(3-4) (1962), 181–186.

10. Charnes, A., Cooper, W.W. and Rhodes, E. *Measuring the efficiency of decision making units*, Eur. J. Oper. Res. 2(6) (1978), 429–444.
11. Craven, B.D. *Fractional programming*, Sigma Series in Applied Mathematics, Heldermann Verlag, Berlin, 1988.
12. Goldman, A.J. and Tucker, A.W. Theory of linear programming, *Linear inequalities and related systems*, edited by Kuhn, H.W. and Tucker, A.W., Princeton University Press, NJ (1956) 53–97.
13. Greenberg, H. J. *The use of the optimal partition in a linear programming solution for postoptimal analysis*, Oper. Res. Lett. 15(4) (1994), 179–186.
14. Hasan, M.B. and Acharjee, S. *Solving LFP by converting it into a single LP*, Int. J. Oper. Res. 8(3) (2011), 1–14.
15. Jahanshahloo, G.R., Hosseinzadeh Lotfi, F., Mehdiloozad, M. and Roshdi, I. *Connected directional slack-based measure of efficiency in DEA*, Appl. Math. Sci. 6(5) (2012), 237–246.
16. Lai, K.K., Mishra, S.K., Panda, G., Chakraborty, S.K., Samei, M.E. and Ram, B. *A limited memory q -BFGS algorithm for unconstrained optimization problems*, J. Appl. Math. Comput. 66 (2021), 183–202.
17. Mehdiloozad, M. *Identifying the global reference set in DEA: A mixed 0-1 LP formulation with an equivalent LP relaxation*, Oper. Res. 17 (2017), 205–211.
18. Mehdiloozad, M., Mirdehghan, S.M., Sahoo, B.K. and Roshdi, I. *On the identification of the global reference set in data envelopment analysis*, Eur. J. Oper. Res. 245(3) (2015), 779–788.
19. Mehdiloozad, M. and Sahoo, B.K. Identifying the global reference set in DEA: An application to the determination of returns to scale, *Handbook of Operations Analytics Using Data Envelopment Analysis*, edited by S.N., Hwang, H.S., Lee and J., Zhu, International Series in Operations Research & Management Science, Vol. 239, Springer, Boston, MA (2016) 299–330.
20. Mehdiloozad, M., Tone, K., Askarpour, R. and Ahmadi, M.B. *Finding a maximal element of a nonnegative convex set through its characteristic cone: An application to finding a strictly complementary solution*, Comput. Appl. Math. 37 (2018), 53–80.
21. Mishra, S.K., Samei, M.E. and Chakraborty, S.K. *On q -variant of Dai–Yuan conjugate gradient algorithm for unconstrained optimization problems*, Nonlinear Dyn. (2021). <https://doi.org/10.1007/s11071-021-06378-3>.
22. Pastor, J.T., Ruiz, J.L. and Sirvent, I. *An enhanced DEA Russell graph efficiency measure*, Eur. J. Oper. Res. 115(3) (1999), 596–607.

23. Rockafellar, R.T. *Convex analysis*, Princeton University Press, Princeton, NJ, 1970.
24. Schaible, S. *Fractional programming: Applications and algorithms*, Eur. J. Oper. Res. 7(2) (1981), 111–120.
25. Stancu-Minasian, I.M. *Fractional programming: Theory, methods and applications*, Kluwer Academic publishers, Boston, 1997.
26. Tone, K. *A slacks-based measure of efficiency in data envelopment analysis*, Eur. J. Oper. Res. 130(3) (2001), 498–509.
27. Tone, K. *A slacks-based measure of super-efficiency in data envelopment analysis*, Eur. J. Oper. Res. 143(1) (2002), 32–41.
28. Tone, K. *Variations on the theme of slacks-based measure of efficiency in DEA*, Eur. J. Oper. Res. 200(3) (2009), 901–907.
29. Tone, K. and Tsutsui, M. *Network DEA: A slacks-based measure approach*, Eur. J. Oper. Res. 197(1) (2009), 243–252.
30. Tone, K., Toloo, M. and Izadikhah, M. *A modified slacks-based measure of efficiency in data envelopment analysis*, Eur. J. Oper. Res. 287(2) (2020), 560–571.
31. Tone, K. and Tsutsui, M. *Dynamic DEA: A slacks-based measure approach*, Omega 38(3-4) (2010), 145–156.



Solving quantum optimal control problems by wavelets method

M. Rahimi, S.M. Karbassi* and M.R. Hooshmandasl

Abstract

We present the quantum equation and synthesize an optimal control procedure for this equation. We develop a theoretical method for the analysis of quantum optimal control system given by the time depending Schrödinger equation. The Legendre wavelet method is proposed for solving this problem. This can be used as an efficient and accurate computational method for obtaining numerical solutions of different quantum optimal control problems. The distinguishing feature of this paper is that it makes the method, previously used to solve non-quantum control equations based on Legendre wavelets, usable by using a change of variables for quantum control equations.

AMS subject classifications (2020): 49M25; 35J10.

Keywords: Quantum Equations; Optimal Control Problems; Legendre Wavelets Methods.

1 Introduction

With the advent of the twentieth century, the inability of classical physics in the fields of relativity and microscopy led to the outburst of quantum physics. After the development of quantum physics, the issue of quantum control was inspired by experimental advances and issues raised in sciences such as

*Corresponding author

Received 19 October 2020; revised 5 May 2021; accepted 30 May 2021

Mojtaba Rahimi

Department of Mathematical Sciences, Yazd University, Yazd, Iran. e-mail: rahimimojtaba1358@gmail.com

Seyed mehdi Karbassi

Department of Mathematical Sciences, Yazd University, Yazd, Iran. e-mail: smkarbassi@yazd.ac.ir

Mohammad Reza Hooshmandasl

Department of Computer Science, University of Mohaghegh Ardabili, Ardabil, Iran. hooshmandasl@uma.ac.ir

quantum chemistry, quantum optics, quantum information, and atomic and molecular physics [7, 10, 24, 27, 34]. Recently, classical control methods such as optimal control, robust control, Lyapunov control, and feedback control for quantum systems have been studied and expanded [5, 16, 19, 20, 22, 25]. In quantum control, the main goal is to effectively control the system from an initial state to a desired final state using external control fields. However, control in quantum systems is at the beginning of the road, and further research is needed. In particular, optimal control in quantum systems is of particular importance as one of the most widely used issues [3, 9, 12, 15, 17, 23, 26, 31].

Along with the development of analytical control methods in quantum systems, numerical methods have also been developed. Numerical methods that have been considered in solving classical equations and problems are also studied [3, 11, 18, 28]. One of the most useful numerical methods for solving differential problems is wavelets-based numerical method [2, 14, 21, 29, 33]. This article tries to provide a useful way to solve optimal quantum control based on wavelets method.

Let \mathcal{H} be a finite- or infinite-dimensional Hilbert space of a quantum system and let Ψ denote the state of this system. Then the Schrödinger equation can be found as follows [13, 32]:

$$i\hbar \frac{\partial \Psi}{\partial t} = H\Psi, \quad (1)$$

where $\Psi \in H$ is the state variable, \hbar is the Planck constant, and H is a self-adjoint Hamiltonian operator in \mathcal{H} .

In every physical system, energy is an important quantity. In quantum systems, Hamiltonian H is corresponding to energy, then we can write $He_i = E_ie_i$, $i = 1, \dots, N$, where E_i are eigenvalues, e_i are eigenvectors of the physical system concerned, and N is the dimension of \mathcal{H} . In this paper, we suppose that e_i is an orthogonal basis; then we expand a state vector Ψ as follows:

$$\Psi = \sum_{i=1}^N \psi_i e_i.$$

The population of energy states of level i are the quantities $|\psi_i|^2$. When a quantum system is operated by an external field, the Schrödinger equation in (1) is modified as

$$i\hbar \frac{\partial \Psi}{\partial t} = (H_0 + \sum_k H_k u_k(t))\Psi,$$

where $H_0 : \mathcal{H} \mapsto \mathcal{H}$, which is the internal Hamiltonian, and the Hamiltonian linear operator $H_k : \mathcal{H} \mapsto \mathcal{H}$ describes the coupling of the system to external fields $u_k(t)$. In this paper, for simplicity, we consider a quantum system with one control, and set $k = 1$. Then we can write

$$i\hbar \frac{\partial \Psi}{\partial t} = (H_0 + H_1 u(t))\Psi.$$

The design of controls in quantum systems is considered for energy-efficient population transfer. These controls perform the desired transmission and in addition optimize a specific performance index. The specific energy performance index considered in this section is shown below:

$$\min J[u] = \int_0^T u^2(t) dt.$$

This cost function for control has been widely used in the literature on optimal control of quantum systems as a part of various objective functions. It is a measure of the energy expended to create a control field.

The main object of this paper is to present an efficient numerical algorithm based on the Legendre wavelets methods to solve the following optimal control problems of the form:

$$\min J[u] = \int_0^T u^2(t) dt,$$

subject to the dynamical quantum systems

$$i\dot{\Psi} = (H_0 + V u(t))\Psi.$$

The initial condition for the above equation is

$$\Psi(0) = \psi_0$$

It is worthy to note that $V = H_1$.

2 The hat function

In this section, we introduce a family of basic functions, namely, the hat functions. An n -set of the hat functions is defined on the interval $[0, T]$ as follows:

$$h_0(t) = \begin{cases} \frac{k-t}{k}, & 0 \leq t \leq k, \\ 0, & o.w. \end{cases}$$

$$h_j(t) = \begin{cases} \frac{t-(j-1)k}{k}, & (j-1)k \leq t \leq jk, \\ \frac{(j+1)k-t}{k}, & jk \leq t \leq (j+1)k, \\ 0, & o.w. \end{cases}$$

$$h_{n-1}(t) = \begin{cases} \frac{t-(1-k)}{k}, & T-k \leq t \leq T, \\ 0, & o.w. \end{cases}$$

where $k = \frac{T}{n-1}$. For hat functions, we can write $h_i(jk) = \delta_{ij}$, where δ is the Kronecker delta. By the definition of the hat functions, we can expand any function like $g(t) \in L^2[0, T]$ as follows:

$$g(t) \simeq \sum_{j=0}^{n-1} g_j h_j(t) = G^T H(t) = H^T(t) G, \quad (2)$$

where

$$G \triangleq [g_0, g_1, \dots, g_{n-1}]$$

and

$$H(t) \triangleq [h_0(t), h_1(t), \dots, h_{n-1}(t)]^T.$$

When we use the hat functions for the $g(t)$, it can be observed that

$$g_j = g(jk), \quad j = 0, 1, \dots, n-1. \quad (3)$$

Now, we introduce another family of basic functions, namely, Legendre wavelet functions. The set of these functions is an orthogonal set on the interval $[0, 1]$ with the weight function $w(t) = 1$. If $P_k(t)$ is the Legendre polynomials of degree k that are orthogonal on the interval $[-1, 1]$ with respect to the weight function $w(t) = 1$, then we can write Legendre wavelets as follows [2, 14]:

$$w_{lk}(t) = \begin{cases} \sqrt{2k+1} 2^{\frac{n}{2}} P_k(2^{n+1}t - 2l + 1), & t \in [\frac{l-1}{2^n}, \frac{l}{2^n}], \\ 0, & o.w. \end{cases} \quad (4)$$

In fact, $w_{lk}(t) = w(l, k, n, t)$, where $l = 1, 2, \dots, 2^n$ and n is an arbitrary positive integer.

For any arbitrary function, like $g(t)$ defined over $[0, 1]$ and square-integrable over $[0, 1]$, we can expand $g(t)$ by the Legendre wavelets as follows:

$$g(t) = \sum_{l=1}^{\infty} \sum_{k=0}^{\infty} a_{lk} w_{lk}(t).$$

By approximating the above infinite series, we can write

$$g(t) \simeq \sum_{l=1}^{2^n} \sum_{k=0}^{K-1} a_{lk} w_{lk}(t) = A^T W(t), \quad (5)$$

where A and $W(t)$ are $\hat{k} = 2^n K$ column vectors.

For the index lk , we can write $j = k(l-1) + k + 1$. Then $a_{lk} = a_j$ and

$w_{lk} = w_j$. Thus (5) can be written as

$$g(t) \simeq \sum_{l=1}^{\hat{k}} a_l w_l(t) = A^T W(t),$$

where

$$A \triangleq [a_1, a_2, \dots, a_{\hat{k}}]^T$$

and

$$W(t) \triangleq [w_1(t), w_2(t), \dots, w_{\hat{k}}(t)]. \quad (6)$$

Now by taking $t_j = \frac{j}{\hat{k}-1}$ as the collocation points into (6), we can write

$$P_{\hat{k} \times \hat{k}} \triangleq [W(0), W(\frac{1}{\hat{k}-1}), \dots, W(1)].$$

It can be simplify verified that the Legendre wavelets can be expanded in \hat{k} in terms of the hat function by using (2) and (3) as follows:

$$W(t) \simeq P_{\hat{k} \times \hat{k}} H(t). \quad (7)$$

Theorem 1. Suppose that $f(t) \simeq F^T H(t)$ and that $g(t) \simeq G^T H(t)$. Then

$$f(t)g(t) \simeq S^T H(t), \quad (8)$$

where $S_{ij} = (F.G)_{ij} = F_{ij}G_{ij}$ denotes pointwise product of F and G .

Proof. By applying (2) and (3) for $f(t)$ and $g(t)$, we can write

$$g(t) \simeq G^T H(t) = \sum_{j=0}^{n-1} g_j h_j(t) = \sum_{j=0}^{\hat{k}-1} g(jk) h_j(t),$$

$$f(t) \simeq F^T H(t) = \sum_{j=0}^{n-1} f_j h_j(t) = \sum_{j=0}^{\hat{k}-1} f(jk) h_j(t).$$

Then by using the point wise product, we have

$$f(t)g(t) \simeq \sum_{j=0}^{\hat{k}-1} f(jk)g(jk)h_j(t) = D^T H(t),$$

which completes the proof. \square

Corollary 1. Suppose that $g(t) \simeq G^T H(t)$ by hat functions. Then, for any integer number $n \geq 2$, we have

$$(g(t))^n \simeq [g_0^n, g_1^n, \dots, g_{\hat{k}-1}^n] H(t). \quad (9)$$

Proof. For $n = 2$ by Theorem 1, we have

$$(g(t))^2 \simeq [g_0^2, g_1^2, \dots, g_{k-1}^2]H(t);$$

then by induction for $n > 2$, we have

$$(g(t))^n \simeq [g_0^n, g_1^n, \dots, g_{k-1}^n]H(t),$$

which completes the proof. \square

Theorem 2. Suppose that $g(t) \simeq A^T W(t)$ and $f(t) \simeq B^T W(t)$ by Legendre wavelets. Then we can write

$$f(t)g(t) \simeq Q^T P_{\hat{k} \times \hat{k}}^{-1} W(t), \quad (10)$$

where $A_1^T = A^T P_{\hat{k} \times \hat{k}}$, $B_1^T = B^T P_{\hat{k} \times \hat{k}}$, and $Q = A_1.B_1$.

Proof. By Theorem 1 and equation (7), we have

$$g(t) \simeq A^T W(t) \simeq A^T P_{\hat{k} \times \hat{k}} H(t) = A_1^T H(t),$$

$$f(t) \simeq B^T W(t) \simeq B^T P_{\hat{k} \times \hat{k}} H(t) = B_1^T H(t),$$

and then

$$f(t)g(t) \simeq (A_1.B_1)^T H(t) = Q^T H(t) \simeq Q^T P_{\hat{k} \times \hat{k}}^{-1} W(t). \quad \square$$

In Theorem 1, the multiplication of two functions is obtained according to the hat functions, and in Theorem 2 by using Theorem 1, the multiplication of two functions is obtained according to the Legendre wavelets. If $g(t) \simeq A^T W(t)$ by the Legendre wavelets, then by Theorem 2 and Corollary 1, we can write

$$(g(t))^n \simeq [\tilde{a}_1^n, \tilde{a}_2^n, \dots, \tilde{a}_k^n] P_{\hat{k} \times \hat{k}}^{-1} W(t). \quad (11)$$

3 Analysis of the proposed method

In previous works, using Legendre wavelet, numerical solutions for non-quantum control equations have been obtained. In this section, we try to implement one of these methods, which is based on the Legendre wavelet, for quantum control equations. To overcome this problem, first change of variables to make the equation usable is performed.

In this section, we consider the quantum control systems of the form

$$\min J[u] = \int_0^T u^2(t) dt,$$

subject to the dynamical system

$$i\dot{\Psi} = (H_0 + Vu(t))\Psi, \quad (12)$$

with the initial condition $\Psi(0) = \Psi_0$.

First, we introduce a change of variables that generalizes

$$\Psi = e^{-iH_0 t} x,$$

such that for (12), we obtain

$$\begin{aligned} i(-iH_0 e^{-iH_0 t} x + e^{-iH_0 t} \dot{x}) &= (H_0 + Vu(t))e^{-iH_0 t} x \rightarrow H_0 e^{-iH_0 t} x + e^{-iH_0 t} \dot{x} \\ &= H_0 e^{-iH_0 t} x + Vu(t)e^{-iH_0 t} x. \end{aligned}$$

Thus we can write

$$\dot{x} = e^{iH_0 t} Vu(t)e^{-iH_0 t} x.$$

By considering $E(t) = e^{iH_0 t} V e^{-iH_0 t}$, equation (12) can be written as

$$\dot{x} = E(t)u(t)x(t), \quad (13)$$

and the new initial condition is $x(0) = \Psi_0$.

By the Legendre wavelets for the derivative of the state variable \dot{x} and the control variable $u(t)$, we can write

$$\dot{x} \simeq X^T W(t) \quad (14)$$

and

$$u(t) \simeq U^T W(t), \quad (15)$$

where

$$U^T = [u_1, u_2, \dots, u_{\hat{k}}]$$

and

$$X^T = [x_1, x_2, \dots, x_{\hat{k}}].$$

We can apply integration on both sides of (14) and considering the initial condition

$$x(t) \simeq X^T Q W(t) + \Psi_0.$$

Let η be the coefficients vector of the unit function. Then

$$x(t) \simeq (X^T Q + \Psi_0 \eta^T) W(t). \quad (16)$$

Now (16) can be written as

$$x(t) \simeq A^T W(t) = [a_1, a_2, \dots, a_{\hat{k}}] W(t). \quad (17)$$

In this part, by applying (11) and (9) for the above approximation, we have

$$(x(t))^n \simeq [\tilde{a}_1^n, \tilde{a}_2^n, \dots, \tilde{a}_{\hat{k}}^n] P_{\hat{k} \times \hat{k}}^{-1} W(t) \simeq [\tilde{a}_1^n, \tilde{a}_2^n, \dots, \tilde{a}_{\hat{k}}^n] H(t),$$

and also for (15), we have

$$(u(t))^n \simeq [\tilde{u}_1^n, \tilde{u}_2^n, \dots, \tilde{u}_{\hat{k}}^n] P_{\hat{k} \times \hat{k}}^{-1} W(t) \simeq [\tilde{u}_1^n, \tilde{u}_2^n, \dots, \tilde{u}_{\hat{k}}^n] H(t). \quad (18)$$

We can also approximate $E(t)$ by Legendre wavelets as

$$E(t) \simeq E^T W(t), \quad (19)$$

where

$$E^T = [e_1, e_2, \dots, e_{\hat{k}}].$$

In this part, by using (8) and (10) for $E(t)u(t)x(t)$, we have

$$\begin{aligned} E(t)u(t)x(t) &\simeq [\tilde{e}_1 \tilde{u}_1 \tilde{x}_1, \tilde{e}_2 \tilde{u}_2 \tilde{x}_2, \dots, \tilde{e}_{\hat{k}} \tilde{u}_{\hat{k}} \tilde{x}_{\hat{k}}] P_{\hat{k} \times \hat{k}}^{-1} W(t) \\ &\simeq [\tilde{e}_1 \tilde{u}_1 \tilde{x}_1, \tilde{e}_2 \tilde{u}_2 \tilde{x}_2, \dots, \tilde{e}_{\hat{k}} \tilde{u}_{\hat{k}} \tilde{x}_{\hat{k}}] H(t) \\ &= \Delta_1^T H(t), \end{aligned} \quad (20)$$

also by using (18) for $n = 2$, we have

$$(u(t))^2 \simeq [\tilde{u}_1^2, \tilde{u}_2^2, \dots, \tilde{u}_{\hat{k}}^2] P_{\hat{k} \times \hat{k}}^{-1} W(t) \simeq [\tilde{u}_1^2, \tilde{u}_2^2, \dots, \tilde{u}_{\hat{k}}^2] H(t) = \Delta_2^T H(t). \quad (21)$$

Now by using (21), the index J can be written as

$$J \simeq J[U] = \Delta_2^T \Omega, \quad (22)$$

where

$$\Omega = [\int_0^T h_o(t) dt, \int_0^T h_1(t) dt, \dots, \int_0^T h_{\hat{k}-1}(t) dt]. \quad (23)$$

By applying (20) for (13) and (12), we can write

$$X^T - \Delta_1 P_{\hat{k} \times \hat{k}}^{-1} \simeq 0. \quad (24)$$

In this part, by Lagrange multiplier method for minimization index J in (22) subject to systems of algebraic equation (24), we can write

$$\tilde{J}[X, U, L] = J[U] + X^T - \Delta_1 P_{\hat{k} \times \hat{k}}^{-1} L = \Delta_2^T \Omega + X^T - \Delta_1 P_{\hat{k} \times \hat{k}}^{-1} L, \quad (25)$$

where

$$L = [L_1, L_2, \dots, L_{\hat{k}}]^T.$$

Hence L is the vector of Lagrange multiplier.

For minimizing by the Lagrange method, the necessary conditions are

$$\begin{cases} \frac{\partial \tilde{J}}{\partial X} = 0, \\ \frac{\partial \tilde{J}}{\partial U} = 0, \\ \frac{\partial \tilde{J}}{\partial L} = 0. \end{cases} \quad (26)$$

By the Newton iteration method, we can solve equations of (26) for X , U , and L . Then the approximation of $x(t)$ and $u(t)$ can be determined by (17) and (15).

4 Proposed algorithm

The object of this algorithm is designed to solve the Schrödinger equation:

Input: T (final time), N (dimension of \mathcal{H}), H_0 , and H_1 .

Step1: Make a change of variable $\Psi = e^{-iH_0 t}x$ in dynamical system (12) to obtain (13).

Step2: Define $x(t)$, $u(t)$, and $E(t)$ by (17), (15), and (19), respectively.

Step3: Write the index J as (22) and (23).

Step4: Compute dynamical system (13) by applying (20) to the form (24).

Step5: Compose new index $\tilde{J}[X, U, L]$ as (25) by (24) and (23).

Step6: Solve equation systems in (26) and obtain X and U .

Step7: Compute $x(t)$ and $u(t)$ by (17) and (15).

Step8: Compute $\Psi = e^{-iH_0 t}x$.

Output: The approximate solution $u(t)$ and Ψ .

5 Numerical experiments

Example 1. In this example, we consider the two-level system $i\dot{\Psi} = (H_0 + Vu(t))\Psi$, where $\Psi \in \mathbb{C}^2$ as follows:

$$\mathbf{H}_0 = \begin{pmatrix} E_1 & 0 \\ 0 & E_2 \end{pmatrix}, \quad \mathbf{V} = \begin{pmatrix} 0 & v_{12} \\ v_{12}^* & 0 \end{pmatrix}.$$

The concept of optimal control of two-level quantum systems was presented in [1, 3, 4, 6, 8, 11, 13, 30]. In the most of these researches, optimal control is constructed on the basis of geometric arguments. If we suppose $E_1 = 2$,

$E_2 = -2$, and $v_{12} = 2+3i$, then we can obtain the numerical results as follows:

Table 1: Approximate and exact value of $u(t)$ in Example 1 for different values of t .

t	approximate of $u(t)$	exact of $u(t)$	error
1	-0.0359	-0.0459	0.0120
5	-0.0496	-0.0578	0.0102
10	0.1010	0.0999	0.0051
15	-0.0922	-0.0878	0.0016
20	0.0663	0.0594	0.0119
25	0.0286	0.0066	0.0220
30	-0.0630	-0.0635	0.0020

Table 2: Approximate and exact value of $x(t)$ in Example 1 for different values of t .

t	approximate of $\psi_1(t)$	exact of $\psi_1(t)$	approximate of $\psi_2(t)$	exact of $\psi_2(t)$	error of $\psi_1(t)$	error of $\psi_2(t)$
1	1.0097	1.0078	0.0690	0.0590	0.0109	0.0190
5	0.9869	0.9769	0.2664	0.2564	0.0108	0.0190
10	0.8666	0.8766	0.4681	0.4764	0.0110	0.0113
15	0.7197	0.7297	0.6656	0.6756	0.0120	0.0160
20	0.5513	0.5413	0.8645	0.8445	0.0110	0.0230
25	0.3033	0.3133	0.9270	0.9470	0.0120	0.0220
30	0.0830	0.0737	1.0105	1.0005	0.0123	0.0130

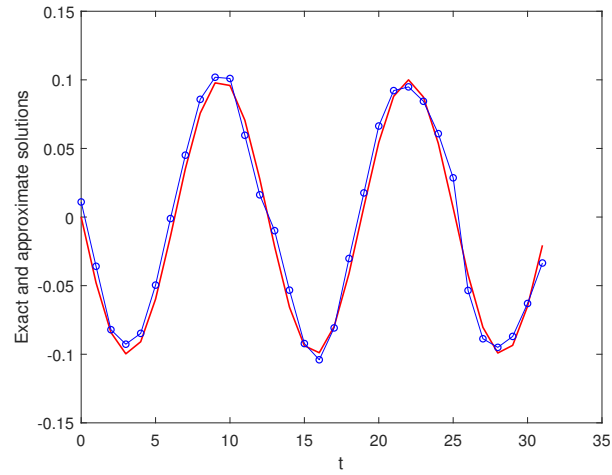


Figure 1: Plots of approximate and exact results of control variable for Example 1

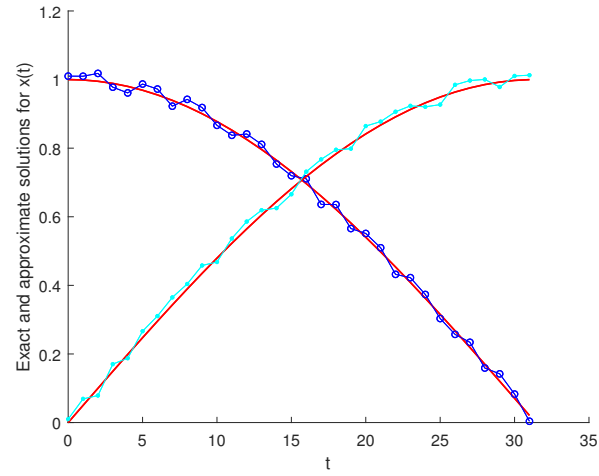


Figure 2: Plots of the approximate and exact results of state variable for Example 1

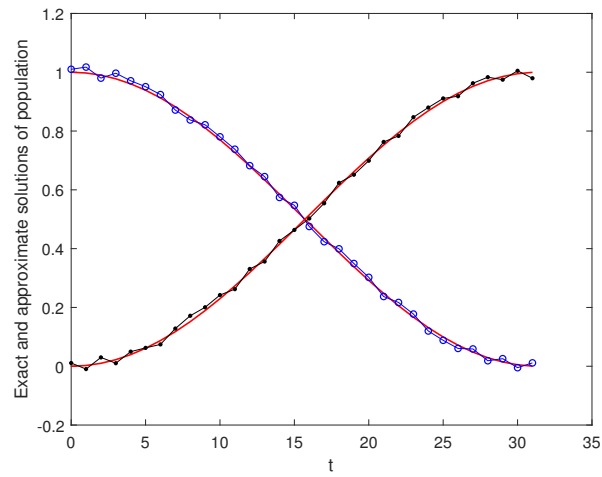


Figure 3: Plots of the approximate and exact results of population for Example 1

We solved the above problem by our proposed algorithm with $n = 2$ and $K = 10$ or $\hat{k} = 40$. The stopping condition is $|u^*(t) - u(t)| < 5 \times 10^{-2}$, where $u^*(t)$ and $u(t)$ are approximate and exact results of control variable, respectively, and by the stopping condition, we have 73 iterations. The plots of the approximate and exact values of $u(t)$ is shown in Figure 1. The plots

of the approximate and exact values of $\psi_1(t)$ and $\psi_2(t)$ are shown in Figure 2. The plots of the approximate and exact values of $|\psi_1(t)|^2$ and $|\psi_2(t)|^2$ are shown in Figure 3.

Approximation and exact results of control variable $u(t)$ and their absolute errors for different t are presented in Table 1. In fact, because the index $J(t)$ in each step depends on the control variable, so by approximating the control variable in each step, the index $J(t)$ is also approximated. Approximation and exact results of state variables $\psi_1(t)$ and $\psi_2(t)$ and their absolute errors for different t are presented in Table 2. The above results show that using the algorithm and method mentioned numerically is very useful and efficient. The main advantage of this method is to provide a simple solution based on classical numerical methods in the field of optimal quantum control. In [32], an index similar to the index used in this article has been used and numerical and graphical results have been obtained. Carefully in these results, it is observed that the solutions obtained in this article have been obtained with less repetition and more accuracy.

Example 2. In this example, we consider the three-level system $i\dot{\Psi} = (H_0 + Vu(t))\Psi$, where $\Psi \in \mathbb{C}^3$ as follows [32]:

$$\mathbf{H}_0 = \begin{pmatrix} E_1 & 0 & 0 \\ 0 & E_2 & 0 \\ 0 & 0 & E_3 \end{pmatrix}, \quad \mathbf{V} = \begin{pmatrix} 0 & v_{12} & v_{13} \\ v_{12}^* & 0 & v_{23} \\ v_{13}^* & v_{23}^* & 0 \end{pmatrix}.$$

The concept of optimal control of three-level quantum systems were presented in [3, 4, 13, 32]. In the most of these works, the optimal control is constructed on the basis of geometric arguments. If we suppose $E_1 = 2$, $E_2 = 0$, $E_3 = 6$, $v_{12} = i + 1$, $v_{13} = 4$, and $v_{23} = 2 + 3i$, then we can obtain the numerical results as follows:

Table 3: Approximate and exact values of $u(t)$ in Example 2 for different values of t .

t	approximate of $u(t)$	exact values of $u(t)$	error
2	0.0074	0.0074	3.3×10^{-4}
10	0.0029	0.0029	5.2×10^{-4}
20	-0.0107	-0.0105	2.0×10^{-4}
30	-0.0002	-0.0001	1.1×10^{-4}
40	0.0062	0.0062	3.3×10^{-4}
50	-0.0032	-0.0032	2.3×10^{-4}
60	0.0060	0.0063	4.1×10^{-4}

Table 4: Approximate and exact values of Ψ in Example 2 for different values of t .

t	approximate of $\psi_1(t)$	approximate of $\psi_2(t)$	approximate of $\psi_3(t)$	error of $\psi_1(t)$	error of $\psi_2(t)$	error of $\psi_3(t)$
2	1.0188	-0.1011	0.0731	0.0200	0.0220	0.0220
10	0.9837	-0.3944	0.2748	0.0140	0.0154	0.0220
20	0.9016	-0.6478	0.5012	0.0210	0.0141	0.0123
30	0.7168	-0.7977	0.6782	0.0220	0.0215	0.0144
40	0.5713	-0.6806	0.8288	0.0180	0.0125	0.0214
50	0.3356	-0.4449	0.9638	0.0000	0.0112	0.0124
60	0.0799	-0.0525	0.9767	0.0225	0.0114	0.0126

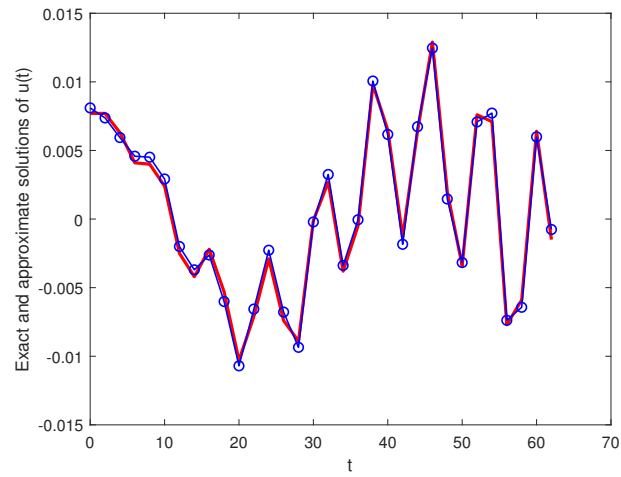


Figure 4: Plots of approximate and exact results of control variable for Example 2

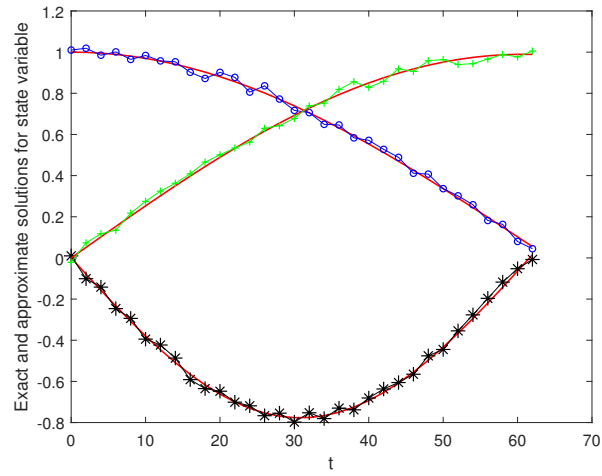


Figure 5: Plots of the approximate and exact results of state variable for Example 2

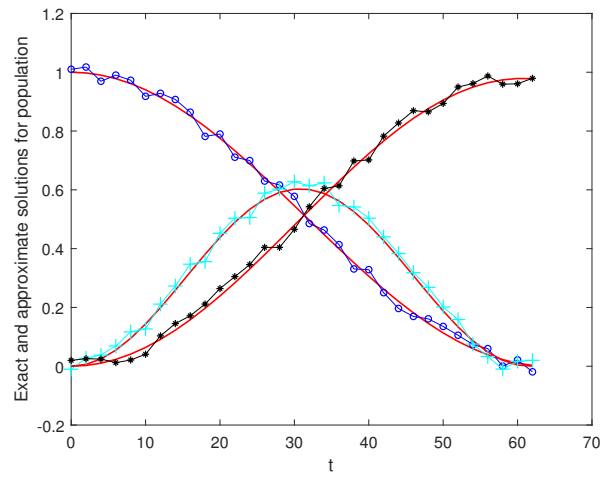


Figure 6: Plots of the approximate and exact results of population for Example 2

We solved the above problem by our proposed algorithm with $n = 2$ and $K = 10$ or $\hat{k} = 40$. The stopping condition is $|u^*(t) - u(t)| < 5 \times 10^{-4}$, where $u^*(t)$ and $u(t)$ are approximate and exact results of control variable, respectively, and by the stopping condition, we have 97 iterations. The plot of the approximate and exact values of $u(t)$ is shown in Figure 4. The plots

of the approximate and exact values of $\psi_1(t)$, $\psi_2(t)$, and $\psi_3(t)$ are shown in Figure 5. The plots of the approximate and exact values of $|\psi_1(t)|^2$, $|\psi_2(t)|^2$ and $|\psi_3(t)|^2$ are shown in Figure 6.

Approximation and exact results of control variable $u(t)$ and their absolute errors for different t are presented in Table 3. Approximation and exact results of state variables $\psi_1(t)$, $\psi_2(t)$, and $\psi_3(t)$ and their absolute errors for different t are presented in Table 4. Figures and tables obtained above show that the numerical method used in this paper is simpler and more useful than other methods. In [32], an index similar to the index used in this article has been used, and numerical and graphical results have been obtained. Investigating these results carefully, it can be observed that the solutions obtained in this article endure less repetition while having more accuracy.

6 Conclusion

Today, the issue of quantum optimal control is one of the most widely used issues in many basic sciences and engineering. At the same time, numerical methods are one of the most useful solutions to these problems. In this paper, a numerical method based on wavelets was proposed to solve the problem of optimal quantum control. This method was benefited by using topics related to applied mathematics in the field of classical equations and presented usable in the field of quantum systems. The above results showed that using the algorithm and method mentioned numerically is very useful and efficient. The merit of this method is to provide a simple solution based on classical numerical methods in the field of optimal quantum control.

References

1. Albertini, F. and D'Alessandro, D. *Time optimal simultaneous control of two level quantum systems*, Automatica, 74 (2016), 55–62.
2. Boggess, A. and Narcowich, F.J. *A first course in wavelets with Fourier analysis*, John Wiley & Sons, 2015.
3. Borzi, A., Salomon, J., and Volkwein, S. *Formulation and numerical solution of finite-level quantum optimal control problems*, J. Comput. Appl. Math. 216(1) (2008), 170–197.
4. Boscain, U., Charlot, G., Gauthier, J., Guérin, S., and Jauslin, H. *Optimal control in laser-induced population transfer for two-and three-level quantum systems*, J. Math. Phys. 43(5) (2002), 2107–2132.

5. Chiu, T.Y. and Lin, K.T. *Optimal control of two-qubit quantum gates in a non-Markovian open system*, 2016 12th IEEE International Conference on Control and Automation (ICCA), (2016), 791–796.
6. Clark, W., Bloch, A., Colombo, L., and Rooney, P. *Optimal control of quantum purity for $n = 2$ systems*, 2017 IEEE 56th Annual Conference on Decision and Control (CDC), (2017), 1317–1322.
7. Cong, Sh. *Control of quantum systems: theory and methods*, John Wiley & Sons, 2014.
8. Cong, Sh., Wen, J., and Zou, X. *Comparison of time optimal control for two level quantum systems*, J. Sys. Eng. Electron. 25(1) (2014), 95–103.
9. Dadashi, M.R., Haghighi, A.R., Soltanian, F., and Yari, A. *On the numerical solution of optimal control problems via Bell polynomials basis*, Iran. J. Numer. Anal. Optim. 10(2) (2020), 197–221.
10. D'alessandro, D. *Introduction to quantum control and dynamics*. Chapman & Hall/CRC Applied Mathematics and Nonlinear Science Series. Chapman & Hall/CRC, Boca Raton, FL, 2008.
11. D'alessandro, D. and Dahleh, M. *Optimal control of two-level quantum systems*, IEEE Transactions on Automatic Control, 46(6) (2001), 866–876.
12. Edrisi-Tabri, Y., Lakestani, M., and Heydari, A. *Two numerical methods for nonlinear constrained quadratic optimal control problems using linear B-spline functions*, Iran. J. Numer. Anal. Optim. 6(2) (2016), 17–38.
13. Grivopoulos, S. and Bamieh, B. *Optimal population transfers for a quantum system in the limit of large transfer time*, Proceedings of the 2004 American Control Conference, volume 3, (2004), 2481–2486.
14. Heydari, M.H., Hooshmandasl, M.R., Maalek Ghaini, F.M., and Cattani, C. *Wavelets method for solving fractional optimal control problems*, Appl. Math. Comput. 286 (2016), 139–154.
15. Itami, T. *Nonlinear optimal control as quantum mechanical eigenvalue problems*, Automatica, 41(9) (2005), 1617–1622.
16. Jacobs, K. and Shabani, A. *Quantum feedback control: how to use verification theorems and viscosity solutions to find optimal protocols*, Contemp. Phys. 49(6) (2008), 435–448.
17. Keller, D. *Optimal control of a linear stochastic Schrödinger equation*, Discrete Contin. Dyn. Syst. 2013, Dynamical systems, differential equations and applications. 9th AIMS Conference. Suppl. 437–446.

18. Koçak, Y., Çelik, E., and Aksoy, N.Y. *A note on optimal control problem governed by Schrödinger equation*, Open Phys. 13, (2015), 407–413.
19. Kuang, S., Dong, D., and Petersen, I.R. *Rapid Lyapunov control of finite-dimensional quantum systems*, Automatica, 81 (2017), 164–175.
20. Li, J., Yang, X., Peng, X., and Sun, Ch.P. *Hybrid quantum-classical approach to quantum optimal control*, Phys. Rev. Lett. 118(15) (2017), 150503.
21. Lotfi, A., Dehghan, M., and Yousefi, A.A. *A numerical technique for solving fractional optimal control problems*, Compute. Math. Appl. 62(3) (2011), 1055–1067.
22. Mirrahimi, M., Rouchon, P., and Turinici, G. *Lyapunov control of bilinear Schrödinger equations*, Automatica, 41(11) (2005), 1987–1994.
23. Mohammadzadeh, F., Tehrani, H.A. and Noori Skandari, M.H. *Chebyshev pseudo-spectral method for optimal control problem of Burgers' equation*, Iran. J. Numer. Anal. Optim. 9(2) (2019), 77–102.
24. Oğuztöreli, M.N., Bellman, R.E., and Oğuztöreli, M.N. *Time-lag control systems*, Mathematics in Science and Engineering, 24 Academic Press, New York-London 1966.
25. Riviello, G., Brif, C., Long, R., Wu, R.B., Tibbetts, K.M., Ho, T.S., and Rabitz, H. *Searching for quantum optimal control fields in the presence of singular critical points*, Phys. Rev. A, 90(1) (2014), 013404.
26. Shi, B., Xu, Ch., and Wu, R. *Time scaling transformation in quantum optimal control computation*, 2018 37th Chinese Control Conference (CCC), 2018, pp. 8138–8143.
27. Thirring, W. *Classical mathematical physics: dynamical systems and field theories*, Springer Science & Business Media, 2013.
28. Toyoglu, F. and Yagubov, G. *Numerical solution of an optimal control problem governed by two dimensional Schrödinger equation*, Appl. Math. Comput. 4(2) (2015), 30–38.
29. Rehman, M. and Khan, R.A. *The Legendre wavelet method for solving fractional differential equations*, Commun. Nonlinear Sci. Numer. Simul. 16(11) (2011), 4163–4173.
30. Damme, L.V., Ansel, Q., Glaser, S.J., and Sugny, D. *Robust optimal control of two-level quantum systems*, Phys. Rev. A, 95(6) (2017), 13 pp.
31. Wang, Q.F. *Quantum optimal control of nonlinear dynamics systems described by Klein-Gordon-Schrödinger equations*, Proceeding of American Control Conference, 2006, 1032–1037.

- 32. Werschnik, J. and Gross, E. *Quantum optimal control theory*, J. Phys. B 40(18) (2007), R175–R211.
- 33. Xu, X. and Xu, D. *Legendre wavelets method for approximate solution of fractional-order differential equations under multi-point boundary conditions*, Int. J. Comput. Math. 95(5) (2018), 998–1014.
- 34. Zhu, W. and Rabitz, H. *Attaining optimal controls for manipulating quantum systems*, Int. J. Quantum Chem. 93(2) (2003), 50–58.



Singularly perturbed robin type boundary value problems with discontinuous source term in geophysical fluid dynamics

B.M. Abagero, G.F. Duressa and H.G. Debela*

Abstract

Singularly perturbed robin type boundary value problems with discontinuous source terms applicable in geophysical fluid are considered. Due to the discontinuity, interior layers appear in the solution. To fit the interior and boundary layers, a fitted nonstandard numerical method is constructed. To treat the robin boundary condition, we use a finite difference formula. The stability and parameter uniform convergence of the proposed method is proved. To validate the applicability of the scheme, two model problems are considered for numerical experimentation and solved for different values of the perturbation parameter, ε , and mesh size, h . The numerical result is tabulated, and it is observed that the present method is more accurate and uniformly convergent with order of convergence of $O(h)$.

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

Keywords: Singularly perturbed problem; Robin type boundary value problems; Discontinuous source term; Nonstandard fitted method.

1 Introduction

Singular perturbation problems model convection-diffusion processes in applied mathematics that arise in diverse areas, including linearized Navier-Stokes equation at high Reynolds number and the drift-diffusion equation

Received 11 May 2021; revised 28 May 2021; accepted 4 June 2021

Bediru Musa Abagero

Department of Mathematics, College of Natural Sciences, Jimma University, Jimma, Ethiopia. e-mail: Bedirumusa6@gmail.com.

Gemechis File Duressa

Department of Mathematics, College of Natural Sciences, Jimma University, Jimma, Ethiopia. e-mail: gammeef@gmail.com.

Habtamu Garoma Debela

Department of Mathematics, College of Natural Sciences, Jimma University, Jimma, Ethiopia e-mail: habte200@gmail.com.

of semiconductor device modeling, heat and mass transfer at high Pe'clet number, and so on; see [6, 13, 18, 19]. The novel aspect of the problem under consideration is that we take a source term in the differential equation that has a jump discontinuity at one or more points in the interior of the domain. This gives rise to an interior layer in the exact solution of the problem, in addition to the boundary layer at the outflow boundary point. Problems with discontinuous data were treated theoretically, in the case of the solution of the convection–diffusion with Dirichlet case problem; see [9, 10]. Authors of [2, 8, 14] discussed a self-adjoint Dirichlet type problem with a discontinuous source term. Authors [14, 15, 17] have examined two parameter singularly perturbed boundary value problems for second-order ordinary differential equations with discontinuous source term. Authors of [5, 7] discussed fitted nonstandard finite difference methods for singularly perturbed second-order ordinary differential equations. Singularly perturbed delay differential equation was examined by Mohapatra and Natesan [12] on an adaptively generated grid. Recently, Shandru and shanthi [3] presented a fitted mesh method to solve singularly perturbed robin type boundary value problems with discontinuous source terms. Indeed, still, there is a room to increase the accuracy and show the parameter uniform convergence because the treatment of singular perturbation problem is not trivial distributions and the solution is pended on perturbation parameter, ε and mesh size, h ; see [6]. Due to this, the numerical treatment of singularly perturbed boundary value problems is need improvement. Therefore, it is important to develop a more accurate and convergent numerical method for solving singularly perturbed boundary value problems under consideration.

2 Definition of the problem

Consider the following singularly perturbed problem with Robin boundary condition of the form

$$Ly(x) \equiv \varepsilon y''(x) + a(x)y'(x) - b(x)y(x) = f(x), \quad x \in \Omega^- \cup \Omega^+. \quad (1)$$

Subject to boundary conditions

$$\begin{cases} L_1 y(0) = \alpha_1 y(0) - \beta_1 \varepsilon y'(0) = p, \\ L_2 y(1) = \alpha_2 y(1) + \beta_2 \varepsilon y'(1) = q, \end{cases} \quad (2)$$

where $\alpha_1, \beta_1 \geq 0$, $\alpha_1 + \beta_1 > 0$, $\alpha_2 > 0, \beta_2 \geq 0$, and $\varepsilon > 0$ is a small parameter. The functions $a(x)$ and $b(x)$ are smooth on $\bar{\Omega}$, such that $a(x) \geq a > 0$ and $b(x) \geq b \geq 0$. Furthermore, the notations for the domain are $\Omega = (0, 1)$, $\Omega^- = (0, d)$, and $\Omega^+ = (d, 1)$, where $d \in \Omega$ stands for the jump in the source function. Boundary value problem of the governing problem

under consideration is a model confinement of a plasma column by reaction pressure and geophysical fluid dynamics; see [4].

The solution $y(x)$ of (1)–(2) has a boundary layer near $x = 0$ due to the perturbation parameter, ε and interior layer due to the discontinuous source term.

3 Properties of continuous solution

The differential operator for (1) is given by

$$L_\varepsilon \equiv \varepsilon \frac{d^2}{dx^2} + a \frac{d}{dx} - b,$$

and it satisfies the following minimum principle for boundary value problems. The following lemmas [6] are necessary for the existence and uniqueness of the solution and for the problem to be well-posed.

Lemma 1 (Continuous minimum principle). Suppose that the function $y \in C^1(\bar{\Omega}) \cap C^2(\Omega^- \cup \Omega^+)$ satisfies $L_1 y(0) \geq 0$, $L_2 y(1) \geq 0$, and $Ly(x) \leq 0$, for all $x \in \Omega^- \cup \Omega^+$ and $[y'](d) \leq 0$. Then, $y(x) \geq 0$ for all $x \in \bar{\Omega}$.

Proof. For the proof, we refer to [3]. □

Lemma 2 (Stability result). Consider the boundary value problem (1)–(2) subject to the conditions $a(x) \geq a > 0$ and $b(x) \geq b \geq 0$. If $y \in C^1(\bar{\Omega}) \cap C^2(\Omega^- \cup \Omega^+)$, then

$$\|y\|_{\bar{\Omega}} \leq C \max\{|L_1 y(0)|, |L_2 y(1)|, |Ly|_{\Omega^- \cup \Omega^+}\}.$$

Proof. For the proof, see [3]. □

Lemma 3. For each integer k satisfying $0 \leq k \leq 4$, the solution of (1)–(2) satisfies the bounds $\|y^{(k)}\|_{\bar{\Omega} \setminus \{d\}} \leq C\varepsilon^{-k}$.

Proof. For the proof, see [3]. □

Lemma 4. Let y_ε be the solution of (P_ε) . Then, for $k = 0, 1, 2, 3$,

$$|y_\varepsilon^{(k)}(x)| \leq C(1 + \varepsilon^{-k} \exp(\frac{-a}{\varepsilon}x)) \quad \text{for all } x \in [0, l].$$

Proof. For the proof, see [1]. □

4 Formulation of the method

The theoretical basis of the nonstandard discrete numerical method is based on the development of the exact finite difference method. The author of [11] presented techniques and rules for developing nonstandard finite difference methods for different problem types. In Mickens's rules, to develop a discrete scheme, the denominator function for the discrete derivatives must be expressed in terms of more complicated functions of step sizes than those used in the standard procedures. These complicated functions constitute a general property of the schemes, which is useful while designing reliable schemes for such problems.

For the problem of the form in (1)–(2), in order to construct the exact finite difference scheme, we follow the procedures used in [1]. Let us consider the following singularly perturbed differential equation of the form

$$\varepsilon y''(x) + a(x)y'(x) - b(x)y(x) = f(x). \quad (3)$$

The constant coefficient homogeneous problems corresponding to (3) are

$$\varepsilon y''(x) + ay'(x) - by(x) = 0, \quad (4)$$

$$\varepsilon y''(x) + ay'(x) = 0, \quad (5)$$

where $a(x) \geq a$ and $b(x) \geq b$. Two linear independent solutions of (4) are $\exp(\lambda_1 x)$ and $\exp(\lambda_2 x)$, where

$$\lambda_{1,2} = \frac{-a \pm \sqrt{a^2 + 4\varepsilon b}}{2\varepsilon}. \quad (6)$$

We discretize the domain $[0, 1]$ using the uniform mesh length $\Delta x = h$ such that $\Omega^N = \{x_i = x_0 + ih, 1, 2, \dots, N, x_0 = 0, x_N = 1, h = \frac{1}{N}\}$, where N denotes the number of mesh points. We denote the approximate solution to $y(x)$ at the grid point x_i by Y_i . Now our main objective is to calculate the difference equation, which has the same general solution as the differential equation (4) has at the grid point x_i given by $Y_i = A_1 \exp(\lambda_1 x_i) + A_2 \exp(\lambda_2 x_i)$. Using the theory of difference equations and the procedures used in [1], we have

$$\det \begin{bmatrix} Y_{i-1} & \exp(\lambda_1 x_{i-1}) & \exp(\lambda_2 x_{i-1}) \\ Y_i & \exp(\lambda_1 x_i) & \exp(\lambda_2 x_i) \\ Y_{i+1} & \exp(\lambda_1 x_{i+1}) & \exp(\lambda_2 x_{i+1}) \end{bmatrix} = 0. \quad (7)$$

Simplifying (7), we obtain

$$-\exp\left(\frac{ah}{2\varepsilon}\right)Y_{i-1} + 2 \cosh\left(\frac{h\sqrt{a^2 + 4\varepsilon b}}{2\varepsilon}\right)Y_i - \exp\left(-\frac{ah}{2\varepsilon}\right)Y_{i+1} = 0, \quad (8)$$

which is an exact difference scheme for (4).

After doing the arithmetic manipulation and rearrangement on (8), for the constant coefficient problem (5), we get

$$\varepsilon \frac{Y_{i-1} - 2Y_i + Y_{i+1}}{\frac{h\varepsilon}{a}(\exp(\frac{ah}{\varepsilon}) - 1)} + a \frac{Y_{i+1} - Y_i}{h} = 0. \quad (9)$$

The denominator function becomes $\Psi^2 = \frac{h\varepsilon}{a} \left(\exp\left(\frac{ha}{\varepsilon}\right) - 1 \right)$. Adopting this denominator function for the variable coefficient problem, we write it as

$$\Psi_i^2 = \frac{h\varepsilon}{a_i} \left(\exp\left(\frac{ha_i}{\varepsilon}\right) - 1 \right), \quad (10)$$

where Ψ_i^2 is the function of ε , a_i , and h .

By using the denominator function Ψ_i^2 in to the main scheme, we obtain the difference scheme as

$$L_\varepsilon^N Y_i \equiv \varepsilon \frac{Y_{i+1} - 2Y_i + Y_{i-1}}{\Psi_i^2} + a_i \frac{Y_{i+1} - Y_i}{h} - b_i Y_i = f_i. \quad (11)$$

This can be written as three term recurrence relation as

$$E_i Y_{i-1} + F_i Y_i + G_i Y_{i+1} = H_i, \quad i = 1, 2, \dots, N-1, \quad (12)$$

where $E_i = \frac{\varepsilon}{\Psi_i^2}$, $F_i = \frac{-2\varepsilon}{\Psi_i^2} - \frac{a_i}{h} - b_i$, $G_i = \frac{\varepsilon}{\Psi_i^2} + \frac{a_i}{h}$, and $H_i = f_i$.

To treat the boundary condition, we use the forward finite difference formula for $i = 0$ and the backward difference formula for $i = N$, respectively, for the first derivative term.

That is, for $i = 0$, from (2), we have $\alpha_1 y(0) - \beta_1 \varepsilon y'_0 = p$ implies $\alpha_1 y_0 - \beta_1 \varepsilon y'_0 = p$, which yields

$$(\alpha_1 + \frac{\beta_1 \varepsilon}{h}) y_0 - \frac{\beta_1 \varepsilon}{h} y_1 = p, \quad (13)$$

Similarly, for $i = N$, from (2), we have $\alpha_2 y(N) + \beta_2 y'_N = q$ implies $\alpha_2 y_N + \beta_2 y'_N = q$, which yields

$$(\alpha_2 + \frac{\beta_2}{h}) y_N - \frac{\beta_2}{h} y_{N-1} = q. \quad (14)$$

Therefore, Equation (1) with the given boundary conditions (2), can be solved using the schemes in (12), (13), and (14) which gives the $N \times N$ system of algebraic equations.

5 Uniform convergence analysis

In this section, we need to show that the discrete scheme in (12) satisfies the discrete minimum principle and uniform convergence. Let us define the forward, backward, and second-order central finite difference operators as

$$D^+Y_j = \frac{Y_{j+1} - Y_j}{h}, \quad D^-Y_j = \frac{Y_j - Y_{j-1}}{h}, \quad \delta^2Y_j = D^+D^-Y_j = \frac{D^+Y_j - D^-Y_j}{h}.$$

Lemma 5 (Discrete Minimum principle). Let V_i be any mush function that satisfies $v_0 \geq 0$, $v_N \geq 0$, and $L^h v_i \leq 0$, $i = 1, 2, \dots, N-1$. Then $v_i \geq 0$, $i = 1, 2, \dots, N$.

Proof. The proof is obtained by contradiction. Let j be such that $V_j = \min V_i$, and suppose that $V_j < 0$. Clearly, $j \notin \{0, N\}$, $V_{j+1} - V_j \geq 0$, and $V_j - V_{j-1} \leq 0$. Therefore,

$$\begin{aligned} L^h V_j &= \frac{\varepsilon}{\Psi_i^2} (V_{j+1} - 2V_j + V_{j-1}) + \frac{a_j}{h} (V_{j+1} - V_j) - bV_j \\ &= \frac{\varepsilon}{\Psi_j^2} [(V_{j+1} - V_j) - (V_j - V_{j-1})] + \frac{a_i}{h} (V_{j+1} - V_j) - bV_j \\ &\geq 0, \end{aligned}$$

where the strict inequality holds if $V_{j+1} - V_j > 0$. This is a contradiction and therefore $V_j \geq 0$. Since j is arbitrary, we have $V_i \geq 0$, $i = 1, 2, \dots, N$. \square

We proved above that the discrete operator L^h satisfies the minimum principle. Next, we analyze the uniform convergence analysis.

Using the Taylor series expansion, the bound for $y(x_{i-1})$ and $y(x_{i+1})$ at x_i are as

$$\begin{cases} y(x_{i-1}) = y(x_i) - hy'(x_i) + \frac{h^2}{2!}y''(x_i) - \frac{h^3}{3!}y^{(3)}(x_i) + \frac{h^4}{4!}y^{(4)}(x_i) + O(h^5), \\ y(x_{i+1}) = y(x_i) + hy'(x_i) + \frac{h^2}{2!}y''(x_i) + \frac{h^3}{3!}y^{(3)}(x_i) + \frac{h^4}{4!}y^{(4)}(x_i) + O(h^5). \end{cases}$$

We obtain the bound for

$$\begin{cases} |D^+D^-y(x_i)| \leq C|y''(x_i)|, \\ |y''(x_i) - D^+D^-y(x_i)| \leq Ch^2|y^{(4)}(x_i)|. \end{cases} \quad (15)$$

Similarly, for the first derivative term, we have

$$|y'(x_i) - D^+y(x_i)| \leq Ch|y^{(2)}(x_i)|, \quad (16)$$

where $|y^{(k)}(x_i)| = \sup_{x_i \in (x_0, x_N)} |y^{(k)}(x_i)|$, $k = 2, 3, 4$.

Theorem 1. Let the coefficients functions $a(x)$ and the source function $f(x)$ in (1)–(2) of the domain Ω be sufficiently smooth, so that $y(x) \in C^4[0, 1]$. Then, the discrete solution Y_i satisfies

$$|L^N(y_i - Y_i)| \leq Ch \left(1 + \sup_{x \in (0,1)} \left(\frac{\exp(\frac{-ax_i}{\varepsilon})}{\varepsilon^3} \right) \right).$$

Proof. We consider the truncation error discretization as

$$\begin{aligned} |L^N(y_i - Y_i)| &= |L^N y_i - L^N Y_i|, \\ &\leq C|\varepsilon y_i'' + a_i y_i' - \{\varepsilon \frac{D^+ D^- h^2}{\Psi_i^2} y_i + a_i D^+ y_i\}|, \\ &\leq C|\varepsilon(y_i'' - \frac{D^+ D^- h^2}{\Psi_i^2} y_i) + a_i(y_i' - D^+ y_i)|, \\ &\leq C\varepsilon|y_i'' - D^+ D^- y_i| + C\varepsilon|(\frac{h^2}{\Psi_i^2} - 1)D^+ D^- y_i| + Ch|y_i''|, \\ &\leq C\varepsilon h^2 |y_i^{(4)}| + Ch|y_i''| + Ch|y_i''|, \\ &\leq C\varepsilon h^2 |y_i^{(4)}| + Ch|y_i''|. \end{aligned}$$

We use the estimate $\varepsilon|\frac{h^2}{\Psi^2} - 1| \leq Ch$, which can be derived from (10). Indeed, define $\rho = \frac{a_i h}{\varepsilon}, \rho \in (0, \infty)$. Then,

$$\varepsilon|\frac{h^2}{\Psi^2} - 1| = a_i h |\frac{1}{\exp(\rho) - 1} - \frac{1}{\rho}| =: a_i h Q(\rho).$$

By simplifying and writing the above equation explicitly, we obtain

$$Q(\rho) = \frac{\exp(\rho) - \rho - 1}{\rho(\exp(\rho) - 1)},$$

and we obtain that the limit is bounded as

$$\lim_{\rho \rightarrow 0} Q(\rho) = \frac{1}{2}, \quad \lim_{\rho \rightarrow \infty} Q(\rho) = 0.$$

Hence, for all $\rho \in (0, \infty)$, we have $Q(\rho) \leq C$. So, the error estimate in the discretization is bounded as

$$|L^N(y_i - Y_i)| \leq C\varepsilon h^2 |y_i^{(4)}| + Ch|y_i''|. \quad (17)$$

From (17) and boundedness of derivatives of solution in Lemma 4, we obtain

$$\begin{aligned} |L^N(y(x_i) - Y_i)| &\leq C\varepsilon h^2 \left| \left(1 + \varepsilon^{-4} \exp\left(\frac{-ax_i}{\varepsilon}\right) \right) \right| \\ &\quad + Ch \left| \left(1 + \varepsilon^{-2} \exp\left(\frac{-ax_i}{\varepsilon}\right) \right) \right| \end{aligned}$$

$$\begin{aligned}
&\leq Ch^2 \left| \left(\varepsilon + \varepsilon^{-3} \exp\left(\frac{-ax_i}{\varepsilon}\right) \right) \right| \\
&\quad + Ch \left| \left(1 + \varepsilon^{-2} \exp\left(\frac{-ax_i}{\varepsilon}\right) \right) \right| \\
&\leq Ch \left(1 + \sup_{x \in (0,1)} \left(\frac{\exp\left(\frac{-ax_i}{\varepsilon}\right)}{\varepsilon^3} \right) \right),
\end{aligned}$$

since $\varepsilon^{-3} > \varepsilon^{-2}$. \square

Most of the time during analysis, one encounters with exponential terms involving divided by the power function in ε , which are always the main cause of worry. For their careful consideration while proving the ε -uniform convergence, we prove the following lemma.

Lemma 6. For a fixed mesh and for $\varepsilon \rightarrow 0$, it holds

$$\begin{aligned}
\lim_{\varepsilon \rightarrow 0} \max_{1 \leq i \leq N-1} \left(\frac{\exp\left(\frac{-ax_i}{\varepsilon}\right)}{\varepsilon^m} \right) &= 0, \quad m = 1, 2, 3, \dots, \\
\lim_{\varepsilon \rightarrow 0} \max_{1 \leq i \leq N-1} \left(\frac{\exp\left(\frac{-a(1-x_i)}{\varepsilon}\right)}{\varepsilon^m} \right) &= 0, \quad m = 1, 2, 3, \dots,
\end{aligned}$$

where $x_i = ih$, $h = \frac{1}{N}$, $i = 1, 2, \dots, N-1$.

Proof. Consider the partition $[0, 1] := \{0 = x_0 < x_1 < \dots < x_{N-1} < x_N = 1\}$. For the interior grid points, we have

$$\begin{aligned}
\max_{1 \leq i \leq N-1} \frac{\exp\left(\frac{-ax_i}{\varepsilon}\right)}{\varepsilon^m} &\leq \frac{\exp\left(\frac{-ax_1}{\varepsilon}\right)}{\varepsilon^m} = \frac{\exp\left(\frac{-ah}{\varepsilon}\right)}{\varepsilon^m}, \\
\max_{1 \leq i \leq N-1} \frac{\exp\left(\frac{-a(1-x_i)}{\varepsilon}\right)}{\varepsilon^m} &\leq \frac{\exp\left(\frac{-a(1-x_{N-1})}{\varepsilon}\right)}{\varepsilon^m} = \frac{\exp\left(\frac{-ah}{\varepsilon}\right)}{\varepsilon^m},
\end{aligned}$$

as $x_1 = 1 - x_{N-1} = h$.

Then, applying L'Hospital's rule m times gives

$$\lim_{\varepsilon \rightarrow 0} \frac{\exp\left(\frac{-ah}{\varepsilon}\right)}{\varepsilon^m} = \lim_{r=\frac{1}{\varepsilon} \rightarrow \infty} \frac{r^m}{\exp(ahr)} = \lim_{r=\frac{1}{\varepsilon} \rightarrow \infty} \frac{m!}{(ah)^m \exp(ahr)} = 0.$$

\square

Theorem 2. Under the hypothesis of boundedness of discrete solution (i.e., it satisfies the discrete minimum principle), Lemma 6, and Theorem 1, the discrete solution satisfies the following bound:

$$\sup_{0 \leq \varepsilon \leq 1} \max_i |y_i - Y_i| \leq CN^{-1}. \quad (18)$$

Proof. Results from boundedness of solution, Lemma 6, and Theorem 1 give the required estimates. \square

6 Numerical examples and results

To validate the established theoretical results, we perform numerical experiments using the model problems of the form in (1)–(2).

Example 1. Consider the following problem:

$$\begin{cases} \varepsilon y''(x) + y'(x) = f(x), & x \in \Omega^- \cup \Omega^+, \\ y(0) - \varepsilon y'(0) = 1, & y(1) - y'(1) = -1, \end{cases}$$

where

$$f(x) = \begin{cases} 0.7, & 0 \leq x \leq 0.5, \\ -0.6, & 0.5 < x \leq 1. \end{cases}$$

Example 2. Consider the following problem:

$$\begin{cases} \varepsilon y''(x) + \frac{1}{1+x} y'(x) = f(x), & x \in \Omega^- \cup \Omega^+, \\ y(0) - \varepsilon y'(0) = 1, & y(1) - y'(1) = 1, \end{cases}$$

where

$$f(x) = \begin{cases} 1+x, & 0 \leq x \leq 0.5, \\ 4, & 0.5 < x \leq 1. \end{cases}$$

Having $y_j \equiv y_j^N$ (the approximated solution is obtained via the fitted operator finite difference method) for different values of h and ε , the maximum errors. Since the exact solution is not available, the maximum errors (denoted by E_ε^N) are evaluated using the double mesh principle [6], for fitted operator finite difference methods using the formula

$$E_\varepsilon^N := \max_{0 \leq j \leq n} |y_j^N - y_{2j}^{2N}|.$$

Furthermore, we will tabulate the ε -uniform error

$$E^N = \max_{0 < \varepsilon \leq 1} E_\varepsilon^N.$$

The numerical rate of convergence is computed using the formula [6]

$$r_\varepsilon^N := \frac{\log(E_\varepsilon^N) - \log(E_\varepsilon^{2N})}{\log(2)}.$$

and the ε -uniform rate of convergence is computed using

$$R^N = \frac{\log(E^N) - \log(E^{2N})}{\log(2)}.$$

Table 1: Maximum absolute errors for different values of ε and number of mesh size, N for Example 1.

ε	N=32	N=64	N=128	N=256	N=512
10^{-4}	2.0313e-02	1.0156 e-02	5.0781e-03	2.5391e-03	1.2695e-03
10^{-8}	2.0313e-02	1.0156 e-02	5.0781e-03	2.5391e-03	1.2695e-03
10^{-12}	2.0313e-02	1.0156 e-02	5.0781e-03	2.5391e-03	1.2695e-03
10^{-16}	2.0313e-02	1.0156 e-02	5.0781e-03	2.5391e-03	1.2695e-03
10^{-20}	2.0313e-02	1.0156 e-02	5.0781e-03	2.5391e-03	1.2695e-03
E^N	2.0313e-02	1.0156 e-02	5.0781e-03	2.5391e-03	1.2695e-03
R^N	1.0001	1.0000	1.0000	1.0001	

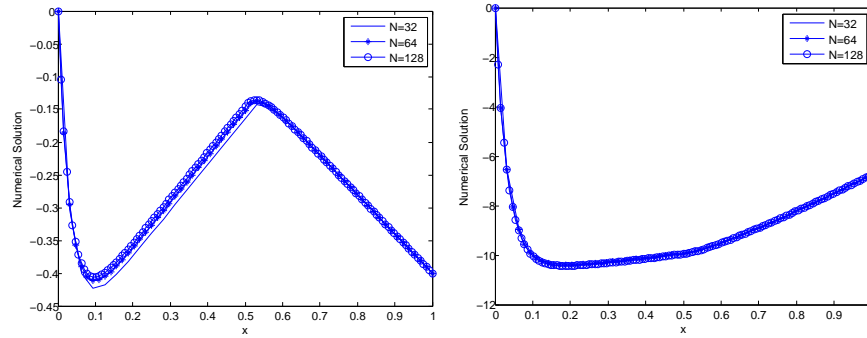


Figure 1: Behavior of numerical solution at $\varepsilon = 2^{-5}$ and different values of N for Examples 1 and 2, respectively.

Table 2: Comparison of maximum absolute errors and order of convergence for Example 1 at number of mesh points N .

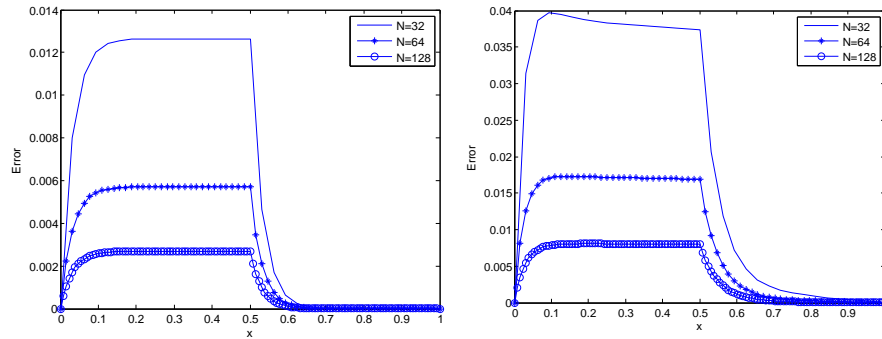
ε	N=64	N=128	N=256	N=512	N=1024
Present method					
E^N	1.0156e-02	5.0781e-03	2.5382e-03	1.2467e-03	5.5910e-04
R^N	1.0000	1.0000	1.0257	1.0257	
Method in [3]					
E^N	2.5658e-02	1.4128e-02	7.5359e-03	3.8225e-03	1.7536e-03
R^N	0.8605	0.90671	0.9793	1.1242	

Table 3: Maximum absolute errors for different values of ε and number of mesh size, N for Example 2.

ε	N=32	N=64	N=128	N=256	N=512
10^{-4}	8.3431e-02	4.1854e-02	2.0962e-02	1.0489e-02	5.2329e-03
10^{-8}	8.3431e-02	4.1854e-02	2.0962e-02	1.0489e-02	5.2329e-03
10^{-12}	8.3431e-02	4.1854e-02	2.0962e-02	1.0489e-02	5.2329e-03
10^{-16}	8.3431e-02	4.1854e-02	2.0962e-02	1.0489e-02	5.2329e-03
10^{-20}	8.3431e-02	4.1854e-02	2.0962e-02	1.0489e-02	5.2329e-03
E_N	8.3431e-02	4.1854e-02	2.0962e-02	1.0489e-02	5.2329e-03
R^N	0.9952	0.9976	0.9989	1.0032	

Table 4: Comparison of maximum absolute errors and order of convergence for Example 2 at number of mesh points N .

ε	N=64	N=128	N=256	N=512	N=1024
Present method					
E^N	4.1854e-02	2.0960e-02	1.0408e-02	4.7906e-03	2.0192e-03
R^N	0.9977	1.0099	1.1194	1.2464	
Method in [3]					
E^N	9.6698e-01	5.8056e-01	3.2795e-01	1.7313e-01	8.1501e-02
R^N	0.7364	0.8239	0.0.9216	1.0870	

Figure 2: Pointwise absolute error plot at $\varepsilon = 2^{-5}$ and different values of N for Examples 1 and 2, respectively.

7 Discussion and conclusion

This study introduced a uniformly convergent numerical method based on nonstandard finite difference method for solving singularly perturbed second-order ordinary differential equations of Robin type boundary value problems with discontinuous source term. Due to discontinuity in the source term,

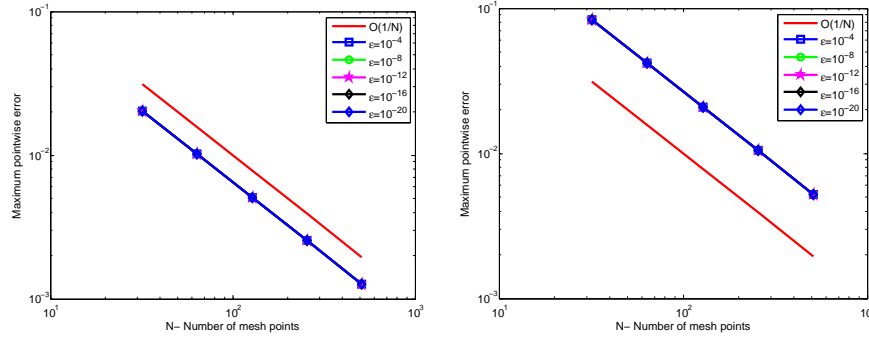


Figure 3: ε -uniform convergence with NSFDM in Log-Log scale for Examples 1 and 2, respectively.

there is an interior layer occurring. To fit the interior and boundary layer, a suitable nonstandard finite difference method on uniform mesh is constructed. The numerical results are tabulated in terms of maximum absolute errors, numerical rate of convergence, and uniform errors (see Tables 1–4) and compared with the results of the previously developed numerical methods existing in the literature (Tables 2 and 4). Furthermore, to see the position of the boundary layer, we plot the behavior of the numerical solution (see Figure 1), as the number of mesh points increases, the maximum pointwise errors decrease (see Figure 2) and the ε -uniform convergence of the method was shown using the log-log plot (see Figure 3). Unlike other fitted operator finite difference methods constructed in standard ways, the method that we presented in this paper is fairly simple to construct.

Acknowledgment.

The authors wish to express their thanks to Jimma University, College of Natural Sciences for financial support and the authors of literatures for the provided scientific aspects and idea for this work.

References

1. Bansal, K. and Sharma, K.K. *Parameter uniform numerical scheme for time dependent singularly perturbed convection-diffusion-reaction problems with general shift arguments*, Numer. Algorithms, 75(1) (2017) 113–145.
2. Chandru, M., Prabha, T. and Shanthi, V. *A hybrid difference scheme for a second-order singularly perturbed reaction-diffusion problem with nonsmooth data*, Int. J. Appl. Comput. Math. 1(1) (2015) 87–100.

3. Chandru, M. and Shanthi, V. *Fitted mesh method for singularly perturbed robin type boundary value problem with discontinuous source term*, Int. J. Appl. Comput. Math. 1(3) (2015) 491–501.
4. Chin, R.C.Y. and Krasny, R. *A hybrid asymptotic-finite element method for stiff two-point boundary value problems*, SIAM J. Sci. Statist. Comput. 4(2)2 (1983) 229–243.
5. Debela, H.G. and Duessa, G.F. *Uniformly Convergent Numerical Method for Singularly Perturbed Convection-Diffusion Type Problems with Non-local Boundary Condition*, Int. J. Numer. Methods Fluids. 92 (2020) 1914–1926.
6. Doolan, E.P., Miller, J.J.H. and Schilders, W.H.A. *Uniform numerical methods for problems with initial and boundary layers*, Boole Press, 1980.
7. Duessa, G.F. and Debela, H.G. *Numerical solution of singularly perturbed differential difference equations with mixed parameters*, Journal of Mathematical Modeling, (2021), 1–15.
8. Farrell, P.A., Miller, J.J.H., O’Riordan, E. and Shishkin, G.I. *Singularly perturbed differential equations with discontinuous source terms: In Proceedings of Analytical and Numerical Methods for Convection-Dominated and Singularly Perturbed Problems*, Lozenetz, Bulgaria, (1998) 23–32.
9. Farrell, P.A., Hegarty, A.F., Miller, J.J.H., O’Riordan, E. and Shishkin, G.I. *Singularly perturbed convection-diffusion problems with boundary and weak interior layers*, J. Comput. Appl. Math. 166(1) (2004) 133–151.
10. Farrell, P.A., Hegarty, A.F., Miller, J.J.H., O’Riordan, E. and Shishkin, G.I. *Global maximum norm parameter-uniform numerical method for a singularly perturbed convection-diffusion problem with discontinuous convection coefficient*, Mathematical and Computer Modelling, 40(11-12) (2004) 1375–1392.
11. Mickens, R.E. *Advances in the applications of nonstandard finite difference schemes*, World Scientific, 2005.
12. Mohapatra, J. and Natesan, S. *Uniform convergence analysis of finite difference scheme for singularly perturbed delay differential equation on an adaptively generated grid*, Numer. Math. Theory Methods Appl. 3(1) (2010) 1–22.
13. Roos, H.G., Stynes, M. and Tobiska, L. *Numerical methods for singularly perturbed differential equations*, Springer Series in Computational Mathematics, New York, 1996.
14. Roos, H.G. and Zarin, H. *A second-order scheme for singularly perturbed differential equations with discontinuous source term*, J. Numer. Math. 10(4) (2002) 275–289.

15. Shanthi, V. and Ramanujam, N. *Asymptotic numerical methods for singularly perturbed fourth order ordinary differential equations of convection-diffusion type*, Appl. Math. Comput. 133(2-3) (2002), 559–579.
16. Shanthi, V. and Ramanujam, N. *A boundary value technique for boundary value problems for singularly perturbed fourth-order ordinary differential equations*, Comput. Math. Appl. 47(10-12) (2004), 1673–1688.
17. Shanthi, V., Ramanujam, N. and Natesan, S. *Fitted mesh method for singularly perturbed reaction-convection-diffusion problems with boundary and interior layers*, J. Appl. Math. Comput. 22(1-2) (2006), 49–65.
18. Turkyilmazoglu, M. *Analytic approximate solutions of parameterized unperturbed and singularly perturbed boundary value problems*, Applied Mathematical Modelling, 35(8) (2011) 3879–3886.
19. Turkyilmazoglu, M. *Effective computation of exact and analytic approximate solutions to singular nonlinear equations of Lane–Emden–Fowler type*, Applied Mathematical Modelling, 37(14-15) (2013) 7539–7548.



Two new approximations to Caputo–Fabrizio fractional equation on non-uniform meshes and its applications

Z. Soori and A. Aminataei*

Abstract

We present two numerical approximations with non-uniform meshes to the Caputo–Fabrizio derivative of order α ($0 < \alpha < 1$). First, the L1 formula is obtained by using the linear interpolation approximation for constructing the second-order approximation. Next, the quadratic interpolation approximation is used for improving the accuracy in the temporal direction. Besides, we discretize the spatial derivative using the compact finite difference scheme. The accuracy of the suggested schemes is not dependent on the fractional α . The coefficients and the truncation errors are carefully investigated for two schemes, separately. Three examples are carried out to support the convergence orders and show the efficiency of the suggested scheme.

AMS subject classifications (2020): 26A33, 35K57.

Keywords: Numerical approximations, Caputo–Fabrizio fractional derivative, Diffusion equation, Advection equation, Non-uniform meshes.

1 Introduction

In recent years, some approximations have been proposed for the fractional derivative of order α , such as Grünwald–Letnikov, Lubich, and Caputo approximations [15, 16, 24, 28, 30, 27]. The schemes that are proposed until now to discretize the Caputo fractional derivative have been limited to the

*Corresponding author

Received 5 May 2021; revised 21 June 2021; accepted 8 July 2021

Zoliekha Soori

Faculty of Mathematics, K. N. Toosi University of Technology, P. O. Box: 16765-3381, Tehran, Iran. e-mail: zsoori@mail.kntu.ac.ir

Azim Aminataei

Faculty of Mathematics, K. N. Toosi University of Technology, P. O. Box: 16765-3381, Tehran, Iran. e-mail: ataei@kntu.ac.ir

accuracy of order $2 - \alpha$ ($0 < \alpha < 1$) on uniform meshes. One disadvantage of previously proposed methods is that when $\alpha \approx 1$, its accuracy may lead to poor accuracy. Thus, in this point of view, the numerical solutions of high dimensional partial fractional differential equations require a large number of computations. Besides, from the truncation error estimate of the methods on uniform meshes verify that the accuracy is dependent on the fractional order α . We are able to overcome these difficulties using Caputo–Fabrizio fractional derivative with a non-singular kernel. Afterward, we will obtain the second and third-orders accuracy in time that is independent of the fractional order α .

Let us consider the following time fractional diffusion and advection equations, respectively:

$$\begin{aligned} {}_0^{\text{CF}}\mathcal{D}_t^\alpha u(X, t) &= \frac{\partial^2 u(x, t)}{\partial x^2} + f(x, t), \quad x \in \Omega, \quad 0 \leq t \leq T, \\ {}_0^{\text{CF}}\mathcal{D}_t^\alpha u(X, t) &= \frac{\partial u(x, t)}{\partial x} + f(x, t), \quad x \in \Omega, \quad 0 \leq t \leq T, \end{aligned} \quad (1)$$

in which ${}_0^{\text{CF}}\mathcal{D}_t^\alpha$ is the α th Caputo–Fabrizio fractional derivative defined by

$${}_0^{\text{CF}}\mathcal{D}_t^\alpha u(X, t) = \frac{M(\alpha)}{1 - \alpha} \int_0^t u'(X, s) \exp\left(-\alpha \frac{t-s}{1-\alpha}\right) ds, \quad 0 < \alpha < 1, \quad (2)$$

where $M(\alpha)$ is a normalization function such that $M(0) = M(1) = 1$.

In 2015, Caputo and Fabrizio [7] suggested a new definition of fractional derivative based on the exponential kernel. They considered two different representations for the temporal and the spatial variables. It is important and interesting that this approach describes the behaviour of classical viscoelastic materials, electromagnetic systems, thermal media, and so on. Another interesting property of this definition is that it opens up new avenues in the mechanical phenomena, related to plasticity, fatigue, damage, and electromagnetic hysteresis [7].

Recently, studies of Caputo–Fabrizio fractional derivatives have been carried out by some authors. Authors of [6] investigated the existence of a solution for two high-order fractional integro-differential equations including the Caputo–Fabrizio derivative. Atangana and Alqahtani [4] considered a numerical approximation of the space and time Caputo–Fabrizio fractional derivative in connection with ground water pollution equation. In [18], the authors presented a Crank–Nicolson finite difference scheme to solve fractional Cattaneo equation by a new fractional derivative. Furthermore, they analyzed the stability and convergence order of the scheme. The main aim of [9] is to prove the existence and uniqueness of the flow of water within a confined aquifer with Caputo–Fabrizio fractional diffusion for the spatial and the temporal variables. In [12], the authors applied the Ritz method with known basis functions for a type of Fokker–Planck equation with Caputo–Fabrizio fractional derivative. In 2017, Mirza and Vieru [21] proposed the fundamen-

tal solutions to time-fractional advection-diffusion equation without singular kernel. They applied the Laplace transform and Fourier transforms with respect to the temporal variable and the space coordinates, respectively. In [19], the authors constructed the shifted Legendre polynomials operational matrix in order to solve problems with left-sided Caputo–Fabrizio operator. A second-order scheme for the space fractional diffusion equation with Caputo–Fabrizio is provided in [26]. The main aim of [10] is to solve two problems in nonlocal quantum mechanics wherein the nonlocal Schrödinger equation has been transformed to an ordinary linear differential equation. Other interesting papers in the field of the Caputo–Fabrizio derivative are found in [1, 2, 3, 5, 8, 11, 13, 14, 20, 22, 23, 25, 17].

The main goal of this paper is to derive two new formulas to approximate the Caputo–Fabrizio derivative of order α ($0 < \alpha < 1$). For this purpose, we use the linear and the quadratic interpolation approximations on non-uniform meshes for obtaining the second and the third orders accuracy. Besides, we discretize the spatial derivative using the compact finite difference scheme. The advantages of the present paper are in the following two aspects, that is, the obtained accuracy is independent of the fractional α and gives a new high-order accuracy to the time fractional derivative in Caputo–Fabrizio’s sense on non-uniform meshes. In this paper, much attention is paid to the numerical aspects. To our knowledge, our interest in Caputo–Fabrizio derivative is due to the necessity of using a model describing the behaviour of classical viscoelastic materials, thermal media, electromagnetic systems, and so on. In fact, the original definition of fractional derivative appears to be particularly convenient for those mechanical phenomena, related to plasticity, fatigue, damage and with electromagnetic hysteresis. In fact, the Caputo–Fabrizio derivative fits to describe material heterogeneities and structures with different scales. Hence, we have focused on non-uniform meshes.

The rest of the paper is organized as follows. In Section 2, the derivation of the new method on any non-uniform meshes for the Caputo–Fabrizio fractional derivative of order α ($0 < \alpha < 1$) in both cases of the second and the third order is developed. Three examples are given in Section 3 to support the theoretical analysis. Finally concluding remarks are given in Section 4.

2 Derivation of new method on non-uniform meshes

2.1 The second-order approximation

In this section, we focus our attention on deriving the new fractional numerical differentiation formula in details. By Caputo–Fabrizio fractional derivative, we have

$$\begin{aligned}
{}_0^{\text{CF}}D_t^\alpha u(t)|_{t=t_k} &= \frac{1}{1-\alpha} \int_0^t u'(s) \exp\left(-\alpha \frac{t-s}{1-\alpha}\right) ds \\
&= \frac{1}{1-\alpha} \sum_{k=1}^n \int_{t_{k-1}}^{t_k} u'(s) \exp\left(-\alpha \frac{t_n-s}{1-\alpha}\right) ds,
\end{aligned} \quad (3)$$

where $0 = t_0 < t_1 < \dots < t_N = T$. We denote the time step by $\tau_n = t_n - t_{n-1}$, $1 \leq n \leq N$, and let $\tau_{Max} = \max_{1 \leq l \leq N} \tau_l$.

To explain the process of deriving formula, we apply the linear interpolation polynomial using points $(t_{k-1}, u(t_{k-1}))$ and $(t_k, u(t_k))$ for approximating $u'(s)$ as follows:

$$\begin{aligned}
\Pi_{1,k}u(t) &= u(t_k) + (t - t_k) \frac{u(t_k) - u(t_{k-1})}{t_k - t_{k-1}} \\
&= u(t_{k-1}) \frac{t_k - t}{\tau_k} + u(t_k) \frac{t - t_{k-1}}{\tau_k}
\end{aligned} \quad (4)$$

and

$$(\Pi_{1,k}u(t))' = \frac{u(t_k) - u(t_{k-1})}{\tau_k}. \quad (5)$$

Then, the interpolation error formula is given by

$$\begin{aligned}
u(t) - \Pi_{1,k}u(t) &= \frac{u''(\xi_k)}{2!} (t - t_{k-1})(t - t_k), \quad t \in [t_{k-1}, t_k], \quad \xi_k \in (t_{k-1}, t_k), \\
&1 \leq k \leq n.
\end{aligned} \quad (6)$$

Now, substituting (5) into (3), we obtain the L1 formula as follows:

$$\begin{aligned}
{}_0^{\text{CF}}D_t^\alpha u(t)|_{t=t_k} &= \frac{1}{1-\alpha} \sum_{k=1}^n \frac{u(t_k) - u(t_{k-1})}{\tau_k} \int_{t_{k-1}}^{t_k} \exp\left(-\alpha \frac{t_n-s}{1-\alpha}\right) ds \\
&= \frac{1}{\alpha} \sum_{k=1}^n (u_k - u_{k-1}) M_k^n,
\end{aligned} \quad (7)$$

where $M_k^n = \frac{1}{\tau_k} \left(\exp\left(-\alpha \frac{t_n-t_k}{1-\alpha}\right) - \exp\left(-\alpha \frac{t_n-t_{k-1}}{1-\alpha}\right) \right)$.

Lemma 1. For any $1 \leq n \leq N$, we have $M_k^n > 0$ and $M_{k+1}^n > M_k^n$.

Proof. Note that $\frac{-\alpha}{1-\alpha}(t_n - t_k) > \frac{-\alpha}{1-\alpha}(t_n - t_{k-1})$ for $0 < \alpha < 1$ and $\exp(x)$ is a monotone increasing function. Thus one can verify that

$$M_k^n = \frac{1}{\tau_k} \left(\exp\left(-\alpha \frac{t_n-t_k}{1-\alpha}\right) - \exp\left(-\alpha \frac{t_n-t_{k-1}}{1-\alpha}\right) \right) > 0.$$

From the mean value theorem for integrals, we have

$$M_k^n = \frac{1}{\tau_k} \int_{t_{k-1}}^{t_k} \exp\left(-\alpha \frac{t_n - s}{1 - \alpha}\right) ds = \exp\left(-\alpha \frac{t_n - \xi_k}{1 - \alpha}\right), \quad \xi_k \in (t_{k-1}, t_k).$$

Clearly, $\exp\left(-\alpha \frac{t_n - s}{1 - \alpha}\right)$ is a monotone increasing function, then the second statement of the lemma follows immediately. \square

Theorem 1. Suppose $u(t) \in C^2[0, T]$. For any $0 < \alpha < 1$, it holds that

$$|R(u(t_k))| \leq \frac{1}{1 - \alpha} \max_{t_0 \leq t \leq t_n} \frac{|u''(t)|}{8} \tau_{\max}^2. \quad (8)$$

Proof. From (2) and (6), we get

$$\begin{aligned} R(u(t_k)) &= \frac{1}{1 - \alpha} \left[\sum_{k=1}^n \int_{t_{k-1}}^{t_k} (u(s) - \Pi_{1,k} u(s))' \exp\left(-\alpha \frac{t_n - s}{1 - \alpha}\right) ds \right] \\ &= \frac{1}{1 - \alpha} \sum_{k=1}^n \underbrace{\left[(u(s) - \Pi_{1,k} u(s)) \exp\left(-\alpha \frac{t_n - s}{1 - \alpha}\right) \right]_{t_{k-1}}^{t_k}}_{=0} \\ &\quad - \int_{t_{k-1}}^{t_k} (u(s) - \Pi_{1,k} u(s)) \exp\left(-\alpha \frac{t_n - s}{1 - \alpha}\right) \frac{\alpha}{1 - \alpha} ds \\ &= \frac{-\alpha}{(1 - \alpha)^2} \left[\sum_{k=1}^n \int_{t_{k-1}}^{t_k} \frac{u''(\nu_k)}{2} (s - t_{k-1})(s - t_k) \exp\left(-\alpha \frac{t_n - s}{1 - \alpha}\right) ds \right] \\ &\leq \frac{\alpha}{(1 - \alpha)^2} \frac{|u''(\nu)|}{2} \frac{\tau_k^2}{4} \underbrace{\int_{t_0}^{t_n} \exp\left(-\alpha \frac{t_n - s}{1 - \alpha}\right) ds}_{\leq \frac{1 - \alpha}{\alpha}} \\ &\leq \frac{\alpha}{(1 - \alpha)^2} \frac{|u''(\nu)|}{8} \tau_{\max}^2 \frac{1 - \alpha}{\alpha}, \quad \nu \in (t_0, t_n) \quad \nu_k \in (t_{k-1}, t_k). \end{aligned}$$

This proves the desired formula. \square

We apply the time step over the non-uniform mesh defined as [29]

$$\tau_n = (N + 1 - n)\mu, \quad 1 \leq n \leq N, \quad (9)$$

where $\mu = \frac{2T}{N(N+1)}$.

2.2 The third-order approximation

Adding an additional point $(t_{k-2}, u(t_{k-2}))$ for $k \geq 2$, we obtain a quadratic interpolation function $\Pi_{2,k} u(t)$ of $u(t)$ as follows:

$$\begin{aligned}
\Pi_{2,k}u(t) &= \Pi_{1,k}u(t) + \frac{1}{t_k - t_{k-2}} \left[\frac{u(t_k) - u(t_{k-1})}{t_k - t_{k-1}} - \frac{u(t_{k-1}) - u(t_{k-2})}{t_{k-1} - t_{k-2}} \right] \\
&\quad \times (t - t_k)(t - t_{k-1}) \\
&= \Pi_{1,k}u(t) + \frac{1}{\tau_k + \tau_{k-1}} \left[\frac{u(t_k) - u(t_{k-1})}{\tau_k} - \frac{u(t_{k-1}) - u(t_{k-2})}{\tau_{k-1}} \right] \\
&\quad \times (t - t_k)(t - t_{k-1}), \quad t \in [t_{k-1}, t_k], \\
(\Pi_{2,k}u(t))' &= \frac{u(t_k) - u(t_{k-1})}{\tau_k} + (2t - (t_k + t_{k-1}))\mathcal{A}_t u(t_k).
\end{aligned} \tag{10}$$

For simplicity in what follows, we define:

$$\mathcal{A}_t u_k = \frac{1}{\tau_k + \tau_{k-1}} \left[\frac{u_k - u_{k-1}}{\tau_k} - \frac{u_{k-1} - u_{k-2}}{\tau_{k-1}} \right],$$

and

$$\begin{aligned}
u(t) - \Pi_{2,k}u(t) &= \frac{u'''(\eta_k)}{6} (t - t_{k-2})(t - t_{k-1})(t - t_k), \\
t &\in [t_{k-1}, t_k], \quad \eta_k \in (t_{k-2}, t_k), \quad 2 \leq k \leq n.
\end{aligned} \tag{11}$$

We substitute (10) into (2) to obtain a new approximation of the Caputo-Fabrizio derivative as follows:

$$\begin{aligned}
{}_0^{\text{CF}}\mathcal{D}_t^\alpha u(t)|_{t=t_n} &= \frac{1}{1-\alpha} \sum_{k=1}^n \int_{t_{k-1}}^{t_k} u'(s) \exp\left(-\alpha \frac{t_n - s}{1-\alpha}\right) ds \\
&\approx \frac{1}{1-\alpha} \left[\int_{t_0}^{t_1} (\Pi_{1,1}u(s))' \exp\left(-\alpha \frac{t_n - s}{1-\alpha}\right) ds \right. \\
&\quad \left. + \sum_{k=2}^n \int_{t_{k-1}}^{t_k} (\Pi_{2,k}u(s))' \exp\left(-\alpha \frac{t_n - s}{1-\alpha}\right) ds \right] \\
&= \frac{1}{1-\alpha} \left[\frac{u_1 - u_0}{\tau_1} \int_{t_0}^{t_1} \exp\left(-\alpha \frac{t_n - s}{1-\alpha}\right) ds \right. \\
&\quad \left. + \sum_{k=2}^n \int_{t_{k-1}}^{t_k} \left[\frac{u_k - u_{k-1}}{\tau_k} + \mathcal{A}_t u_k (2s - (t_{k-1} + t_k)) \right] \right. \\
&\quad \left. \times \exp\left(-\alpha \frac{t_n - s}{1-\alpha}\right) ds \right]
\end{aligned}$$

$$\begin{aligned}
&= \frac{1}{1-\alpha} \left[\sum_{k=1}^n \frac{u_k - u_{k-1}}{\tau_k} \int_{t_{k-1}}^{t_k} \exp\left(-\alpha \frac{t_n - s}{1-\alpha}\right) ds \right. \\
&\quad \left. + \sum_{k=2}^n \mathcal{A}_t u_k \int_{t_{k-1}}^{t_k} (2s - (t_{k-1} + t_k)) \exp\left(-\alpha \frac{t_n - s}{1-\alpha}\right) ds \right] \\
&= {}_0^{\text{CF}} D_t^\alpha u(t)|_{t=t_n} + \frac{1}{\alpha^2} \sum_{k=2}^n B_k^n \mathcal{A}_t u_k,
\end{aligned} \tag{12}$$

where

$$\begin{aligned}
B_k^n &= 2(\alpha - 1) \left[\exp\left(-\alpha \frac{t_n - t_k}{1-\alpha}\right) - \exp\left(-\alpha \frac{t_n - t_{k-1}}{1-\alpha}\right) \right] \\
&\quad + \alpha \tau_k \left[\exp\left(-\alpha \frac{t_n - t_k}{1-\alpha}\right) + \exp\left(-\alpha \frac{t_n - t_{k-1}}{1-\alpha}\right) \right].
\end{aligned} \tag{13}$$

Moreover, ${}_0^{\text{CF}} D_t^\alpha$ is the $L1$ method for non-uniform time grid in Section 2, where, we define

$${}_0^{\text{CF}} \mathbb{D}_t^\alpha u(t)|_{t=t_n} = {}_0^{\text{CF}} D_t^\alpha u(t)|_{t=t_n} + \frac{1}{\alpha^2} \sum_{k=2}^n B_k^n \mathcal{A}_t u_k, \tag{14}$$

wherein, we define a new operator $\mathcal{D}_t^\alpha u(t)$, which is the new fractional numerical differentiation operator for the Caputo–Fabrizio fractional derivative ${}_0^{\text{CF}} \mathbb{D}^\alpha$.

The following lemma states the property of coefficients B_k^n .

Lemma 2. For any α ($0 < \alpha < 1$), let

$$\begin{aligned}
B_k^n &= 2(\alpha - 1) \left[\exp\left(-\alpha \frac{t_n - t_k}{1-\alpha}\right) - \exp\left(-\alpha \frac{t_n - t_{k-1}}{1-\alpha}\right) \right] \\
&\quad + \alpha \tau_k \left[\exp\left(-\alpha \frac{t_n - t_k}{1-\alpha}\right) + \exp\left(-\alpha \frac{t_n - t_{k-1}}{1-\alpha}\right) \right], \quad 2 \leq k \leq n.
\end{aligned}$$

It holds that

$$B_n^n > B_{n-1}^n > \cdots > B_k^n > B_{k-1}^n > \cdots > B_2^n > 0.$$

Proof. To prove our statement, we apply the error representation of the trapezoidal formula. For this purpose, we consider the function $-2\alpha \exp\left(-\alpha \frac{t_n - s}{1-\alpha}\right)$ on the interval $[t_{k-1}, t_k]$. Then

$$\begin{aligned}
B_k^n &= -2\alpha \left[\int_{t_{k-1}}^{t_k} \exp\left(-\alpha \frac{t_n - s}{1 - \alpha}\right) ds - \frac{\tau_k}{2} \left(\exp\left(-\alpha \frac{t_n - t_k}{1 - \alpha}\right) \right. \right. \\
&\quad \left. \left. + \exp\left(-\alpha \frac{t_n - t_{k-1}}{1 - \alpha}\right) \right) \right] \\
&= -2\alpha \left(-\frac{1}{12}\right) \left(\exp\left(-\alpha \frac{t_n - s}{1 - \alpha}\right) \right)'' \Big|_{s=\xi_k} \\
&= \frac{\alpha}{6} \frac{\alpha^2}{(1 - \alpha)^2} \exp\left(-\alpha \frac{t_n - \xi_k}{1 - \alpha}\right), \quad \xi_k \in (t_{k-1}, t_k).
\end{aligned}$$

Since $\exp\left(-\alpha \frac{t_n - s}{1 - \alpha}\right) > 0$ and $\frac{\alpha}{6} \frac{\alpha^2}{(1 - \alpha)^2} > 0$, these imply that $B_k^n > 0$ for $2 \leq k \leq n$. Besides, $\exp\left(-\alpha \frac{t_n - s}{1 - \alpha}\right)$ is a monotone increasing function with respect to s on $[0, T]$ and this completes the proof. \square

Theorem 2. Suppose $u(t) \in C^3[0, T]$. For any $0 < \alpha < 1$, it holds that

$$|R(u(t_k))| \leq \left[\frac{\alpha}{(1 - \alpha)^2} \max_{t_0 \leq t \leq t_1} \frac{|u''(t)|}{8} + \max_{t_0 \leq t \leq t_n} \frac{|u'''(t)|}{12} \frac{\alpha}{1 - \alpha} \right] \tau_{\max}^3. \quad (15)$$

Proof. To prove this, we see from (2), (6), and (11) that

$$\begin{aligned}
R(u(t_k)) &= \frac{1}{1 - \alpha} \left[\int_{t_0}^{t_1} (u(s) - \Pi_{1,1}u(s))' \exp\left(-\alpha \frac{t_n - s}{1 - \alpha}\right) ds \right. \\
&\quad \left. + \sum_{k=2}^n \int_{t_{k-1}}^{t_k} (u(s) - \Pi_{2,k}u(s))' \exp\left(-\alpha \frac{t_n - s}{1 - \alpha}\right) ds \right] \\
&= \frac{1}{1 - \alpha} \left[\underbrace{(u(s) - \Pi_{1,1}u(s)) \exp\left(-\alpha \frac{t_n - s}{1 - \alpha}\right)}_{=0} \Big|_{t_0}^{t_1} \right. \\
&\quad \left. - \int_{t_0}^{t_1} (u(s) - \Pi_{1,1}u(s)) \exp\left(-\alpha \frac{t_n - s}{1 - \alpha}\right) \frac{\alpha}{1 - \alpha} ds \right] \\
&\quad + \frac{1}{1 - \alpha} \left[\underbrace{\sum_{k=2}^n \int_{t_{k-1}}^{t_k} (u(s) - \Pi_{2,k}u(s))' \exp\left(-\alpha \frac{t_n - s}{1 - \alpha}\right) ds}_{=0} \right. \\
&\quad \left. - \sum_{k=2}^n \int_{t_{k-1}}^{t_k} (u(s) - \Pi_{2,k}u(s)) \exp\left(-\alpha \frac{t_n - s}{1 - \alpha}\right) \frac{\alpha}{1 - \alpha} ds \right]
\end{aligned}$$

$$\begin{aligned}
&\leq \frac{\alpha}{(1-\alpha)^2} \left[\frac{|u''(\xi_1)|}{2} \frac{\tau^2}{4} \underbrace{\int_{t_0}^{t_1} \exp\left(-\alpha \frac{t_n-s}{1-\alpha}\right) ds}_{\tau_1} \right. \\
&\quad \left. + \frac{|u'''(\nu)|}{6} \frac{\tau_k^2}{4} (\tau_{k-1} + \tau_k) \underbrace{\int_{t_1}^{t_n} \exp\left(-\alpha \frac{t_n-s}{1-\alpha}\right) ds}_{\leq \frac{1-\alpha}{\alpha}} \right], \\
&\xi_1 \in (t_0, t_1), \nu \in (t_0, t_n).
\end{aligned}$$

The second term of last inequality results by the following remark in [29]:

Remark 1. With regard to $\max_{t_0 \leq s \leq t_n} |(s-t_k)(s-t_{k-1})| = \frac{\tau_k^2}{4}$ is obtained at $s = t_{k-1} + \frac{\tau_k}{2}$, consequently we have:

$$\begin{aligned}
(s-t_{k-2})(s-t_{k-1})(s-t_k) &= \frac{\tau_k^2}{4} (t_{k-1} + \frac{\tau_k}{2} - t_{k-2}) \\
&= \frac{\tau_k^2}{4} (\tau_{k-1} + \frac{\tau_k}{2}) \leq \frac{\tau_k^2}{4} (\tau_{k-1} + \tau_k).
\end{aligned}$$

Besides, we know that $\tau_{Max} = \max_{1 \leq l \leq N} \tau_l$. This proves (15).

□

3 Numerical applications of the examples

In the current section, the efficiency of the suggested scheme for the time Caputo–Fabrizio fractional diffusion and advection equations are presented on three numerical examples in one dimension. The accuracy and the stability of the suggested scheme in the paper for different values of M and N are tested. In order to carry out our numerical examples, we have used the Maple 18 software with a PC of 4 GHz CPU and 6 GB memory. The accuracy of the proposed scheme is measured by the following error norm

$$e(N, M) = \max_{1 \leq i \leq M-1} |u(x_i, t_N) - u_i^N|.$$

We denote the numerical convergence orders by

$$Rate = \log_2 \left(\frac{e(N/2, M)}{e(N, M)} \right).$$

Example 1. Suppose $0 < \alpha < 1$. Let $u(t) = \sin(4t)$. The exact solution is obtained from the definition of the Caputo–Fabrizio (2) without the variable x . Denote $e(N) = |u(t_N) - u_N|$.

Table 1: Numerical convergence orders in temporal direction for Example 1

α	N	The second-order		The third-order	
		$e(N)$	Rate	$e(N)$	Rate
0.5	5	5.6188×10^{-2}	—	2.0699×10^{-2}	—
	10	1.5043×10^{-2}	1.9012	2.3793×10^{-3}	3.2100
	20	3.9207×10^{-3}	1.9399	2.6437×10^{-4}	3.1699
	40	1.0028×10^{-3}	1.9671	3.1297×10^{-5}	3.0785
	80	$53722. \times 10^{-4}$	1.9827	3.6530×10^{-6}	3.0989
0.9	5	3.4485×10^{-3}	—	5.9336×10^{-2}	—
	10	9.3840×10^{-4}	1.8777	7.9779×10^{-3}	2.8946
	20	2.4994×10^{-4}	1.9086	1.0192×10^{-3}	2.9886
	40	6.4364×10^{-5}	1.9572	1.2836×10^{-4}	2.9892
	80	1.6314×10^{-5}	1.9801	1.5964×10^{-5}	3.0073

From Table 1, this fact is extracted that the computational orders of our schemes are independent of the fractional order α . This means that with changing values α , the computational orders do not change and the orders of convergence of these two cases the second-order and three-order schemes should be 2 and 3, respectively.

Example 2. Consider the time fractional diffusion equation [29]

$$\begin{cases} {}^C_0\mathcal{D}_t^\alpha u(x, t) = \frac{\partial^2 u(x, t)}{\partial x^2} + f(x, t), & 0 < x < 1, \ 0 < t \leq 1, \\ u(x, 0) = u^0(x), & 0 \leq x \leq 1, \\ u(0, t) = \Phi(t), \ u(1, t) = \varphi(t), & 0 < t \leq 1. \end{cases} \quad (16)$$

The exact solution of (16) is $u(x, t) = \sin(\pi x)t^2$. The functions can be obtained by substituting $u(x, t)$ into (16).

Remark 2. For discretizing the term $\frac{\partial^2 u(x, t)}{\partial x^2}$, we apply the fourth-order CFD scheme as follows:

$$\frac{\partial^2 u(x, t)}{\partial x^2} \Big|_{x=x_i} = \frac{\delta_x^2}{1 + \frac{h^2}{12}\delta_x^2} u(x_i, t) + O(h^4).$$

Now, we consider the fourth-order approximation of the second derivative of u at point x_i as follows:

$$\frac{\partial^2 u(x, t)}{\partial x^2} \Big|_{i, n} = \frac{\delta_x^2}{1 + \frac{h^2}{12}\delta_x^2} u_i^n,$$

where u_i^n denotes the numerical solution at (x_i, t_n) . Then, a difference scheme using third-order formula can be obtained as

$$\mathcal{H}_0^{\text{CF}} \mathbb{D}_t^\alpha u_i^n = \delta_x^2 u_i^n + \mathcal{H} f_i^n, \quad 1 \leq i \leq M-1, \quad 1 \leq n \leq N, \quad (17)$$

$$u_0^n = \Phi(t_n), \quad u_1^n = \varphi(t_n), \quad 1 \leq n \leq N, \quad (18)$$

$$u_i^0 = u^0(x_i), \quad 0 \leq i \leq M, \quad (19)$$

where

$$\delta_x^2 v_i^n = \frac{1}{h} (\delta_x v_{i+\frac{1}{2}}^n - \delta_x v_{i-\frac{1}{2}}^n).$$

Moreover

$$\mathcal{H} v_i = \begin{cases} \frac{1}{12} (v_{i+1} + 10v_i + v_{i-1}), & 1 \leq i \leq M-1, \\ v_i, & i = 0 \text{ or } M. \end{cases}$$

It can be seen that

$$\mathcal{H} v_i = \left(I + \frac{h^2}{12} \delta_x^2 \right) v_i, \quad 1 \leq i \leq M-1.$$

Lemma 3. (See [31]). consider the function $g(x) \in \mathcal{C}^6[x_{i-1}, x_{i+1}]$, and let $\xi(s) = 5(1-s)^3 - 3(1-s)^5$. Then

$$\begin{aligned} & \frac{g''(x_{i+1}) + 10g''(x_i) + g''(x_{i-1}))}{12} \\ &= \frac{g(x_{i+1}) - 2g(x_i) + g(x_{i-1}))}{h^2} + \frac{h^4}{360} \int_0^1 [g^6(x_i - sh) + g^6(x_i + sh)] \xi(s) ds. \end{aligned}$$

Having seen Tables 2 and 3, we observe that the third-order scheme produces better results than the second-order scheme. In the Caputo's sense, the accuracy of the presented method is dependent on α . In this case, the computational orders for $\alpha = 0.4, 0.6$ and 0.8 are 1.6, 1.4 and 1.2, respectively, when the theoretical order is $3 - \alpha$, whereas the accuracy of the presented method is not dependent on the fractional α . Table 4 illustrates the error and CPU time of the third-order and the second-order schemes. Having seen Table 4, we conclude that the third-order scheme produces more accurate results than the second-order scheme. Besides, the third-order scheme needs fewer temporal grid size and less CPU time for bigger N .

Figure 1 exhibits the solution curves at final time $T = 1$ for different values of $\alpha = 0.1, 0.5$ and 0.9 with $M = N = 50$ for Example 2. Figure 2 shows the comparison of the absolute errors wherein the third-order scheme is more accurate than the second order-scheme. The plots of absolute error and the numerical solution for $\alpha = 0.1$ with $M = N = 50$ for Example 2 are shown in Figure 3.

Table 2: Numerical convergence orders in temporal direction with $M = 50$ for Example 2

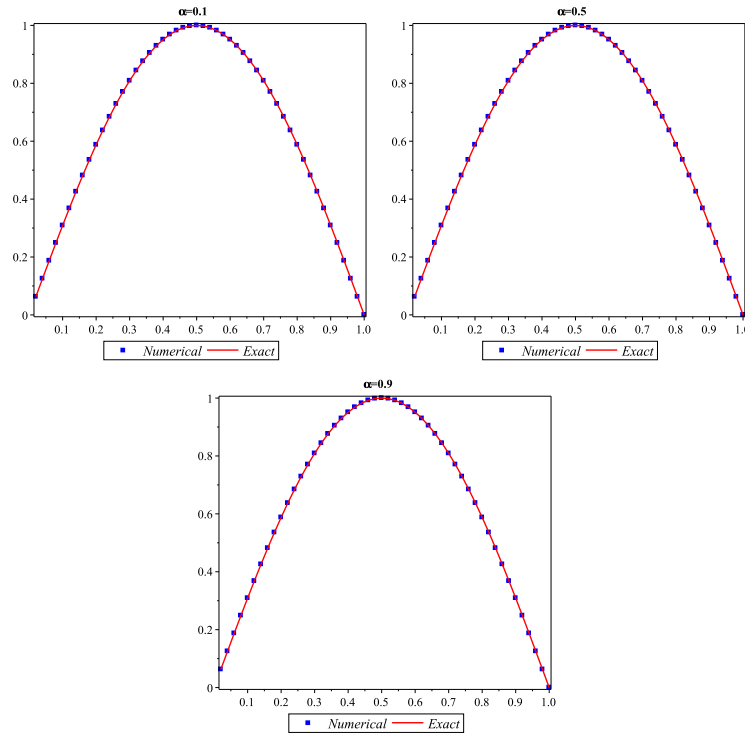
α	N	The second-order		The third-order	
		$e(N, M)$	$Rate$	$e(N, M)$	$Rate$
0.25	5	3.6444×10^{-4}	—	1.9185×10^{-4}	—
	10	9.9230×10^{-5}	1.8768	3.0437×10^{-5}	2.6561
	20	2.5987×10^{-5}	1.8299	4.4476×10^{-6}	2.7747
	40	6.6615×10^{-6}	1.9639	5.4377×10^{-7}	3.0320
0.5	5	1.1051×10^{-3}	—	5.2234×10^{-4}	—
	10	2.9975×10^{-4}	1.8823	7.9398×10^{-5}	2.7206
	20	7.8515×10^{-5}	1.9327	1.1188×10^{-5}	2.8272
	40	2.0126×10^{-5}	1.9639	1.5315×10^{-6}	2.8689
0.75	5	2.6813×10^{-3}	—	9.2499×10^{-4}	—
	10	7.2606×10^{-4}	1.8848	1.2632×10^{-4}	2.8724
	20	1.8975×10^{-4}	1.9360	1.6398×10^{-5}	2.9455
	40	4.8687×10^{-5}	1.9625	2.2262×10^{-6}	2.8809

Table 3: Numerical convergence orders in temporal direction with $M = 50$ for Example 2

α	N	The second-order		The third-order	
		$e(N, M)$	$Rate$	$e(N, M)$	$Rate$
0.4	5	7.4249×10^{-4}	—	3.7053×10^{-4}	—
	10	2.0170×10^{-4}	1.8802	5.7480×10^{-5}	2.8434
	20	5.2738×10^{-5}	1.9353	8.0087×10^{-6}	2.8434
	40	1.3679×10^{-5}	1.9469	1.2630×10^{-7}	2.6647
0.6	5	1.5958×10^{-3}	—	6.9588×10^{-4}	—
	10	4.3238×10^{-4}	1.8839	1.0289×10^{-4}	2.7577
	20	1.1306×10^{-4}	1.9352	1.3961×10^{-5}	2.8816
	40	2.8973×10^{-5}	1.9643	1.7908×10^{-6}	2.9627
0.8	5	3.1781×10^{-3}	—	9.4828×10^{-4}	—
	10	8.6131×10^{-4}	1.8836	1.2325×10^{-4}	2.9437
	20	2.2530×10^{-4}	1.9347	1.5675×10^{-5}	2.9751
	40	5.7678×10^{-5}	1.9658	1.9745×10^{-6}	2.9889

Table 4: The errors and CPU time (seconds) of the third-order and the second-order schemes for Example 2

α	The third-order ($M = 50$)			The second-order ($M = 50$)		
	N	$e(N, M)$	$CPU(s)$	N	$e(N, M)$	$CPU(s)$
0.7	8	9.3694×10^{-4}	0.03	5	8.6414×10^{-4}	0.09
	24	1.1214×10^{-4}	0.06	12	7.2084×10^{-5}	0.1
	72	1.2902×10^{-5}	13.10	24	9.6323×10^{-6}	5.75
0.8	8	1.3174×10^{-3}	0.06	5	9.4828×10^{-4}	0.09
	24	1.5769×10^{-4}	0.09	10	1.2325×10^{-4}	0.1
	72	1.8033×10^{-5}	14.30	20	1.5675×10^{-5}	3.86
0.9	8	1.9637×10^{-3}	0.07	4	1.6778×10^{-3}	0.07
	24	2.3611×10^{-4}	0.09	8	1.7168×10^{-4}	0.09
	72	2.7014×10^{-5}	14.53	16	1.9223×10^{-5}	3.05

Figure 1: The solution curves at $T = 1$ with $M = N = 50$ for Example 2

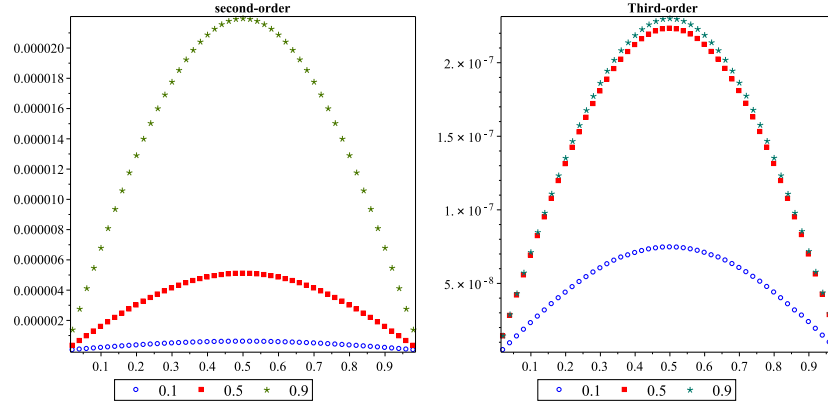


Figure 2: Comparison of the absolute errors for the second-order (left) and the third-order (right) schemes with $M = N = 50$ for Example 2

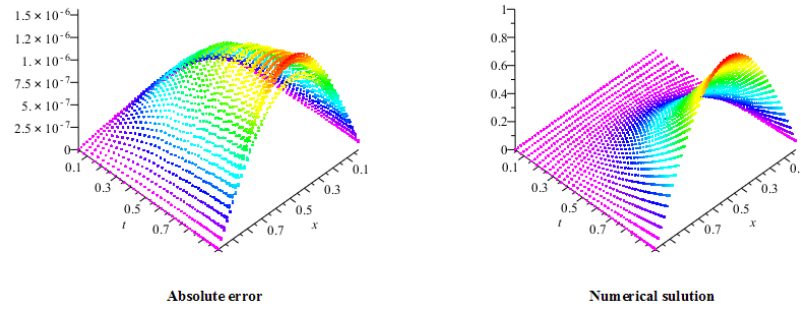


Figure 3: Plots of the absolute error (left) with $M = N = 50$ and the numerical solution (right) for Example 2

Example 3. Consider the time fractional advection equation [29]

$$\begin{cases} {}^C_0\mathcal{D}_t^\alpha u(x, t) = \frac{\partial u(x, t)}{\partial x} + f(x, t), & 0 < x < 1, \quad 0 < t \leq 1, \\ u(x, 0) = u^0(x), & 0 \leq x \leq 1, \\ u(0, t) = \Phi(t), \quad u(1, t) = \varphi(t), & 0 < t \leq 1. \end{cases} \quad (20)$$

Remark 3. For discretizing the term $\frac{\partial u(x, t)}{\partial x}$, we apply the fourth-order CFD scheme [29] as follows:

Table 5: Numerical convergence orders in temporal direction with $\alpha = 0.5$ and $M = 50$ for Example 3

N	The second-order		The third-order	
	$e(N, M)$	Rate	$e(N, M)$	Rate
5	1.3158×10^{-3}	—	5.7401×10^{-3}	—
10	3.4968×10^{-3}	1.9118	8.8998×10^{-4}	2.6892
20	9.0919×10^{-4}	1.9434	1.2614×10^{-4}	2.8187
40	2.3234×10^{-4}	1.9683	1.7195×10^{-5}	2.8750

$$\frac{\partial u(x, t)}{\partial x} \Big|_{i, n} = \frac{\delta_{\hat{x}}}{1 + \frac{h^2}{6} \delta_x^2} u_i^n.$$

Difference scheme using third-order formula can be obtained as

$$\mathcal{A}_0^{\text{CF}} \mathbb{D}_t^\alpha u_i^n = \delta_{\hat{x}} u_i^n + \mathcal{A} f_i^n, \quad 1 \leq i \leq M-1, \quad 1 \leq n \leq N, \quad (21)$$

$$u_0^n = \Phi(t_n), \quad u_1^n = \varphi(t_n), \quad 1 \leq n \leq N, \quad (22)$$

$$u_i^0 = u^0(x_i), \quad 0 \leq i \leq M, \quad (23)$$

where

$$\delta_{\hat{x}} v_i = \frac{1}{2h} (v_{i+1} - v_{i-1}),$$

and

$$\mathcal{A} v_i = \begin{cases} \frac{1}{6} (v_{i+1} + 4v_i + v_{i-1}), & 1 \leq i \leq M-1, \\ v_i, & i = 0 \text{ or } M. \end{cases}$$

It can be seen that

$$\mathcal{A} v_i = \left(I + \frac{h^2}{6} \delta_x^2 \right) v_i, \quad 1 \leq i \leq M-1.$$

Like Remark 2 but with slight change, we consider the fourth-order approximation of the first derivative of u at point x_i as follows:

$$\frac{\partial u(x, t)}{\partial x} \Big|_{x=x_i} = \frac{\delta_{\hat{x}}}{1 + \frac{h^2}{6} \delta_x^2} u(x_i, t) + O(h^4). \quad (24)$$

The exact solution of (20) is $u(x, t) = \sin(\pi x) t^5$. Functions can be obtained by substituting $u(x, t)$ into (20).

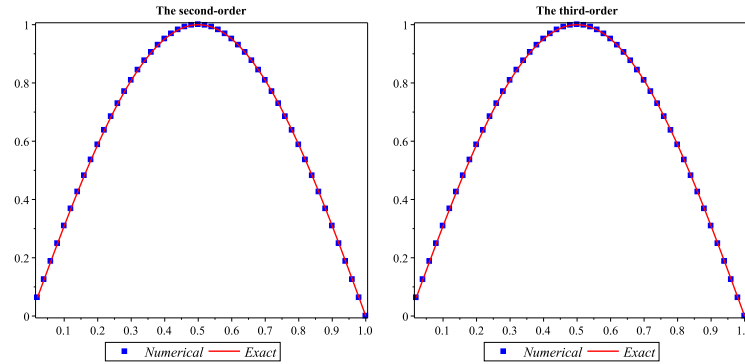


Figure 4: The solution curves at $T = 1$ with $M = 50, N = 40$ for Example 3

Table 5 confirms that the numerical convergence orders in both the second-order and the third-order schemes are close to theoretical results. Having seen Table 5, we conclude that the third-order produces better results for $e(N, M)$ than that the second-order both in error and accuracy. In [29], the accuracy of the presented method is dependent on α . In this case, the computational orders for $\alpha = 0.5$ is 1.5 with the theoretical order of $3 - \alpha$, whereas the accuracy of our schemes is not dependent on the fractional α . Figure 4 illustrates the plot of numerical solution and exact solution with $\alpha = 0.5, M = 50$ and $N = 40$ at final time $T = 1$ for Example 3.

4 Conclusions

In the current paper, we have obtained two new fractional numerical differentiation formulas to approximate the time Caputo–Fabrizio fractional derivative of order α ($0 < \alpha < 1$) on non-uniform meshes. First, the linear and quadratic interpolation approximations are considered for the integrand $u(t)$ because of obtaining the new formulas. Then, a fourth-order CFD scheme is employed for spatial discretization. This difference scheme is led to the third-order (second-order) and the fourth-order accuracy in the temporal and the spatial variables, respectively. Numerical results are carried out to support the convergence orders and show the efficiency of the suggested scheme. What distinguishes this paper from our previous studies is its accuracy aspect because the accuracy of the suggested schemes is not dependent on the fractional α .

Acknowledgements

The authors are very grateful to the reviewers for carefully reading the paper and for their comments and suggestions which have led to improvements of the paper. The computation was carried out at the Computing Center of Faculty of Mathematics, K. N. Toosi University of Technology.

References

1. Akman, T., Yildiz, B. and Baleanu, D. *New discretization of Caputo–Fabrizio derivative*, Comp. Appl. Math. 37 (2018) 3307–3333.
2. Atangana, A. *On the new fractional derivative and application to nonlinear fishers reaction-diffusion equation*, Appl. Math. Comput. 273 (2016) 948–956.
3. Atangana, A. and Alkahtani, B.S.T. *Extension of the resistance, inductance, capacitance electrical circuit to fractional derivative without singular kernel*, Adv. Mech. Eng. 7 (2015) 2015.
4. Atangana, A. and Alqahtani, R.T. *Numerical approximation of the space-time Caputo–Fabrizio fractional derivative and application to groundwater pollution equation*, Adv. Difference Equ. (2016), Paper No. 156, 13 pp.
5. Atangana, A. and Nieto, J.J. *Numerical solution for the model of RLC circuit via the fractional derivative without singular kernel*, Adv. Mech. Eng. 7 (2015) 1–7.
6. Aydogan, S.M., Baleanu, D., Mousalou, A. and Rezapour, S. *On approximate solutions for two higher-order Caputo–Fabrizio fractional integro-differential equations*, Adv. Differ. Equ. 221 (2017) 1–11.
7. Caputo, M. and Fabrizio, M. *A new definition of fractional derivative without singular kernel*, Progr. Fract. Differ. Appl. 1 (2015) 73–85.
8. Caputo, M. and Fabrizio, M. *Applications of new time and spatial fractional derivatives with exponential kernels*, Progr. Fract. Differ. Appl. 2 (2016) 1–11.
9. Djida, J.D. and Atangana, A. *More generalized groundwater model with space-time Caputo–Fabrizio fractional differentiation*, Numer. Methods Partial Differ. Equ. 33 (2017) 1616–1627.
10. El-Ghenbazia, F. and Tayeb Meftah, M. *Solution of nonlocal schrodinger equation via the Caputo–Fabrizio definition for some quantum systems*, Reports. Math. Physics, 85 (2020) 57–67.

11. Fakhr Kazemi, B. and Jafari, H. *Error estimate of the MQ-RBF collocation method for fractional differential equations with Caputo–Fabrizio derivative*, Math. Sci. 11 (2017) 297–305.
12. Firoozjaee, M.A., Jafari, H., Lia, A. and Baleanu, D. *Numerical approach of Fokker-Planck equation with Caputo–Fabrizio fractional derivative using Ritz approximation*, J. Compu. Appl. Math. 339 (2018) 367–373.
13. Gao, F., Li, X., Li, W. and Zhou, X. *Stability analysis of a fractional-order novel hepatitis B virus model with immune delay based on Caputo–Fabrizio derivative*, Chaos, Solitons, Fractals, 142 (2021) 110436.
14. Goufo, E.F. D. *Application of the Caputo–Fabrizio fractional derivative without singular kernel to Korteweg-de Vries-Burgers equation*, Math. Modell. Anal. 211 (2016) 88–98.
15. Guo, B., Pu, X. and Huang, F. *Fractional Partial Differential Equations and Their Numerical Solutions*, Science Press, Beijing, China, 2011.
16. Hilfer, R. *Applications of Fractional Calculus in Physics*, World Scientific, Singapore, 2000.
17. Leilei, W. and Li, W. *Local discontinuous Galerkin approximations to variable-order time-fractional diffusion model based on the Caputo-Fabrizio fractional derivative*, Math. Comput. Simulation, 188 (2021) 280–290.
18. Liu, Z., Cheng, A. and Li, X. *A second order Crank-Nicolson scheme for fractional Cattaneo equation based on new fractional derivative*, Appl. Math. Comput. 311 (2017) 361–374.
19. Loh, J.R. and Toh, Y.T. *On the new properties of Caputo–Fabrizio operator and its application in deriving shifted Legendre operational matrix*, Appl. Numer. Math. 132 (2018) 138–153.
20. Losada, J. and Nieto, J.J. *Properties of a new fractional derivative without singular kernel*, Progr. Fract. Differ. Appl. 1 (2015) 87–92.
21. Mirza, I.A. and Vieru, D. *Fundamental solutions to advection-diffusion equation with time-fractional Caputo–Fabrizio derivative*, Comput. Math. Appl. 73 (2017) 1–10.
22. Owolabi, K.M. and Atangana, A. *Analysis and application of new fractional Adams-Bashforth scheme with Caputo–Fabrizio derivative*, Chaos, Solitons, Fractals, 105 (2017) 111–119.
23. Owolabi, K.M. and Atangana, A. *Numerical approximation of nonlinear fractional parabolic differential equations with Caputo–Fabrizio derivative in Riemann-Liouville sense*, Chaos, Solitons, Fractals, 99 (2017) 171–179.

24. Podlubny, I. *Fractional Differential Equations*, Academic Press, New York, 1999.
25. Rubbab, Q., Nazeer, M., Ahmad, F., Chu, Y.M., Khan, M.I. and Kadry, S. *Numerical simulation of advection-diffusion equation with Caputo–Fabrizio time fractional derivative in cylindrical domains: Applications of pseudo-spectral collocation method*, Alex. Eng. J. 60 (2021) 1731–1738.
26. She, J. and Chen, M. *A second-order accurate scheme for two-dimensional space fractional diffusion equations with time Caputo–Fabrizio fractional derivative*, Appl. Numer. Math. 151 (2020) 246–262.
27. Soori, Z. and Aminataei, A. *Sixth-order non-uniform combined compact difference scheme for multi-term time fractional diffusion-wave equation*, Appl. Numer. Math. 131 (2018) 72–94.
28. Soori, Z. and Aminataei, A. *Effect of the nodes near boundary points on the stability analysis of sixth-order compact finite difference ADI scheme for the two-dimensional time fractional diffusion-wave equation*, Trans. A. Razmadze Math. Inst. 172 (2018) 582–605.
29. Soori, Z. and Aminataei, A. *A new approximation to Caputo-type fractional diffusion and advection equations on non-uniform meshes*, Appl. Numer. Math. 144 (2019) 21–41.
30. Soori, Z. and Aminataei, A. *Numerical solution of space fractional diffusion equation by spline method combined with Richardson extrapolation*, Comput. Appl. Math. 139 (2020) 1–18.
31. Sun, Z.Z. *The Method of Order Reduction and Its Application to the Numerical Solutions of Partial Differential Equations*, Science Press, Beijing, 2009.



Application of Newton–Cotes quadrature rule for nonlinear Hammerstein integral equations

A. Shahsavaran

Abstract

A numerical method for solving Fredholm and Volterra integral equations of the second kind is presented. The method is based on the use of the Newton–Cotes quadrature rule and Lagrange interpolation polynomials. By the proposed method, the main problem is reduced to solve some nonlinear algebraic equations that can be solved by Newton's method. Also, we prove some statements about the convergence of the method. It is shown that the approximated solution is uniformly convergent to the exact solution. In addition, to demonstrate the efficiency and applicability of the proposed method, several numerical examples are included, which confirms the convergence results.

AMS subject classifications (2020): 45B05; 45D05; 65R20; 45G10.

Keywords: Fredholm integral equation; Volterra integral equation; Newton–Cotes quadrature rule; Lagrange interpolation; Convergence.

1 Introduction

Integral equations have lots of applications in science and engineering. Fredholm integral equations arising in the theory of signal processing, which is a subfield of mathematics, information, and electrical engineering. In physics, the solution of such integral equations allows for experimental spectra to be related to various underlying distributions, for example, the mass distribution of polymers in a polymeric melt. They also prevalently appear in linear forward modeling and inverse problems. Also, Volterra integral equations arise in many scientific applications such as the population dynamics, the spread of epidemics, and semi-conductor devices. It was also shown that

Received 13 April 2021; revised 24 June 2021; accepted 27 June 2021

Ahmad Shahsavaran

Department of Mathematics, Borujerd Branch, Islamic Azad University, Borujerd, Iran.
e-mail: a.shahsavaran@iaub.ac.ir

Volterra integral equations can be derived from initial value problems; see [21]. Therefore, providing effective methods for solving such equations is consequential. Fredholm and Volterra integral equations have been extensively studied in many papers, and numerical methods have been widely used to solve such equations, for instance, Newton–Kantorovich and Haar wavelets for nonlinear Fredholm and Volterra integral equations [1], hat basis functions for the system of linear and nonlinear integral equations [2], Haar wavelets for nonlinear Fredholm integral equations [3], alternative Legendre polynomials for nonlinear Volterra–Hammerstein integral equations [5], operational matrix approach for nonlinear Volterra–Fredholm integral equations [4], cubic spline method for Fredholm integral equations [6], implicitly linear collocation method for nonlinear Volterra equations [7], Chebyshev approximation for nonlinear Fredholm–Volterra integral equations [9], Chebyshev collocation method for the class of Fredholm integral equations with highly oscillatory kernels [10], collocation-type method for Hammerstein integral equations [12], Sinc-collocation method for nonlinear Fredholm integral equations with weakly singular kernel [15], triangular functions for nonlinear Volterra–Fredholm integral equations [13], wavelets–Galerkin method and wavelets precondition for first kind Fredholm integral equations [14], spectral collocation method for Fredholm integral equations on the half-line [16], hybrid Legendre Block-Pulse functions for the system of nonlinear Fredholm–Hammerstein integral equations [17], single-term Walsh series for nonlinear Volterra–Hammerstein integral equations [18], piecewise constant functions method for nonlinear Fredholm–Volterra integral equations [19], and so on. The existence and uniqueness of such equations were discussed in [8, 11].

In this work, we consider the Fredholm and Volterra integral equations of Hammerstein type

$$u(x) = \nu(x) + \int_a^b \kappa(x, t) \psi(t, u(t)) dt, \quad x \in [a, b],$$

$$u(x) = \nu(x) + \int_a^x \kappa(x, t) \psi(t, u(t)) dt, \quad x \in [a, b],$$

where $\nu \in L^2[a, b]$ and $\kappa \in L^2[a, b]^2$ are known functions, ψ is a given nonlinear function defined on $[a, b]$, and u is unknown to be determined.

2 Method of solution

In this section, first, we describe the Newton–Cotes quadrature rule.

Let the interval $[a, b]$ be partitioned into n sub-intervals, by n equally spaced points. That is,

$$x_0 = a, \quad x_i = x_0 + ih, \quad \text{for } i = 0, 1, \dots, n, \quad (1)$$

where the step size h is defined by $h = \frac{x_n - x_0}{n} = \frac{b-a}{n}$.

The Newton–Cotes quadrature rule for a function f defined on $[a, b]$ with known values at equally spaced points x_i , $i = 0, 1, \dots, n$, is as follows:

$$\int_a^b f(x)dx \approx \sum_{i=0}^n \omega_i f(x_i), \quad (2)$$

for the set of weights $\{\omega_i\}_{i=0}^n$. Now consider the Lagrange basis polynomials $l_i(t) = \prod_{j=0, j \neq i}^n \left(\frac{t-t_j}{t_i-t_j}\right)$ and let $\rho_n(t)$ be the interpolation polynomial in the Lagrange form for the given data points $(t_0, f(t_0)), (t_1, f(t_1)), \dots, (t_n, f(t_n))$. Then

$$\begin{aligned} \int_a^b f(t)dt &\approx \int_a^b \rho_n(t)dt \\ &= \int_a^b \left(\sum_{i=0}^n f(t_i) l_i(t) \right) dt \\ &= \sum_{i=0}^n f(t_i) \int_a^b l_i(t)dt \\ &= \sum_{i=0}^n \omega_i f(t_i), \end{aligned} \quad (3)$$

where $\omega_i = \int_a^b l_i(t)dt$.

Note that we take the Lagrange interpolation points t_i to be the same as the points x_i for the Newton–Cotes quadrature rule. Now consider the following Fredholm and Volterra integral equations of the second kind

$$u(x) = \nu(x) + \int_a^b \kappa(x, t) \psi(t, u(t)) dt, \quad x \in [a, b], \quad (4)$$

$$u(x) = \nu(x) + \int_a^x \kappa(x, t) \psi(t, u(t)) dt, \quad x \in [a, b]. \quad (5)$$

These can be written as

$$u(x) = \nu(x) + \int_a^b \kappa(x, t) \Psi(t) dt, \quad (6)$$

$$u(x) = \nu(x) + \int_a^x \kappa(x, t) \Psi(t) dt, \quad (7)$$

where $\Psi(t) = \psi(t, u(t))$. By considering $u_n(x)$ as an approximation for $u(x)$, we can turn equations (6) and (7) into the following equations:

$$u_n(x) = \nu(x) + \int_a^b \kappa(x, t) \Psi_n(t) dt, \quad (8)$$

$$u_n(x) = \nu(x) + \int_a^x \kappa(x, t) \Psi_n(t) dt, \quad (9)$$

where $\Psi_n(t) = \psi(t, u_n(t))$, which immediately implies

$$\Psi_n(t) = \psi \left(t, \nu(t) + \int_a^b \kappa(t, x) \Psi_n(x) dx \right), \quad (10)$$

$$\Psi_n(t) = \psi \left(t, \nu(t) + \int_a^t \kappa(t, x) \Psi_n(x) dx \right). \quad (11)$$

Using quadrature formula (2) to evaluate the integral part of (10), we obtain

$$\begin{aligned} \int_a^b \kappa(t, x) \Psi_n(x) dx &= \sum_{i=0}^n \omega_i \kappa(t, x_i) \Psi_n(x_i) \\ &= \sum_{i=0}^n \omega_i \kappa(t, x_i) \Psi_{n,i}, \end{aligned} \quad (12)$$

where $\Psi_{n,i} = \Psi_n(x_i)$ and $\omega_i = \int_a^b l_i(t) dt$. Similarly, using the Lagrange interpolation for the integrand of (11) gives

$$\begin{aligned} \int_a^t \kappa(t, x) \Psi_n(x) dx &= \int_a^t \sum_{i=0}^n \kappa(t, x_i) \Psi_n(x_i) l_i(x) dx \\ &= \sum_{i=0}^n \kappa(t, x_i) \Psi_n(x_i) \int_a^t l_i(x) dx \\ &= \sum_{i=0}^n \omega_i(t) \kappa(t, x_i) \Psi_n(x_i) \\ &= \sum_{i=0}^n \omega_i(t) \kappa(t, x_i) \Psi_{n,i}, \end{aligned} \quad (13)$$

where $\omega_i(t) = \int_a^t l_i(x) dx$ and $\Psi_{n,i} = \Psi_n(x_i)$. Substituting (12) into (10) and (13) into (11), we obtain

$$\Psi_n(t) = \psi \left(t, \nu(t) + \sum_{i=0}^n \omega_i \kappa(t, x_i) \Psi_{n,i} \right) \quad (14)$$

and

$$\Psi_n(t) = \psi \left(t, \nu(t) + \sum_{i=0}^n \omega_i(t) \kappa(t, x_i) \Psi_{n,i} \right), \quad (15)$$

respectively. Evaluating (14) and (15) at the points $t = x_j$, $j = 0, 1, \dots, n$ (the points for Newton–Cotes quadrature rule), respectively, gives

$$\Psi_{n,j} = \psi \left(x_j, \nu(x_j) + \sum_{i=0}^n \omega_i \kappa(x_j, x_i) \Psi_{n,i} \right), \quad j = 0, 1, \dots, n, \quad (16)$$

and

$$\Psi_{n,j} = \psi \left(x_j, \nu(x_j) + \sum_{i=0}^n \omega_{i,j} \kappa(x_j, x_i) \Psi_{n,i} \right), \quad j = 0, 1, \dots, n, \quad (17)$$

where $\omega_{i,j} = \omega_i(x_j) = \int_a^{x_j} l_i(x) dx$. Nonlinear systems of algebraic equations (16) and (17) can be solved by numerical methods such as Newton's method. By solving the above systems, the values $\Psi_{n,i}$, $i = 0, 1, \dots, n$, will be known. Finally, by substituting (12) into (8) and (13) into (9), we find the numerical solutions of the integral equations (4) and (5) by

$$u_n(x) = \nu(x) + \sum_{i=0}^n \omega_i \kappa(x, t_i) \Psi_{n,i} \quad (18)$$

and

$$u_n(x) = \nu(x) + \sum_{i=0}^n \omega_i(x) \kappa(x, t_i) \Psi_{n,i}, \quad (19)$$

respectively, where $t_i = x_i$, $i = 0, 1, \dots, n$.

3 Convergence of the method

In this section, we analyze the convergence of the prescribed method in Section 2, which enables us to control the estimated errors. First, we provide an interpolation polynomial error bound, which is given in the following theorem.

Theorem 1. [20] Suppose that $f \in C^{n+1}[a, b]$, and let p_n be a polynomial of degree $\leq n$ that interpolates the function f at $n + 1$ distinct points $x_0, x_1, \dots, x_n \in [a, b]$. Then for each $x \in [a, b]$, there exists a point $\zeta_x \in [a, b]$ such that

$$f(x) - p_n(x) = \frac{\pi(x)}{(n+1)!} f^{(n+1)}(\zeta_x), \quad (20)$$

where $\pi(x) = \prod_{i=0}^n (x - x_i)$.

Theorem 2. In the case of equally spaced interpolation points $x_0 = a$ and $x_i = x_0 + ih$, for $i = 0, 1, \dots, n$, where $h = \frac{b-a}{n}$, we have

$$|\pi(x)| \leq \frac{n!}{4} h^{n+1}. \quad (21)$$

Proof. Suppose $x \in [x_i, x_{i+1}]$. Then for the first i terms of $\pi(x)$, that is $\Pi_{j=0}^{i-1}(x - x_j) = (x - x_0)(x - x_1) \dots (x - x_{i-1})$, and due to equally spaced points x_i , we have

$$\begin{aligned} |x - x_{i-j}| &\leq x_{i+1} - x_{i-j} \\ &= (j+1)h, \quad j = 1, 2, \dots, i. \end{aligned} \quad (22)$$

Thus

$$|\Pi_{j=0}^{i-1}(x - x_j)| \leq (i+1)! h^i. \quad (23)$$

For the next two terms of $\pi(x)$, that is $(x - x_i)(x - x_{i+1})$, and using the simple identity $\alpha\beta \leq \left(\frac{\alpha+\beta}{2}\right)^2$, we can write

$$\begin{aligned} |(x - x_i)(x - x_{i+1})| &= (x - x_i)(x_{i+1} - x) \quad (x_i \leq x \leq x_{i+1}) \\ &\leq \left(\frac{x_{i+1} - x_i}{2}\right)^2 \\ &= \frac{h^2}{4}. \end{aligned} \quad (24)$$

For the $n - i - 1$ remaining terms of $\pi(x)$, that is $\Pi_{j=i+2}^n(x - x_j) = (x - x_{i+2})(x - x_{i+3}) \dots (x - x_n)$, we may proceed as follows:

$$\begin{aligned} |x - x_{i+j}| &= x_{i+j} - x \quad (x_i \leq x \leq x_{i+1}) \\ &\leq x_{i+j} - x_0 \\ &= (i+j)h, \quad j = 2, 3, \dots, n-i. \end{aligned} \quad (25)$$

Thus

$$|\Pi_{j=i+2}^n(x - x_j)| \leq \frac{n!}{(i+1)!} h^{n-i-1}. \quad (26)$$

Therefore combining (23), (24), and (26) leads to

$$\begin{aligned} |\pi(x)| &= |\Pi_{i=0}^n(x - x_i)| \\ &\leq \frac{n!}{4} h^{n+1}. \end{aligned} \quad (27)$$

□

Theorem 3. Suppose that $\kappa \in C^{n+1}[a, b]^2$ and that ψ in (4) is a function in $C^{n+1}[a, b]$ with $n \geq 0$. If $u(x)$, the exact solution, and $u_n(x)$, the approximate solution defined by (18), are both in $C^{n+1}[a, b]$, then $u_n(x)$ is uniformly convergent to $u(x)$.

Proof. From (6) and (18), for every $x \in [a, b]$, we have

$$\begin{aligned}
u(x) - u_n(x) &= \int_a^b \kappa(x, t) \Psi(t) dt - \sum_{i=0}^n \omega_i \kappa(x, t_i) \Psi_{n,i} \quad (\omega_i = \int_a^b l_i(t) dt) \\
&= \int_a^b \left(\kappa(x, t) \Psi(t) - \sum_{i=0}^n \kappa(x, t_i) \Psi_{n,i} l_i(t) \right) dt, \quad (28)
\end{aligned}$$

but $\sum_{i=0}^n \kappa(x, t_i) \Psi_{n,i} l_i(t)$ interpolates $\kappa(x, t) \Psi(t)$ at the points t_i , $i = 0, 1, \dots, n$, ($t_i = x_i$, the points for Newton–Cotes quadrature rule). If we set $F(x, t) = \kappa(x, t) \Psi(t)$, from (28) and Theorem 1, then

$$u(x) - u_n(x) = \int_a^b \left(\frac{\pi(t)}{(n+1)!} \frac{\partial^{n+1} F}{\partial t^{n+1}}(x, \zeta_t) \right) dt, \quad \zeta_t \in [a, b], \quad (29)$$

where $\pi(t) = \prod_{i=0}^n (t - t_i)$. Since $\psi, u \in C^{n+1}[a, b]$, there is some $M_1 > 0$ with $|\Psi^{(n+1)}(t)| \leq M_1$ for all $t \in [a, b]$, where $\Psi(t) = \psi(t, u(t))$. Also $\kappa \in C^{n+1}[a, b]^2$; thus there is some $M_2 > 0$ with $|\frac{\partial^{n+1} \kappa}{\partial t^{n+1}}(x, t)| \leq M_2$ for all $x, t \in [a, b]$. Then $F(x, t) = \kappa(x, t) \Psi(t)$, necessitates that $|\frac{\partial^{n+1} F}{\partial t^{n+1}}(x, t)| \leq M$ for a real number M and for all $x, t \in [a, b]$. Therefore from (27) and (29), we obtain

$$\begin{aligned}
|u(x) - u_n(x)| &\leq \frac{1}{(n+1)!} \int_a^b \left(|\pi(t)| \left| \frac{\partial^{n+1} F}{\partial t^{n+1}}(x, \zeta_t) \right| \right) dt \\
&\leq \frac{(b-a)M}{4(n+1)} h^{n+1} \quad (h = \frac{b-a}{n}) \\
&\leq \frac{M}{4} \left(\frac{b-a}{n} \right)^{n+2}, \quad (30)
\end{aligned}$$

where we used $(n+1)n^{n+1} \geq nn^{n+1} = n^{n+2}$ to get the last inequality. We note that

$$\left(\frac{b-a}{n} \right)^{n+2} \leq \left(\frac{b-a}{n} \right)^2 \left(1 + \frac{b-a}{n} \right)^n.$$

Thus from (30), we obtain

$$0 \leq |u(x) - u_n(x)| \leq \frac{M}{4} \left(\frac{b-a}{n} \right)^2 \left(1 + \frac{b-a}{n} \right)^n. \quad (31)$$

It is clear to see that

$$\lim_{n \rightarrow \infty} \left(\frac{b-a}{n} \right)^2 = 0, \quad (32)$$

$$\lim_{n \rightarrow \infty} \left(1 + \frac{b-a}{n} \right)^n = e^{b-a}. \quad (33)$$

Considering the limit of the both sides of (31) as n approaches infinity and using (32)–(33), yield

$$\lim_{n \rightarrow \infty} u_n(x) = u(x), \quad \text{for every } x \in [a, b];$$

that is, $u_n(x)$ is uniformly convergent to $u(x)$. \square

As a result of Theorem 3 and from (30), it is also clearly seen that

$$|u(x) - u_n(x)| = O(h^{n+2}), \text{ where } h = \frac{b-a}{n}. \quad (34)$$

4 Illustrative examples

In this section, we apply the method proposed in Section 2 to some test examples. All numerical calculations are performed by Maple 13.

4.1 Fredholm integral equation

Example 1. Consider

$$u(x) = \frac{7}{8}x + \int_0^1 \frac{1}{2}xtu^2(t)dt, \quad 0 \leq x \leq 1,$$

with exact solution $u(x) = x$.

For this example, by solving the nonlinear system (16) for $n = 4$, we have

$$\Psi_{4,0} = 0, \Psi_{4,1} = 0.0625, \Psi_{4,2} = 0.25, \Psi_{4,3} = 0.5625, \Psi_{4,4} = 1.$$

Also the nodes x_i and the weights ω_i of the Newton–Cotes quadrature rule for the same n are

$$t_0 = 0, t_1 = \frac{1}{4}, t_2 = \frac{1}{2}, t_3 = \frac{3}{4}, t_4 = 1,$$

$$\omega_0 = \frac{7}{90}, \omega_1 = \frac{16}{45}, \omega_2 = \frac{2}{15}, \omega_3 = \frac{16}{45}, \omega_4 = \frac{7}{90}.$$

Substituting the values of x_i , ω_i and $\Psi_{4,i}$ for $i = 0, \dots, 4$ into (18), we have

$$\begin{aligned}
u_4(x) &= \nu(x) + \sum_{i=0}^4 \omega_i \kappa(x, t_i) \Psi_{4,i} \\
&= \frac{7}{8}x + \frac{7}{90} \left(\frac{1}{2}x \times 0 \right) \times 0 \\
&\quad + \frac{16}{45} \left(\frac{1}{2}x \times \frac{1}{4} \right) \times 0.0625 \\
&\quad + \frac{2}{15} \left(\frac{1}{2}x \times \frac{1}{2} \right) \times 0.25 \\
&\quad + \frac{16}{45} \left(\frac{1}{2}x \times \frac{3}{4} \right) \times 0.5625 \\
&\quad + \frac{7}{90} \left(\frac{1}{2}x \times 1 \right) \times 1 \\
&= x,
\end{aligned}$$

which is the exact solution of Example 1

Example 2. [9] Consider

$$u(x) = ex + 1 - \int_0^1 (x+t)e^{u(t)} dt, \quad 0 \leq x \leq 1,$$

with the exact solution $u(x) = x$.

The absolute error $|u_n(x) - u(x)|$ of Example 2 for $n = 2, 4, 6$ is shown in Table 1.

Table 1: Absolute error of Example 2

x	$n = 2$	$n = 4$	$n = 6$	method of [9] ($N = 7$)
0.0	$1.7e-3$	$3.8e-6$	$6.0e-9$	$2.5e-6$
0.2	$1.5e-3$	$3.2e-6$	$5.2e-9$	$7.3e-6$
0.4	$1.2e-3$	$2.6e-6$	$4.4e-9$	$7.9e-6$
0.6	$1.0e-3$	$2.1e-6$	$3.6e-9$	$2.5e-6$
0.8	$7.7e-4$	$1.5e-6$	$2.8e-9$	$3.9e-6$
1.0	$5.3e-4$	$9.6e-7$	$2.0e-9$	$2.6e-6$

Example 3. [1] Consider

$$u(x) = e^{x+1} - \int_0^1 e^{x-2t}(u(t))^3 dt, \quad 0 \leq x \leq 1,$$

with exact solution $u(x) = e^x$.

The absolute error $|u_n(x) - u(x)|$ of Example 3 for $n = 2, 4, 6$ is shown in Table 2.

Table 2: Absolute error of Example 3

x	$n = 2$	$n = 4$	$n = 6$	method of [1] ($N = 16$)
0.1	$1.0e-4$	$1.5e-7$	0.0	$4.3e-4$
0.2	$1.1e-4$	$1.7e-7$	$1.0e-9$	$3.3e-7$
0.3	$1.2e-4$	$1.8e-7$	0.0	$7.9e-4$
0.4	$1.4e-4$	$2.0e-7$	0.0	$2.9e-4$
0.5	$1.5e-4$	$2.3e-7$	$1.0e-9$	$1.2e-3$
0.6	$1.7e-4$	$2.5e-7$	0.0	$7.1e-4$
0.7	$1.8e-4$	$2.8e-7$	$1.0e-9$	$5.4e-7$
0.8	$2.0e-4$	$3.1e-7$	$1.0e-9$	$1.3e-3$
0.9	$2.3e-4$	$3.4e-7$	$1.0e-9$	$4.8e-4$

Example 4. [12] In this example, the proposed method in Section 2 is used to solve an integral equation reformulation of the nonlinear two-point boundary value problem

$$u''(t) - \exp(u(t)) = 0, \quad t \in (0, 1), \quad u(0) = u(1) = 0.$$

This problem has the unique solution

$$u(t) = -\ln(2) + 2 \ln \left(\frac{c}{\cos \left(\frac{c(t-\frac{1}{2})}{2} \right)} \right),$$

where c is the only solution of $\frac{c}{\cos(\frac{c}{4})} = \sqrt{2}$, and may be reformulated as the integral equation

$$u(x) = \int_0^1 \kappa(x, t) \exp(u(t)) dt, \quad x \in [0, 1],$$

where

$$\kappa(x, t) = \begin{cases} -t(1-x), & t \leq x, \\ -x(1-t), & t > x. \end{cases}$$

The uniform norm $\|u_n - u\| = \sup\{|u_n(t) - u(t)|, t \in [0, 1]\}$ of Example 4 for $n = 5, 9$ is shown in Table 3.

Table 3: Numerical results of Example 4

n	$\ u_n - u\ $ (presented method)	$\ u_n - u\ $ (method of [12])
5	$5.94e-3$	$5.19e-4$
9	$2.19e-3$	$1.28e-4$

4.2 Volterra integral equation

Example 5. [5] Consider

$$u(x) = \frac{3}{2} - \frac{1}{2}e^{-2x} - \int_0^x (u^2(t) + u(t))dt, \quad 0 \leq x \leq 1,$$

with the exact solution $u(x) = e^{-x}$.

For this example, we define $e_n(x) = |u_n(x) - u(x)|$. The maximum norm $\|e_n\|_\infty = \max\{|e_n(x_i)|, x_i = .1 * i, i = 0, 1, \dots, 10\}$ for $n = 1, 3, 5, 7, 9$ is presented in Table 4.

Table 4: Numerical results of Example 5

n	$\ e_n\ _\infty$ (presented method)	$\ e_n\ _\infty$ (method of [5])
1	$1.174e-1$	$6.282e-2$
3	$5.235e-4$	$1.001e-3$
5	$3.829e-6$	$8.294e-6$
7	$2.020e-8$	$3.913e-8$
9	$1.000e-10$	$1.163e-10$

Example 6. [18] Consider

$$u(x) = 1 + \sin^2 x - 3 \int_0^x \sin(x-t)(u(t))^2 dt, \quad 0 \leq x \leq 1,$$

with exact solution $u(x) = \cos x$.

The absolute error $|u_n(x) - u(x)|$ of Example 6 for $n = 3, 5, 7, 9$ is shown in Table 5.

Table 5: Absolute error of Example 6

x	$n = 3$	$n = 5$	$n = 7$	$n = 9$	method of [18] ($m = 60$)
0.2	$4.6e-3$	$8.4e-5$	$8.4e-7$	$6.3e-9$	0.0
0.4	$3.8e-3$	$2.2e-5$	$4.2e-7$	$3.9e-9$	$1.0e-5$
0.6	$2.1e-3$	$1.0e-6$	$1.1e-7$	$5.0e-10$	$1.0e-5$
0.8	$3.4e-3$	$6.3e-5$	$6.1e-7$	$3.9e-9$	$2.0e-5$
1.0	$1.1e-3$	$2.2e-5$	$2.9e-7$	$2.1e-9$	$1.0e-5$

Example 7 (Constructed by author). Consider

$$u(x) = e^{-x} - e^x(x+1) + \int_{-1}^x e^{x+t}u(t)dt, \quad -1 \leq x \leq 1. \quad (35)$$

To calculate the error in the interval $[-1, 1]$, we define the error function $e_n(x)$ as

$$e_n(x) = u_n(x) - \nu(x) - \int_{-1}^x \kappa(x, t) u_n(t) dt.$$

Actually, on the right-hand side of the above equation, we put the approximated solution $u_n(x)$ instead of the exact solution $u(x)$ for (4). Now the absolute error $|e_n(x)|$ of Example 7 for $n = 2, 3, 5, 7, 9$ and some $x \in [-1, 1]$ is shown in Table 6.

Table 6: Absolute error of Example 7

x	$e_2(x)$	$e_3(x)$	$e_5(x)$	$e_7(x)$	$e_9(x)$
-1.0	0.0	0.0	0.0	0.0	0.0
-0.8	$2.0e-3$	$3.4e-4$	$2.3e-4$	$5.7e-6$	$1.0e-9$
-0.6	$1.0e-2$	$1.2e-3$	$4.5e-4$	$6.9e-6$	$1.0e-9$
-0.4	$2.8e-2$	$2.2e-3$	$4.4e-4$	$3.6e-7$	$1.2e-9$
-0.2	$5.2e-2$	$2.5e-3$	$4.2e-4$	$1.2e-5$	$9.0e-10$
0.0	$7.4e-2$	$1.6e-3$	$6.4e-4$	$1.9e-5$	$3.0e-9$
0.2	$7.3e-2$	$5.9e-5$	$9.4e-4$	$3.2e-5$	$6.1e-9$
0.4	$2.0e-2$	$3.7e-4$	$9.0e-4$	$1.8e-5$	$9.2e-9$
0.6	$1.0e-1$	$1.5e-3$	$7.1e-4$	$6.9e-6$	$2.1e-8$
0.8	$2.9e-1$	$2.2e-4$	$1.4e-3$	$3.8e-5$	$6.9e-8$
1.0	$4.4e-1$	$1.3e-3$	$1.4e-3$	$5.1e-5$	$2.2e-7$

Example 8. Consider

$$u(x) = -\frac{x^5}{4} - \frac{2x^4}{3} - \frac{5x^3}{6} - x^2 + 1 + \int_0^x (xt + 1)(u(t))^2 dt, \quad 0 \leq x \leq 1.$$

Similar to Example 7, to calculate the error in the interval $[0, 1]$, we define the error function $e_n(x)$ as

$$e_n(x) = u_n(x) - \nu(x) - \int_0^x \kappa(x, t)(u_n(t))^2 dt.$$

Now the absolute error $|e_n(x)|$ of Example 8 for $n = 2, 3, 5, 7$ and some $x \in [0, 1]$ is shown in Table 7.

5 Conclusion

In this work, the Newton–Cotes quadrature rule together with the Lagrange interpolation were used to transform Fredholm and Volterra integral equations to a system of algebraic equations. As shown in Section 4, via some

Table 7: Absolute error of Example 8

x	$e_2(x)$	$e_3(x)$	$e_5(x)$	$e_7(x)$
0.0	0.0	0.0	0.0	0.0
0.1	$4.7e-4$	$9.5e-3$	0.0	0.0
0.2	$1.1e-3$	$2.7e-2$	$1.0e-10$	$1.0e-10$
0.3	$1.4e-3$	$4.0e-2$	$1.0e-10$	$3.0e-10$
0.4	$1.1e-3$	$4.1e-2$	$1.0e-10$	$2.0e-10$
0.5	$4.4e-4$	$3.1e-2$	$1.0e-10$	0.0
0.6	$9.4e-4$	$1.9e-2$	0.0	$2.0e-9$
0.7	$3.4e-3$	$1.7e-2$	0.0	$8.0e-9$
0.8	$7.4e-3$	$3.2e-2$	$1.0e-9$	$1.0e-8$
0.9	$1.2e-2$	$5.1e-2$	$1.0e-9$	$4.0e-8$
1.0	$1.6e-2$	$5.7e-2$	$3.0e-9$	$7.7e-8$

test examples, as n increases, the error decreases. The calculated errors in the test examples were compatible with the presented error bound in (30). It was stated that a high accuracy is achieved even by using a small number of n . Also the method can be extended to solve a system of such equations.

References

1. Abdul Sathar, M.H., Rasedee, A.F.N., Ahmedov, A.A. and Bachok, N. *Numerical solution of nonlinear Fredholm and Volterra integral equations by Newton–Kantorovich and Haar wavelets methods*, Symmetry, 12 (2020) 1–13.
2. Babolian, E. and Mordad, M. *A numerical method for solving system of linear and nonlinear integral equations of the second kind by hat basis functions*, Comput. Math. Appl., 62 (2011) 187–198.
3. Babolian, E. and Shamsavaran, A. *Numerical solution of nonlinear Fredholm integral equations of the second kind using Haar wavelet*, J. Comput. Appl. Math. 225 (2009) 87–95.
4. Basirat, B., Maleknejad, K. and Hashemizadeh, E. *Operational matrix approach for the nonlinear Volterra–Fredholm integral equations: Arising in physics and engineering*, Int. J. Phys. Sci. 7 (2012) 226–233.
5. Bazm, S. *Solution of nonlinear Volterra–Hammerstein integral equations using alternative Legendre collocation method*, Sahand Communications in Mathematical Analysis, 4 (2016) 57–77.

6. Bellour, A., Sbibi, D. and Zidna, A. *Two cubic spline methods for solving Fredholm integral equations*, Appl. Math. Comput. 276 (2016) 1–11.
7. Brunner, H. *Implicitly linear collocation method for nonlinear Volterra equations*, J. Appl. Num. Math. 9 (1982) 235–247.
8. Delves, L.M. and Mohamed, J.L. *Computational methods for integral equations*, Cambridge University Press, 1985.
9. Fattahzadeh, F. *Numerical solution of general nonlinear Fredholm-Volterra integral equations using Chebyshev approximation*, Int. J. Ind. Math. 8 (2016) 81–86.
10. He, G., Xiang, S. and Xu, Z. *A Chebyshev collocation method for a class of Fredholm integral equations with highly oscillatory kernels*, J. Comput. Appl. Math. 300 (2016) 354–368.
11. Ibrahim, I.A. *On the existence of solutions of functional integral equations of Urysohn type*, Comput. Math. with Appl. 57 (2009) 1609–1614.
12. Kumar, S. and Sloan, I.H. *A New collocation-type method for Hammerstein integral equations*, Math. Comput. 48 (1987) 585–593.
13. Maleknejad, K., Almasieh, H. and Roodaki, M. *Triangular functions (TF) method for the solution of nonlinear Volterra–Fredholm integral equations*, Commun. Nonlinear Sci. Numer. Simul. 15 (2010) 3293–3298.
14. Maleknejad, K., Lotfi, T. and Mahdiani, K. *Numerical solution of first kind Fredholm integral equations with wavelets-Galerkin method and wavelets precondition*, Appl. Math. Comput. 186 (2007) 794–800.
15. Maleknejad, K., Mollapourasl, R. and Ostadi, A. *Convergence analysis of Sinc-collocation method for nonlinear Fredholm integral equations with a weakly singular kernel*, J. Comput. Appl. Math. 278 (2015) 1–11.
16. Rahmoune, A. *Spectral collocation method for solving Fredholm integral equations on the half-line*, Appl. Math. Comput. 219 (2013) 9254–9260.
17. Sahu, P.K. and Ray, S.S. *Hybrid Legendre Block-Pulse functions for the numerical solutions of system of nonlinear Fredholm–Hammerstein integral equations*, Appl. Math. Comput. 270 (2015) 871–878.
18. Sepehrian, B. and Razzaghi, M. *A new method for solving nonlinear Volterra–Hammerstein integral equations via single-term Walsh series*, Mathematical Analysis and Convex Optimization, 1 (2020) 59–69.
19. Shahsavaran, A. *Numerical solution of nonlinear Fredholm-Volterra integral equations via piecewise constant functions by collocation method*, Am. J. Comput. Math. 1 (2011) 134–138.

20. Stoer, J. and Bulirsch, R. *Introduction to numerical analysis*, Springer-Verlag, 1991.
21. Wazwaz, A.M. *Linear and nonlinear integral equations: Methods and applications*, Higher education, Springer, 2011.



Investigating a claim about resource complexity measure

H.R. Yousefzadeh

Abstract

The utilization factor (UF) measures the ratio of the total resources' amount required to the availability of resources' amount during the life cycle of a project. In 1982, in the journal of Management Science, Kurtulus and Davis claimed that "If two resource-constrained problems for each type of resource have the same UF's value in each period of time, then each problem is subjected to the same amount of delay provided that the same sequencing rule is used (If different tie-breaking rules are used, a different schedule may be obtained)". In this paper, with a counterexample, we show that the claim of authors cannot be justified.

AMS subject classifications (2020): 90B35, 68M20.

Keywords: Scheduling scheme; Priority rule; Multi-project environment; Resource measure.

1 Brief description

For the resource-constrained project scheduling problem (RCPSP) in single-project environments, several resource measures and distributions have been proposed in the literature. Notable among them are resource factor, resource strength, resource density, and resource constrained-ness (see, e.g., [2, 3]). However, the related RCPSP's measures are not suitable for the multi-project environments [4]. Hence, for the multi-project environments, some resource measures such as average loading factor, average resource loading factor (ARLF), and average utilization factor (AUF) are proposed for resource-constrained multi-project scheduling problems (RCMPSPs).

The AUF is more widely used than the others, and first, we describe it as blow (for more details, see, e.g., [4, 5]):

Received 16 May 2021; revised 9 July 2021; accepted 9 July 2021

Hamid Reza Yousefzadeh

Department of Mathematics, Payame Noor University (PNU), P.O. Box, 19395-4697, Tehran, Iran.

E-mail: usefzadeh.math@pnu.ac.ir

The AUF measures the degree of dependency for each of the required resources. As the AUF considers the ratio of the amount of resource required to the level of available resource, it is complementary to the ARLF.

To explain more, suppose that CP_i is the length of the critical path of project i ($i = 1, \dots, M$) in the RCMPSP subject to

$$CP_1 \leq CP_2 \leq \dots \leq CP_M.$$

Define

$$S_1 = CP_1, S_2 = CP_2 - CP_1, \dots, S_M = CP_M - CP_{M-1}.$$

Then the total required resource of type k on L th interval, that is, $[a, b] = [CP_{L-1}, CP_L]$ is defined as

$$W_{S_L, k} = \sum_{t=a}^b \sum_{i=1}^M \sum_{j=1}^{N_i} r_{ijk} X_{ijt},$$

where

$$X_{ijt} = \begin{cases} 1 & \text{if activity } j \text{ in project } i \text{ at time } t \text{ is active,} \\ 0 & \text{otherwise,} \end{cases}$$

the amount of required resource of type k for the activity j of the project i is denoted by r_{ijk} , and N_i is the number of activities in the project i .

The AUF for the resource k (AUF_k) is defined as

$$AUF_k = \frac{1}{M} \sum_{L=1}^M \frac{W_{S_L, k}}{R_k \times |S_L|},$$

where R_k is the total amount of renewable resource k per unit of time ($k \in \mathcal{R} = \{1, \dots, K\}$). Hence, the value of AUF for the RCMPSP with K resources is determined by

$$AUF = \{\max\{AUF_1, \dots, AUF_K\}\}.$$

When $AUF_k > 1$, it can be concluded that the resource k is constrained in the RCMPSP [4].

For addressing the AUF measure, a related measure, that is, the utilization factor (UF) is needed to discuss: The UF measures the ratio of the total amount of resources required to the amount of available resources during the life cycle of a project [1]. Hence, for a particular type of resource in a project, if $UF \leq 1$, then it is clear that there is no resource constraint and that the early schedule is optimal [4].

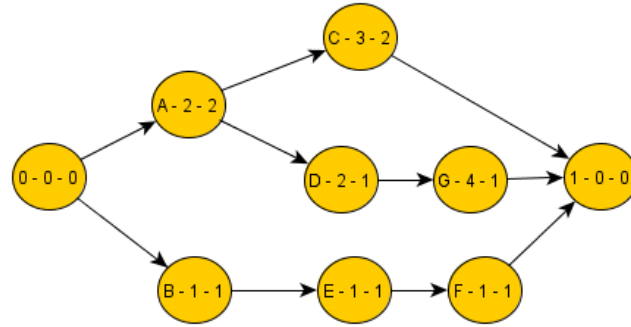
2 Main problem

In 1982, Kurtulus and Davis [4, p.163] claimed that “If two resource-constrained problems for each type of resource have the same UF’s value in each period of time, then each problem is subjected to the same amount of delay provided that the same sequencing rule is used (If different tie-breaking rules are used, a different schedule may be obtained)”.

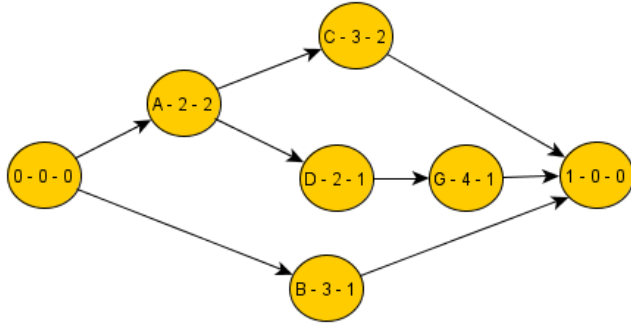
It is worth noting that the above statement is the main basis of many types of research in multi-project environments.

In the following example, we show that such a claim cannot be justified.

Example 1. Consider two projects P_1 and P_2 in Figure 1. They have the same UF distribution during their life cycles (see Figures 2 and 3). In the following, we show that even if the priority rule and the tie-breaking rule in the scheduling of these projects are the same, then the yielded schedules are not necessarily identical.



(a) Project P_1



(b) Project P_2

Figure 1: Two example project networks

Without loss of generality, we consider the case where $|\mathcal{R}| = 1$. Let $R_1 = 3$. Moreover, assume that the priority rule and the tie-breaking rule are the GRD¹ and the LPT², respectively. In other word, they are formulated as (see, e.g., [6])

$$\text{GRD: } \max_j = d_j \sum_{k \in \mathcal{R}} r_{jk}$$

and

$$\text{LPT: } \max_j = d_j,$$

where d_j indicates the processing time of activity j and r_{jk} is the amount of required resource k in the activity j .

The resource distribution for projects P_1 and P_2 is shown in Figures 2a and 2b, respectively.

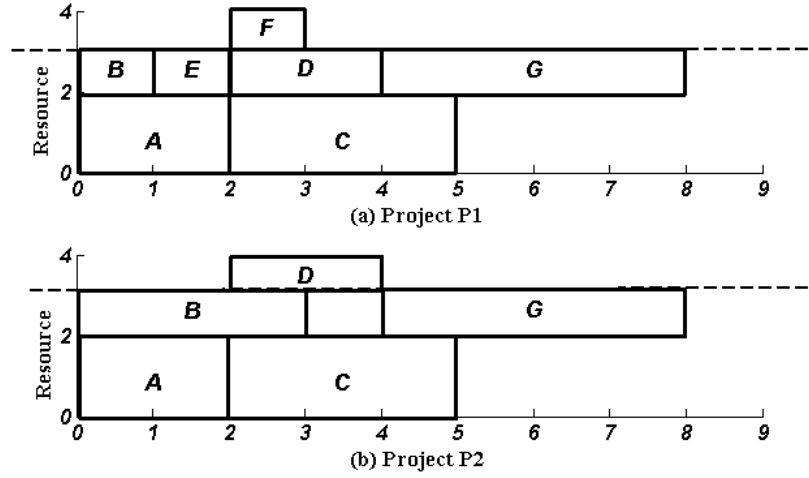
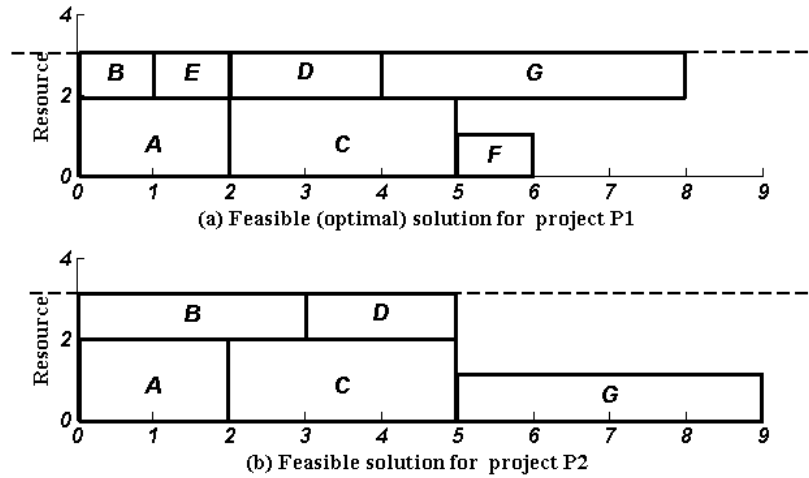


Figure 2: Resource distribution for projects P_1 and P_2

Since the amount of available resources and resource distribution for both P_1 and P_2 are the same, then the corresponding amount of the UFs is the same. Now, by applying the same priority rule (i.e., the GRD rule) and the same tie-breaking rule (i.e., the LPT rule), from Figure 3a, it is observed that the schedule for P_1 is optimal ($T=8$), while for P_2 , as Figure 3b shows, the schedule's makespan is not optimal and equals 9 (i.e., $T=9$). Hence, the authors claim in [4] cannot be verified.

¹ Greatest Resource Demand (GRD)

² Longest Processing Time (LPT)

Figure 3: Feasible solution for projects P_1 and P_2

References

1. Davis, E.W. *Project network summary measures constrained resource scheduling*, IIE Trans. 7(2), (1975), 132–142.
2. Demeulemeester E., Vanhoucke M. and Herroelen W. *RanGen: A random network generator for activity on the node networks*, J. Sched. 6(1) (2003), 17–38.
3. Kolisch, R., Sprecher, A. and Drexel, A. *Characterization and generation of a general class of resource constrained project scheduling problems*, Manage Sci. 41 (10) (1975), 1693–1703.
4. Kurtulus, I. and Davis E.W. *Multi Project Scheduling: Categorization of Heuristic Rules Performance*, Manage Sci. 28(2) (1982), 161–172.
5. Kurtulus I. and Narula S.C. *Multi Project Scheduling: Analysis of Project Management*, IIE Trans. 17(1) (1985), 58–65.
6. Yoosefzadeh, H.R., Tareghian, H.R., and Farahi, M.H. *Tri-directional Scheduling Scheme: Theory and Computation*, J. Math. Model. Algorithms. 9(4) (2010), 357–373.



A new algorithm for solving linear programming problems with bipolar fuzzy relation equation constraints

S. Aliannezhadi and A. Abbasi Molai*

Abstract

This paper studies the linear optimization problem subject to a system of bipolar fuzzy relation equations with the max-product composition operator. Its feasible domain is briefly characterized by its lower and upper bound, and its consistency is considered. Also, some sufficient conditions are proposed to reduce the size of the search domain of the optimal solution to the problem. Under these conditions, some equations can be deleted to compute the minimum objective value. Some sufficient conditions are then proposed which under them, one of the optimal solutions of the problem is explicitly determined and the uniqueness conditions of the optimal solution are expressed. Moreover, a modified branch-and-bound method based on a value matrix is proposed to solve the reduced problem. A new algorithm is finally designed to solve the problem based on the conditions and modified branch-and-bound method. The algorithm is compared to the methods in other papers to show its efficiency.

AMS subject classifications (2020): 90C70, 90C46, 90C26.

Keywords: Bipolar Fuzzy Relation Equation; Linear Optimization; Max-Product Composition; Modified Branch-and-Bound Method.

1 Introduction

Sanchez [31] has firstly studied Fuzzy Relation Equations (FREs) and their associated problems. Then, many researchers investigated them from a theoretical standpoint and in view of applications [26, 28, 32].

*Corresponding author

Received 4 November 2020; revised 6 July 2021; accepted 21 July 2021

Samaneh Aliannezhadi

School of Mathematics and Computer Sciences, Damghan University, P.O.Box 36715-364, Damghan, Iran. e-mail: s.aliannezhadi@std.du.ac.ir

Ali Abbasi Molai

School of Mathematics and Computer Sciences, Damghan University, P.O.Box 36715-364, Damghan, Iran. e-mail: a_abbasimolai@du.ac.ir

Various approaches were designed to solve a system of FREs such as the algebraic method [21], the matrix pattern method [25], the universal algorithm [27], and the improved Lichun and Boxing's method [39]. A comprehensive review of their resolution methods has been expressed in [7] and references therein. An extended kind of FREs is a system of fuzzy Relation Inequalities (FRIs), which its solution set can be completely determined by finding its minimal solutions and maximum solution similar to FREs [16]. Their applications can be seen in the supply chain [38] and Peer-to-Peer data transmission network system [22, 36, 4].

The above FREs and FRIs are increasing in each of the variables. In some applications, for example, in an application of product public awareness in revenue management, we need variables with a bipolar characterization [9]. Some researchers introduced such bipolarity effects with the max-min composition operator in the application [9]. Li and Jin [19] showed that checking the consistency of the system of bipolar max-min FREs is NP-complete. Therefore, the resolution of the linear optimization problem with constraints of bipolar FREs will be NP-hard. Yang [37] proposed a bipolar path approach to finding the complete solution set of the system. The characterization of the solvability of bipolar max-product FREs was investigated with the standard negation [5] and the product negation [6].

The optimization of objective functions with FRE constraints is an interesting research topic [8, 14, 17, 24, 29, 30, 33, 34, 35]. Fang and Li [8] firstly studied the linear programming problem provided to a system of the max-min FREs. They proposed an algorithm based on the branch-and-bound method with the jump-tracking technique for its resolution. The algorithm was extended to the problem with the max-product composition operator in [24]. Wu, Guu, and Liu, [35] improved their approach by designing an efficient procedure. In the procedure, the branch-and-bound method was applied based on an upper bound for its optimal objective value. Hence, the procedure checks much fewer nodes to find the optimal solution with respect to Fang and Li's approach. A necessary condition has been given to find an optimal solution to the problem with the max-min [34] and the max-product [14] composition operator. Then, the necessary condition was extended for fuzzy relation programming problem with the max-strict-t-norm composition [33]. Three rules were presented to simplify the process of finding the optimal solution based on the necessary condition [34]. Li and Fang [17] investigated the resolution and optimization of a system of FREs with Sup-T composition, and they generalized the most known results in literature and provided a unified framework for the resolution and optimization of the Sup-T equation. Recently, fuzzy relation programming was extended to optimize separable functions in [15]. Different kinds of fuzzy relation programming have appeared like the max-min fuzzy relation programming problem with the addition-min FRIs [4], linear optimization with the addition-min FRIs [12], and fuzzy relation lexicographic programming [40]. Recently, Zhou et al. [41] considered the problem of optimizing a nonlinear objective function

subject to a system of bipolar FREs with the max-Lukasiewicz triangular norm composition. They equivalently converted the problem to a 0-1 mixed nonlinear integer programming problem. It is NP-hard, and its resolution has high computational complexity. To increase the efficiency of algorithms, researchers focused on special classes of the bipolar fuzzy relation programming problems. Two important classes from the problem are linear [9, 20, 23, 3] and geometric [1, 2] with bipolar FRE constraints. For the first time, Freson, De Baets, and De Meyer [9] formulated the system of bipolar max-min FREs. They obtained the solution set of each of its equations. Using this point, they determined the solution set of a system of bipolar max-min FREs. This set can be characterized by a finite set of maximal and minimal solution pairs. They then studied the linear optimization problem subject to the system with a potential application of product public awareness in revenue management. They also found its optimal solutions based on the structure of the solution set of the system. Li and Liu [20] considered the problem with the max-Lukasiewicz t -norm. They converted the problem into a 0-1 integer linear optimization problem and solved it by using integer optimization techniques. However, the techniques may involve a high computational complexity. To overcome the point, Liu, Lur, and Wu [23] used the useful property that each component of an optimal solution can either be the corresponding component of lower or upper bound value and they proposed a simple value matrix with some simplification rules to reduce the dimensions of the original problem. To improve computational efficiency, some other rules were presented for more reduction of the dimensions of the problem with the max-parametric Hamacher composition operator in [3]. With regard to the above points, a modified branch-and-bound method was designed for the resolution of the problem in [3].

In this paper, the problem with the max-product operator is investigated. Its simplification procedures are completely different from the rules in [23, 3]. The motivations of this paper are the answers to the following questions:

- 1- How do we can detect and remove the redundancy constraints in the bipolar FRE system?
- 2- What are the sufficient conditions of optimality for a feasible solution to the original problem?
- 3- What are the sufficient conditions for the uniqueness of the optimal solution to the original problem?
- 4- How can we design an efficient algorithm to solve the original problem with respect to the answers to questions 1 and 2?

To find the answers to the above questions, two characteristic matrices are defined based on the components of lower and upper bound vector. Some sufficient conditions are presented to remove some equations. Moreover, some sufficient conditions are proposed which under them, one of the optimal solutions of the problem is determined. Then, the sufficient conditions for the uniqueness of the optimal solution are expressed. Furthermore, a modified branch-and-bound method based on a value matrix is designed to solve the

reduced problem. A new algorithm is proposed to solve the original problem based on the conditions and modified branch-and-bound method.

The structure of this paper is organized as follows: Section 2 introduces the linear optimization problem subject to bipolar max-product FREs. We also investigate the characterizations of its feasible domain. In Section 3, the optimality conditions are presented for a feasible solution to the problem and some theorems are given to simplify and reduce the problem. Section 4 proposes an algorithm to solve the problem. Some numerical examples are presented to illustrate the algorithm in Section 5. A comparative study is done with other methods to show the efficiency of the algorithm in Section 6. Finally, conclusions are given in Section 7.

2 Linear programming problem with bipolar max-product FREs

This section is divided into two subsections. In the first subsection, the linear programming problem subject to bipolar FREs is formulated. The characterizations of its feasible domain are finally illustrated in the second subsection.

2.1 Formulation of the problem

Let $A^+ = (a_{ij}^+)$ and $A^- = (a_{ij}^-)$ be two $m \times n$ fuzzy relation matrices with $0 \leq a_{ij}^+, a_{ij}^- \leq 1$ for each $i \in I = \{1, 2, \dots, m\}$ and $j \in J = \{1, 2, \dots, n\}$. Also, assume that $b = (b_1, \dots, b_m)^T \in [0, 1]^m$ and that $c = (c_1, \dots, c_n)$ is a vector of cost coefficients, where $c_j \geq 0$ for each $j \in J$. In this paper, the following programming problem is considered:

$$\min \quad Z(x) = \sum_{j=1}^n c_j x_j, \quad (1)$$

$$\text{s.t.} \quad A^+ \circ x \vee A^- \circ \neg x = b, \quad (2)$$

where $x = (x_1, \dots, x_n)^T \in [0, 1]^n$ is the vector of decision variables to be determined and $\neg x$ denotes the negation of x , that is, $\neg x = (1 - x_1, \dots, 1 - x_n)^T$. The operator of “ \circ ” represents the max- T_p composition, where T_p denotes the product operator. Moreover, $S(A^+, A^-, b) = \{x \in [0, 1]^n \mid A^+ \circ x \vee A^- \circ \neg x = b\}$, which consists of a set of solution vectors $x \in [0, 1]^n$ such that

$$\max_{j \in J} \max \{a_{ij}^+ x_j, a_{ij}^- (1 - x_j)\} = b_i \quad \text{for all } i \in I. \quad (3)$$

Problem (1)–(2) with the real cost coefficients can be converted to a problem with nonnegative cost coefficients in a similar method to Subsection 3.3 in [9]. Hence, without loss of generality, we assume that $c_j \geq 0$ for each $j \in J$.

2.2 The structure of the feasible domain of problem (1)–(2)

A system of bipolar max- T_p FREs $A^+ \circ x \vee A^- \circ \neg x = b$ is called consistent if its solution set, that is, $S(A^+, A^-, b)$, is nonempty. Otherwise, it is inconsistent. Now, we focus on the system of bipolar max- T_p FREs (3), when $S(A^+, A^-, b) \neq \emptyset$.

Lemma 1. A vector $x \in [0, 1]^n$ is a solution for the system of bipolar max- T_p FREs (3) if and only if $\max\{a_{ij}^+ \cdot x_j, a_{ij}^- \cdot (1 - x_j)\} \leq b_i$ for all $i \in I$ and $j \in J$, and for each $i \in I$, there exists an index $j_i \in J$ such that $\max\{a_{ij_i}^+ \cdot x_{j_i}, a_{ij_i}^- \cdot (1 - x_{j_i})\} = b_i$.

Proof. It is obvious. □

Remark 1. For any a_{ij}^+, a_{ij}^- , and b_i with $i \in I$ and $j \in J$, it is assumed that if $a_{ij}^- = 0$, then we define $\max\{1 - \frac{b_i}{a_{ij}^-}, 0\} = 0$. Also, if $a_{ij}^+ = 0$, then we define $\min\{\frac{b_i}{a_{ij}^+}, 1\} = 1$.

Lemma 2. For any a_{ij}^+, a_{ij}^- , and b_i with $i \in I$ and $j \in J$, the inequality of $\max\{a_{ij}^+ \cdot x_j, a_{ij}^- \cdot (1 - x_j)\} \leq b_i$ holds if and only if $\max\{1 - \frac{b_i}{a_{ij}^-}, 0\} \leq x_j \leq \min\{\frac{b_i}{a_{ij}^+}, 1\}$. Especially, if $\max\{a_{ij}^+ \cdot x_j, a_{ij}^- \cdot (1 - x_j)\} \leq b_i = 0$, then at least one of two statements $a_{ij}^+ = 0$ or $a_{ij}^- = 0$ holds. According to Remark 1, $\max\{a_{ij}^+ \cdot x_j, a_{ij}^- \cdot (1 - x_j)\} \leq b_i = 0$ if and only if $\max\{1 - \frac{b_i}{a_{ij}^-}, 0\} \leq x_j \leq \min\{\frac{b_i}{a_{ij}^+}, 1\}$.

Proof. The proof will be divided into four cases as follows:

Case 1. $a_{ij}^+ \neq 0$ and $a_{ij}^- \neq 0$.

Case 2. $a_{ij}^+ = a_{ij}^- = 0$.

Case 3. $a_{ij}^+ = 0$ and $a_{ij}^- \neq 0$.

Case 4. $a_{ij}^+ \neq 0$ and $a_{ij}^- = 0$.

Case 1. It is obvious that the inequality $\max\{a_{ij}^+.x_j, a_{ij}^-. (1-x_j)\} \leq b_i$ holds if and only if $a_{ij}^+.x_j \leq b_i$ and $a_{ij}^-. (1-x_j) \leq b_i$. If $b_i \neq 0$, then the recent inequalities can equivalently be rewritten as $1 - \frac{b_i}{a_{ij}^-} \leq x_j \leq \frac{b_i}{a_{ij}^+}$. On the other hand, according to the assumption, we have $0 \leq x_j \leq 1$. The inequalities are equivalently concluded $\max\{1 - \frac{b_i}{a_{ij}^-}, 0\} \leq x_j \leq \min\{\frac{b_i}{a_{ij}^+}, 1\}$. Whenever $b_i = 0$, the inequalities imply $x_j \leq 0$ and $x_j \geq 1$ which do not hold for any $x_j \in [0, 1]$.

Case 2. Since $a_{ij}^+ = a_{ij}^- = 0$ and $b_i \geq 0$, the inequality of $\max\{a_{ij}^+.x_j, a_{ij}^-. (1-x_j)\} \leq b_i$ holds if and only if $x_j \in [0, 1]$. On the other hand, we have $\max\{1 - \frac{b_i}{a_{ij}^-}, 0\} = 0$ and $\min\{\frac{b_i}{a_{ij}^+}, 1\} = 1$ with regard to Remark 1. Consequently, the result is also true in this case.

Case 3. Since $a_{ij}^+ = 0$, then $\min\{\frac{b_i}{a_{ij}^+}, 1\} = 1$ with regard to Remark 1.

We now have the following subcases: 1. $b_i = 0$ and 2. $b_i \neq 0$. In the first subcase, we have $\max\{1 - \frac{b_i}{a_{ij}^-}, 0\} = 1$, that is, $\max\{1 - \frac{b_i}{a_{ij}^-}, 0\} \leq x_j \leq \min\{\frac{b_i}{a_{ij}^+}, 1\}$ implies that $x_j = 1$. On the other hand, the inequality $\max\{a_{ij}^+.x_j, a_{ij}^-. (1-x_j)\} \leq b_i = 0$ holds if and only if $a_{ij}^-. (1-x_j) = 0$. This implies that $x_j = 1$. In the second subcase, the inequality of $\max\{a_{ij}^+.x_j, a_{ij}^-. (1-x_j)\} \leq b_i$ holds if and only if $a_{ij}^-. (1-x_j) \leq b_i$, which is equivalent to $\max\{1 - \frac{b_i}{a_{ij}^-}, 0\} \leq x_j$. On the other hand, we have $0 \leq x_j \leq 1$. This implies that $\max\{1 - \frac{b_i}{a_{ij}^-}, 0\} \leq x_j \leq 1 = \min\{\frac{b_i}{a_{ij}^+}, 1\}$ with regard to Remark 1. Hence, the result is true in both subcases. \square

Whenever $S(A^+, A^-, b) \neq \emptyset$, the lower and upper bound on the solution set for the system of equations (3) can be determined using the following lemma.

Lemma 3. [1] The vector of $\tilde{x} = (\tilde{x}_1, \dots, \tilde{x}_n)^T$ is the lower bound on the solution set of equations (3), where $\tilde{x}_j = \max_{i \in I} \{1 - \frac{b_i}{a_{ij}^-} \mid a_{ij}^- > b_i\}$ for each $j \in J$. Also, the vector of $\hat{x} = (\hat{x}_1, \dots, \hat{x}_n)^T$ is the upper bound on the solution set of equations (3), where $\hat{x}_j = \min_{i \in I} \{\frac{b_i}{a_{ij}^+} \mid a_{ij}^+ > b_i\}$, for each $j \in J$. Furthermore, if there exists $j \in J$ such that for each $i \in I$, $a_{ij}^- \leq b_i$, then $\tilde{x}_j = 0$ and if there exists $j \in J$ such that for each $i \in I$, $a_{ij}^+ \leq b_i$, then $\hat{x}_j = 1$, that is, $\max \emptyset = 0$ and $\min \emptyset = 1$ are defined.

Lemma 4. [1] Assume that $S(A^+, A^-, b) \neq \emptyset$ and that its lower and upper bound are \tilde{x} and \hat{x} , respectively. If there exists $j_0 \in J$ such that $\tilde{x}_{j_0} = \hat{x}_{j_0}$, then $x_{j_0} = \tilde{x}_{j_0} = \hat{x}_{j_0}$ for all $x \in S(A^+, A^-, b)$. Also, the solution set of system (2) (or (3)) is the same to the following system:

$$\begin{cases} \max_{j \in J - \{j_0\}} \max\{a_{ij}^+.x_j, a_{ij}^-. (1-x_j)\} = b_i & \text{for all } i \in I - \bar{I}, \\ \tilde{x}_j \leq x_j \leq \hat{x}_j & \text{for all } j \in J - \{j_0\}; x_{j_0} = \tilde{x}_{j_0} = \hat{x}_{j_0}, \end{cases}$$

where \tilde{x}_j and \hat{x}_j for all $j \in J$ are defined on the basis of system (2) (or (3)) and $\bar{I} = \{i \in I \mid \max\{a_{ij_0}^+ \cdot x_{j_0}, a_{ij_0}^- \cdot (1 - x_{j_0})\} = b_i\}$ and $\bar{I} \neq \emptyset$.

We first consider the special case of $b_i = 0$ for $i \in I$. Since $\max\{a_{ij}^+ \cdot x_j, a_{ij}^- \cdot (1 - x_j)\}$ is always nonnegative, the following inequality $\max\{a_{ij}^+ \cdot x_j, a_{ij}^- \cdot (1 - x_j)\} \leq b_i$ can be converted into the equation $\max\{a_{ij}^+ \cdot x_j, a_{ij}^- \cdot (1 - x_j)\} = b_i$. We now express the following lemma.

Lemma 5. Let $b_i = 0$ for $i \in I$. A vector x is a solution for i th equation of the system (3) if and only if $\max\{1 - \frac{b_i}{a_{ij}^-}, 0\} \leq x_j \leq \min\{\frac{b_i}{a_{ij}^+}, 1\}$ for each $j \in J$.

Proof. Considering Remark 1, the proof can be easily obtained by the proofs of Lemmas 1 and 2. \square

Lemma 6. [1] Suppose that $S(A^+, A^-, b) \neq \emptyset$, and that its lower and upper are \tilde{x}_j and \hat{x}_j , respectively. Then the solution set of system (2) (or (3)) is the same to the following system:

$$\begin{cases} \max_{j \in J} \max\{a_{ij}^+ \cdot x_j, a_{ij}^- \cdot (1 - x_j)\} = b_i & \text{for all } i \in I - I_0, \\ \tilde{x}_j \leq x_j \leq \hat{x}_j & \text{for all } j \in J, \end{cases}$$

where $I_0 = \{i \in I \mid b_i = 0\}$. Also, \tilde{x}_j and \hat{x}_j , for all $j \in J$, are defined on the basis of system (2) (or (3)).

Furthermore, if $b_i = 0$ for each $i \in I$, then we can easily obtain the solution set of equations (3) and an optimal solution for problem (1)–(2). These points are expressed in the following corollary.

Corollary 1. If $b = 0$ in the constraint part of problem (1)–(2), then we have

$S(A^+, A^-, b) = \{x \mid \tilde{x} \leq x \leq \hat{x}\}$ and $x^* = \tilde{x}$ is an optimal solution to problem (1)–(2).

Proof. It can be easily seen that $S(A^+, A^-, b) = \{x \mid \tilde{x} \leq x \leq \hat{x}\}$ with regard to Lemmas 3 and 5. The proof is completed by showing that $x^* = \tilde{x}$ is an optimal solution of problem (1)–(2). Since $\tilde{x} \in S(A^+, A^-, b)$, $c \geq 0$, and the problem is minimization, then \tilde{x} is an optimal solution of problem (1)–(2). \square

Without loss of generality, from now on, we will assume that $\tilde{x}_j < \hat{x}_j$, for each $j \in J$, and $b_i > 0$, for each $i \in I$.

Remark 2. In this paper, if $a_{ij}^+ = a_{ij}^- = 0$, then we define $\frac{a_{ij}^+ \cdot a_{ij}^-}{a_{ij}^+ + a_{ij}^-} = 0$.

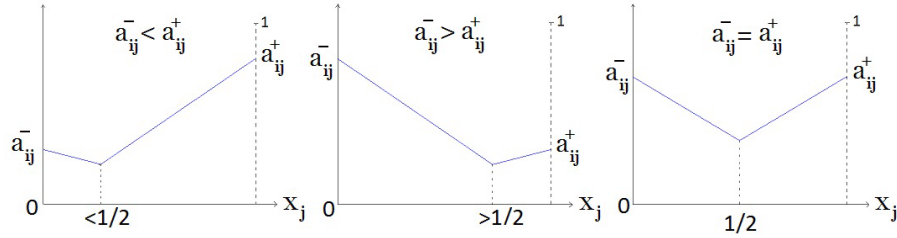


Figure 1: Illustration of function $f(x_j) = \max\{a_{ij}^+ x_j, a_{ij}^- (1 - x_j)\}$

Lemma 7. For any a_{ij}^+, a_{ij}^- , and b_i with $i \in I$ and $j \in J$, the equation of $\max\{a_{ij}^+ x_j, a_{ij}^- (1 - x_j)\} = b_i$ has a solution if and only if $\frac{a_{ij}^+ a_{ij}^-}{a_{ij}^+ + a_{ij}^-} \leq b_i \leq \max\{a_{ij}^+, a_{ij}^-\}$. Also, its solution set is determined with regard to the following cases:

Case 1: If $a_{ij}^- < b_i \leq a_{ij}^+$, then $S(a_{ij}^+, a_{ij}^-, b_i) = \left\{ \min\left(\frac{b_i}{a_{ij}^+}, 1\right) \right\}$;

Case 2: If $a_{ij}^+ < b_i \leq a_{ij}^-$, then $S(a_{ij}^+, a_{ij}^-, b_i) = \left\{ \max\left(1 - \frac{b_i}{a_{ij}^-}, 0\right) \right\}$;

Case 3: If $\frac{a_{ij}^+ a_{ij}^-}{a_{ij}^+ + a_{ij}^-} \leq b_i \leq \min\{a_{ij}^+, a_{ij}^-\}$, then

$$S(a_{ij}^+, a_{ij}^-, b_i) = \left\{ \max\left(1 - \frac{b_i}{a_{ij}^-}, 0\right), \min\left(\frac{b_i}{a_{ij}^+}, 1\right) \right\}.$$

Proof. For given $a_{ij}^+, a_{ij}^- \in [0, 1]$, and $b_i > 0$, the range of the function of $\max\{a_{ij}^+ x_j, a_{ij}^- (1 - x_j)\}$ and the solution set of $S(a_{ij}^+, a_{ij}^-, b_i)$ can be observed from Figure 1 and determined, easily. \square

Note that the vectors of \tilde{x} and \hat{x} are only the lower and upper bound on the solution set of equations (3), respectively. They are not necessarily feasible solutions to the system of equations (3). Moreover, we have $S(A^+, A^-, b) \subseteq \{x \mid \tilde{x} \leq x \leq \hat{x}\}$. Considering Lemmas 3 and 7, each equation in system (3) can be satisfied by \tilde{x}_j or \hat{x}_j . In order to store these facts, the characteristic matrices of Q^+ and Q^- are defined below.

Definition 1. Define two characteristic matrices $Q^+ = (q_{ij}^+)_{m \times n}$ and $Q^- = (q_{ij}^-)_{m \times n}$ such that for each $i \in I$ and $j \in J$, we have

$$q_{ij}^+ = \begin{cases} 1 & \text{if } a_{ij}^+ \hat{x}_j = b_i, \\ 0 & \text{otherwise,} \end{cases} \quad \text{and} \quad q_{ij}^- = \begin{cases} 1 & \text{if } a_{ij}^- (1 - \tilde{x}_j) = b_i, \\ 0 & \text{otherwise.} \end{cases}$$

Also, a series of index sets is defined as follows.

Definition 2. (i) [1] For the matrix Q^+ , define

$$I_j^+(x) = \{i \in I \mid x_j = \hat{x}_j \text{ and } q_{ij}^+ = 1\} \text{ and } J_i^+(x) = \{j \in J \mid x_j = \hat{x}_j \text{ and } q_{ij}^+ = 1\}.$$

Also, for the matrix Q^- , define

$$I_j^-(x) = \{i \in I \mid x_j = \check{x}_j \text{ and } q_{ij}^- = 1\} \text{ and } J_i^-(x) = \{j \in J \mid x_j = \check{x}_j \text{ and } q_{ij}^- = 1\},$$

for each $i \in I$ and $j \in J$. Furthermore, let $I_j(x) = I_j^+(x) \cup I_j^-(x)$, for each $j \in J$.

(ii) Let $I_j^+ = I_j^+(\hat{x})$, $J_i^+ = J_i^+(\hat{x})$, let $I_j^- = I_j^-(\check{x})$, and let $J_i^- = J_i^-(\check{x})$, for each $i \in I$ and $j \in J$.

Considering Lemmas 2 and 7 and the above concepts, we present a necessary and sufficient condition for the solution of system (3) (or (2)).

Theorem 1. A vector $x \in [0, 1]^n$ is a solution for the system of bipolar max- T_p FREs $A^+ \circ x \vee A^- \circ \neg x = b$ if and only if $\check{x} \leq x \leq \hat{x}$ and $\bigcup_{j \in J} I_j(x) = I$.

Proof. With regard to Lemma 1, we show that $\check{x} \leq x \leq \hat{x}$ and $\bigcup_{j \in J} I_j(x) = I$

if and only if $\max \{a_{ij}^+ \cdot x_j, a_{ij}^- \cdot (1 - x_j)\} \leq b_i$, for all $i \in I$ and $j \in J$, and for each $i \in I$ there exists an index $j_i \in J$ such that

$\max \{a_{ij_i}^+ \cdot x_{j_i}, a_{ij_i}^- \cdot (1 - x_{j_i})\} = b_i$. Considering Lemmas 2 and 3, all the inequalities $\max \{a_{ij}^+ \cdot x_j, a_{ij}^- \cdot (1 - x_j)\} \leq b_i$, for $i \in I$ and $j \in J$, hold if and only if $\check{x} \leq x \leq \hat{x}$. We now focus on proving $\bigcup_{j \in J} I_j(x) = I$. The equality

is true if and only if for each $i \in I$, there exists an index $j_i \in J$ such that $\max \{a_{ij_i}^+ \cdot x_{j_i}, a_{ij_i}^- \cdot (1 - x_{j_i})\} = b_i$. Since $I_j(x) = I_j^+(x) \cup I_j^-(x)$, for each $j \in J$, we have $\bigcup_{j \in J} I_j(x) = I$ if and only if $\bigcup_{j \in J} (I_j^+(x) \cup I_j^-(x)) = I$. Moreover,

we have $\bigcup_{j \in J} (I_j^+(x) \cup I_j^-(x)) = I$ if and only if for each $i \in I$ there exists an

index $j_i \in J$ such that $(x_{j_i} = \hat{x}_{j_i} \text{ \& } q_{ij_i}^+ = 1)$ or $(x_{j_i} = \check{x}_{j_i} \text{ \& } q_{ij_i}^- = 1)$, with regard to Definition 2. By Definition 1 and Lemma 7, it can easily be seen that for each $i \in I$, there exists an index $j_i \in J$ such that $(x_{j_i} = \hat{x}_{j_i} \text{ \& } q_{ij_i}^+ = 1)$ or $(x_{j_i} = \check{x}_{j_i} \text{ \& } q_{ij_i}^- = 1)$ if and only if $\max \{a_{ij_i}^+ \cdot x_{j_i}, a_{ij_i}^- \cdot (1 - x_{j_i})\} = b_i$. \square

For each $j \in J$, label the values of \hat{x}_j and \check{x}_j with boolean variables of y_j and $\neg y_j$, respectively. The following theorem is used to determine the consistency of system $A^+ \circ x \vee A^- \circ \neg x = b$.

Theorem 2. A system of bipolar max- T_p FREs $A^+ \circ x \vee A^- \circ \neg x = b$ is consistent if and only if its characteristic boolean formula $C = \bigwedge_{i \in I} C_i$ is well-defined and satisfiable, where $C_i = \bigvee_{j \in J_i^+} y_j \vee \bigvee_{j \in J_i^-} \neg y_j$.

Proof. With regard to Definitions 1 and 2, the proof is similar to the proof of Theorem 2.5 in [18] \square

Theorem 3. Let \tilde{x} and \hat{x} be the lower and upper bound, respectively. Then

$$I_j(x) \subseteq I_j^+ \cup I_j^-, \quad \text{for all } x \in S(A^+, A^-, b), \quad \text{for all } j \in J.$$

Exactly, we have $I_j(x) = I_j^+$ (when $x_j = \hat{x}_j$) or $I_j(x) = I_j^-$ (when $x_j = \tilde{x}_j$) or $I_j(x) = \emptyset$ (when $\tilde{x}_j < x_j < \hat{x}_j$).

Proof. The proof is obtained from Definitions 1 and 2. \square

We are now ready to present some theorems to simplify problem (1)–(2) in the next section.

3 Some optimality sufficient conditions for problem (1)–(2)

This section studies some optimality conditions for problem (1)–(2). One of its optimal solutions is found under the conditions. Moreover, some sufficient conditions are proposed to guarantee its uniqueness, and a closed form is proposed to determine it. In this section, it is assumed that $S(A^+, A^-, b) \neq \emptyset$. The following lemma presents a useful property of the optimal solution of problem (1)–(2).

Lemma 8. Consider the optimization problem of (1)–(2). Then there exists an optimal solution $x^* = (x_1^*, \dots, x_n^*)^T$ such that for each $j \in J$ either $x_j^* = \hat{x}_j$ or $x_j^* = \tilde{x}_j$.

Proof. The proof is similar to the proof of Lemma 4 in [20]. \square

Some theorems are first presented to reduce the search domain of the optimal solution of problem (1)–(2). The dimensions of the matrices of Q^+ and Q^- can be reduced using these theorems.

Theorem 4. Let $T_{i_1} = \{i \in I \setminus \{i_1\} \mid J_i^+ \supseteq J_{i_1}^+ \text{ \& } J_i^- \supseteq J_{i_1}^-\}$, for $i_1 \in I$. If for vector x , where $\tilde{x} \leq x \leq \hat{x}$, $\max_{j \in J} \max\{T_P(a_{ij}^+, x_j), T_P(a_{ij}^-, (1 - x_j))\} = b_i$ holds for $i = i_1$, then for each $i \in T_{i_1}$, we have

$$\max_{j \in J} \max\{T_P(a_{ij}^+, x_j), T_P(a_{ij}^-, (1 - x_j))\} = b_i.$$

Proof. If for the vector x , where $\tilde{x} \leq x \leq \hat{x}$, the following equality $\max_{j \in J} \max\{T_P(a_{ij}^+, x_j), T_P(a_{ij}^-, (1 - x_j))\} = b_i$ holds for $i = i_1$, then there exist some $j_1 \in J$ such that $\max\{T_P(a_{i_1 j_1}^+, x_{j_1}), T_P(a_{i_1 j_1}^-, (1 - x_{j_1}))\} = b_{i_1}$.

Since these equalities hold only at the values of \tilde{x}_{j_1} , for $j_1 \in J_{i_1}^-(x)$, or \hat{x}_{j_1} , for $j_1 \in J_{i_1}^+(x)$, then $j_1 \in J_{i_1}^+(x) \cup J_{i_1}^-(x) \neq \emptyset$. Also, with regard to Definition 2, $J_{i_1}^+(x) \subseteq J_{i_1}^+$ and $J_{i_1}^-(x) \subseteq J_{i_1}^-$. On the other hand, for each $i \in T_{i_1}$, we have $J_i^+ \supseteq J_{i_1}^+$ and $J_i^- \supseteq J_{i_1}^-$. Therefore, for each $i \in T_{i_1}$, $J_{i_1}^+(x) \subseteq J_i^+$ and $J_{i_1}^-(x) \subseteq J_i^-$ because $J_{i_1}^+(x) \subseteq J_{i_1}^+$, $J_{i_1}^-(x) \subseteq J_{i_1}^-$, $J_{i_1}^+ \subseteq J_i^+$, and $J_{i_1}^- \subseteq J_i^-$, for each $i \in T_{i_1}$. Since $j_1 \in J_{i_1}^+(x) \cup J_{i_1}^-(x)$, $J_{i_1}^+(x) \subseteq J_i^+$, and $J_{i_1}^-(x) \subseteq J_i^-$, for each $i \in T_{i_1}$, then it is concluded that $(j_1 \in J_{i_1}^+(x) \text{ and } j_1 \in J_i^+)$ or $(j_1 \in J_{i_1}^-(x) \text{ and } j_1 \in J_i^-)$. So, we can write $(x_{j_1} = \hat{x}_{j_1} \text{ and } q_{ij_1}^+ = 1)$ or $(x_{j_1} = \tilde{x}_{j_1} \text{ and } q_{ij_1}^- = 1)$, for each $i \in T_{i_1}$. With regard to Definition 1, $T_p(a_{ij_1}^+, x_{j_1}) = b_i$ or $T_p(a_{ij_1}^-, (1 - x_{j_1})) = b_i$, for each $i \in T_{i_1}$. Thus, the equality $\max\{T_p(a_{ij_1}^+, x_{j_1}), T_p(a_{ij_1}^-, (1 - x_{j_1}))\} = b_i$ holds true for each $i \in T_{i_1}$. Since $\tilde{x} \leq x \leq \hat{x}$ and $\max\{T_p(a_{ij_1}^+, x_{j_1}), T_p(a_{ij_1}^-, (1 - x_{j_1}))\} = b_i$, for each $i \in T_{i_1}$, then the equality $\max_{j \in J} \max\{T_p(a_{ij}^+, x_j), T_p(a_{ij}^-, (1 - x_j))\} = b_i$ holds true for each $i \in T_{i_1}$. \square

The following corollary is a direct result of Theorem 4.

Corollary 2. Under the conditions of Theorem 4, all the equations with numbers $i \in T_{i_1}$ can be removed from the matrices of Q^+ and Q^- once \tilde{x} and \hat{x} have been obtained.

In the next theorem, an equivalent system with system (2) is presented.

Theorem 5. A binary vector $u \in \{0, 1\}^n$ induces a solution $x = \tilde{x} + Vu$ for the system of bipolar max- T_p FREs $A^+ \circ x \vee A^- \circ \neg x = b$ if and only if $(Q^+ - Q^-)u + Q^-e_n \geq e_m$, where $V = \text{diag}(\hat{x} - \tilde{x})$ and e_k is a k -dimensional vector with the unit components.

Proof. The proof is similar to the proof of Theorem 3 in [20]. \square

Lemma 9. If there exists a pair $i \in I$ and $k \in J$ such that $q_{ik}^+ = q_{ik}^- = 1$, then the reduced system of $(Q^{+'} - Q^{-'})u + Q^{-'}e_n \geq e'_{m-1}$ and the system of $(Q^+ - Q^-)u + Q^-e_n \geq e_m$ have the same solution set, where the matrices of $Q^{+'}$, $Q^{-'}$, and e'_{m-1} are obtained by removing the i th row of the matrices of Q^+ , Q^- , and e_m , respectively.

Proof. Let $S(Q^+, Q^-, u) = \{u \in \{0, 1\}^n \mid (Q^+ - Q^-)u + Q^-e_n \geq e_m\}$ and $S(Q^{+'}, Q^{-'}, u) = \{u \in \{0, 1\}^n \mid (Q^{+'} - Q^{-'})u + Q^{-'}e_n \geq e'_{m-1}\}$, where $Q^{+'}$, $Q^{-'}$, and e'_{m-1} are obtained from Q^+ , Q^- , and e_m by removing their row of i , respectively. We will now show $S(Q^+, Q^-, u) = S(Q^{+'}, Q^{-'}, u)$. Consider the i th equation of system (2). Since $q_{ik}^+ = q_{ik}^- = 1$ and $u_k \in \{0, 1\}$, the i th inequality of $(Q^+ - Q^-)u + Q^-e_n \geq e_m$, or equivalently $Q^+u + Q^-(e_n - u) \geq e_m$, is automatically satisfied. Therefore, $S(Q^+, Q^-, u) = S(Q^{+'}, Q^{-'}, u)$. \square

Theorem 6. If there exists a pair $i \in I$ and $k \in J$ such that $q_{ik}^+ = q_{ik}^- = 1$, then we can remove the i th row in the computation of the minimum objective value.

Proof. It is obvious from Lemma 9 and Theorem 4 in [20]. \square

Now, some sufficient conditions are presented to determine one of the optimal solutions of problem (1)–(2). First of all, we express the following definition.

Definition 3. Define two index sets I_1 and I_2 as follows:

$$I_1 = \bigcup_{j \in J} I_j^- \quad \text{and} \quad I_2 = I \setminus I_1.$$

Corollary 3. If $I_2 = \emptyset$, then $x^* = \tilde{x}$ is an optimal solution of problem (1)–(2).

Proof. If $\bigcup_{j \in J} I_j^- = I$, then \tilde{x} is a feasible solution for the system of equations (2). Hence, we can assign \tilde{x}_j to x_j^* due to $c_j \geq 0$, for each $j \in J$. \square

If the vector of \tilde{x} is a feasible solution of problem (1)–(2), then \tilde{x} is an optimal solution due to $c_j \geq 0$, for each $j \in J$. Otherwise, we have to construct a solution x^* with elements $x_j^* = \tilde{x}_j$ or $x_j^* = \hat{x}_j$ such that $\bigcup_{j \in J} I_j(x^*) = I$ and the value of $Z(x^*)$ is the minimum objective value. With regard to these points, we present the following theorems.

Theorem 7. If there exists an index $k \in \bigcap_{i \in I_2} J_i^+$ such that the following conditions are satisfied

1. for all $j \in \bigcup_{i \in I_2} J_i^+$, $c_k(\hat{x}_k - \tilde{x}_k) \leq c_j(\hat{x}_j - \tilde{x}_j)$ and
2. $I_k^- \setminus I_k^+ \subseteq \bigcup_{\substack{j \in J \\ j \neq k}} I_j^-$,

then there exists an optimal solution $x^* = (x_j^*)_{j \in J}$ for problem (1)–(2) as follows:

$$x_j^* = \begin{cases} \hat{x}_k & \text{if } j = k, \\ \tilde{x}_j & \text{otherwise,} \end{cases} \quad \text{for all } j \in J. \quad (4)$$

Proof. We show that x^* is an optimal solution of problem (1)–(2). It is enough to show that i) $x^* \in S(A^+, A^-, b)$ and ii) $Z(x^*) \leq Z(x)$, for each $x \in S(A^+, A^-, b)$.

i) With regard to the structure of vector x^* and $I_k^- \setminus I_k^+ \subseteq \bigcup_{\substack{j \in J \\ j \neq k}} I_j^-$, the following

equalities hold:

$$\bigcup_{j \in J} I_j(x^*) = \left(\bigcup_{\substack{j \in J \\ j \neq k}} I_j^- \right) \cup I_k^+ = \left(\bigcup_{j \in J} I_j^- \right) \cup I_k^+. \quad (5)$$

On the other hand, $k \in \bigcap_{i \in I_2} J_i^+$ implies that $I_k^+ \supseteq I_2 = I \setminus \bigcup_{j \in J} I_j^-$. Therefore, we have

$$I_k^+ \cup \left(\bigcup_{j \in J} I_j^- \right) = I. \quad (6)$$

With regard to the expressions of (5) and (6), it is concluded that $\bigcup_{j \in J} I_j(x^*) = I$. Since $\check{x} \leq x^* \leq \hat{x}$ and $\bigcup_{j \in J} I_j(x^*) = I$, the vector x^* is a feasible solution for the system of equations (3) with regard to Theorem 1.

ii) For each $x \in S(A^+, A^-, b)$ and $x \neq x^*$, we have $x_k = \hat{x}_k$ or there exists an index $j \in \bigcup_{i \in I_2} J_i^+ \setminus \{k\}$ such that $x_j = \hat{x}_j$.

If $x_k = \hat{x}_k$, then there exists an index $j' \in J \setminus \{k\}$ such that $x_{j'} = \hat{x}_{j'}$ due to $x \neq x^*$. Hence, we have $Z(x^*) \leq Z(x^*) + c_{j'}(\hat{x}_{j'} - \check{x}_{j'}) \leq Z(x)$. If there exists an index $j \in \bigcup_{i \in I_2} J_i^+ \setminus \{k\}$ such that $x_j = \hat{x}_j$, then $Z(x) \geq Z(x^*)$ with regard to the condition 1 in Theorem 7. \square

Note that the optimal solution introduced in the relation (4) is not necessarily unique with regard to condition 1 in Theorem 7 and $c_j \geq 0$, for each $j \in J$. Considering Theorem 7, some sufficient conditions are expressed in Lemma 10 that under them, problem (1)–(2) has a unique optimal solution and the optimal solution is explicitly determined.

Lemma 10. If for each $j \in J \setminus \bigcup_{i \in I_2} J_i^+$, $c_j > 0$ and there exists an index $k \in \bigcap_{i \in I_2} J_i^+$ such that the conditions

1. for all $j \in \bigcup_{i \in I_2} J_i^+ \setminus \{k\}$, $c_k(\hat{x}_k - \check{x}_k) < c_j(\hat{x}_j - \check{x}_j)$, and
2. $I_k^- \setminus I_k^+ \subseteq \bigcup_{\substack{j \in J \\ j \neq k}} I_j^-$

are satisfied, then the optimization problem of (1)–(2) has a unique optimal solution of $x^* = (x_j^*)_{j \in J}$ as relation (4).

Proof. Since the assumptions of Theorem 7 hold, problem (1)–(2) has an optimal solution as the relation (4). We now show its uniqueness. By the assumptions of $\check{x}_j < \hat{x}_j$ and $c_j \geq 0$, for each $j \in J$, the condition 1 in Lemma 10 implies that $c_j > 0$, for each $j \in \bigcup_{i \in I_2} J_i^+ \setminus \{k\}$. This point implies that $c_j >$

0, for each $j \in J \setminus \{k\}$, with regard to the assumption for all $j \in J \setminus \bigcup_{i \in I_2} J_i^+$,

$c_j > 0$. For each $x \in S(A^+, A^-, b)$ and $x \neq x^*$, we have $x_k = \hat{x}_k$ or there exists an index $j \in \bigcup_{i \in I_2} J_i^+ \setminus \{k\}$ such that $x_j = \hat{x}_j$. If $x_k = \hat{x}_k$, then there

exists an index $j' \in J \setminus \{k\}$ such that $x_{j'} = \hat{x}_{j'}$ due to $x \neq x^*$. Hence, we have $Z(x^*) < Z(x^*) + c_{j'}(\hat{x}_{j'} - \check{x}_{j'}) \leq Z(x)$ due to the condition 1. If there exists $j \in \bigcup_{i \in I_2} J_i^+ \setminus \{k\}$ such that $x_j = \hat{x}_j$, then $Z(x) > Z(x^*)$ due to $c_j > 0$,

for each $j \in J \setminus \{k\}$. This shows the uniqueness of optimal solution x^* for problem (1)–(2). \square

The lemmas, theorems, and corollaries of this section are firstly applied to reduce the size of the original problem. If all of the components of the optimal solution of problem (1)–(2) were not determined, then we have to solve the reduced problem. To do this, we will explain the modified branch-and-bound method to find the rest of its components in the next section.

4 A procedure for the resolution of problem (1)–(2)

We first define a simple value matrix in this section. Applying the value matrix and some points, the branch-and-bound method with the jump-tracking technique is modified to solve problem (1)–(2). Finally, an algorithm is proposed for the resolution of problem (1)–(2).

4.1 Modified branch-and-bound method

In this subsection, we rewrite the objective function (1) as follows:

$$Z(x) = \sum_{j=1}^n c_j x_j - \sum_{j=1}^n c_j \tilde{x}_j + \sum_{j=1}^n c_j \tilde{x}_j = \sum_{j=1}^n c_j (x_j - \tilde{x}_j) + \sum_{j=1}^n c_j \tilde{x}_j.$$

Now, the optimal solutions of problem (1)–(2) are the same with the optimal solutions of the following problem:

$$\min \quad \bar{Z}(x) = \sum_{j=1}^n c_j (x_j - \tilde{x}_j), \quad (7)$$

$$\text{s.t.} \quad A^+ \circ x \vee A^- \circ \neg x = b, \quad (8)$$

$$x \in [0, 1]^n. \quad (9)$$

It is necessary to recall that $Z(x^*) = \bar{Z}(x^*) + \sum_{j=1}^n c_j \tilde{x}_j$. Therefore, we try to find the optimal solution of problem (7)–(9). Since $\tilde{x} \leq x \leq \hat{x}$, for each $x \in S(A^+, A^-, b)$ and $c \geq 0$, we have $\bar{Z}(x) \geq 0$ and $Z(x) \geq \sum_{j \in J} c_j \tilde{x}_j$ for each $x \in S(A^+, A^-, b)$. As it was stated before, if $\tilde{x} \in S(A^+, A^-, b)$, then we can set \tilde{x} as an optimal solution. Otherwise, we have $\tilde{x} \notin S(A^+, A^-, b)$. In order to find the optimal solution of problem (1)–(2) (or (7)–(9)) according to Lemma 8, we have to set some \hat{x}_j instead of \tilde{x}_j , $j \in J$, in the vector \tilde{x} such that the new vector \tilde{x}^* satisfy all equations (8)–(9) and minimize the

objective function (7). To do this, we first express a useful property of the objective function (7).

Proposition 1. Let $x = (x_j)_{j \in J}$ be a vector in $S(A^+, A^-, b)$ where $x_t = \tilde{x}_t$ and $x_k = \hat{x}_k$ for two indices t and k such that $t \neq k$ and vectors \tilde{x} and \hat{x} are lower and upper bounds of system (8)–(9), respectively. Construct two new vectors $x' = (x'_j)_{j \in J}$ and $x'' = (x''_j)_{j \in J}$ such that $x'_t = \hat{x}_t$ and $x'_j = x_j$, for each $j \in J \setminus \{t\}$, and $x''_k = \hat{x}_k$ and $x''_j = x_j$, for each $j \in J \setminus \{k\}$. If $c_k(\hat{x}_k - \tilde{x}_k) < c_t(\hat{x}_t - \tilde{x}_t)$, then $\bar{Z}(x'') < \bar{Z}(x')$.

Proof. It is obvious. \square

It is necessary to recall the importance of the form of problem (7)–(9) with respect to problem (1)–(2). With regard to Proposition 1, if $c_k \hat{x}_k > c_t \hat{x}_t$, then we could not conclude that $Z(x'') > Z(x')$ or $Z(x'') < Z(x')$, but problem (7)–(9) gives us more information. If $c_k(\hat{x}_k - \tilde{x}_k) < c_t(\hat{x}_t - \tilde{x}_t)$, then we have $\bar{Z}(x'') < \bar{Z}(x')$. We use the useful property of the objective function (7) to present a modified branch-and-bound method to solve problem (1)–(2). First of all, we consider the following remark.

Remark 3. Rearrange the rows of matrices Q^+ and Q^- such that the first $|I_2|$ rows in these matrices are the rows $i \in I_2$. More precisely, transfer all rows $i \in I_2$ to the top $|I_2|$ of the rows of the matrices Q^+ and Q^- .

According to Lemma 8, there exists an optimal solution $x^* = (x^*_j)_{j \in J}$ such that either $x^*_j = \hat{x}_j$ or $x^*_j = \tilde{x}_j$ for each $j \in J$. Hence, it is concluded that $I_j(x^*) = I_j^+$ or $I_j(x^*) = I_j^-$, for each $j \in J$, with regard to Theorem 3. Since $I_j(x^*) = I_j^+$ or $I_j(x^*) = I_j^-$, for each $j \in J$, we hereinafter focus on these columns. To do this, the following value matrix is defined based on problem (7)–(9).

Definition 4. Define the value matrix of $M = (m_{ij})_{m \times 2n}$, where

$$m_{i,2j-1} = \begin{cases} c_j(\hat{x}_j - \tilde{x}_j) & \text{if } q_{ij}^+ = 1, \\ \infty & \text{otherwise,} \end{cases} \quad \text{and} \quad m_{i,2j} = \begin{cases} 0 & \text{if } q_{ij}^- = 1, \\ \infty & \text{otherwise,} \end{cases}$$

for each $i \in I$ and $j \in J$.

We will employ the branch-and-bound method with the jump-tracking technique to solve problem (1)–(2) using the value matrix M . Since \hat{x}_j and \tilde{x}_j , for each $j \in J$, cannot be selected simultaneously along a branch, we should modify the branch-and-bound method. We consider three modifications on this method similar to [1] as follows:

1. If we choose \hat{x}_j (or \tilde{x}_j) to branch from one node to another node, then we never use \tilde{x}_j (or \hat{x}_j) to branch further on the current node.

2. Under the stated conditions below, we cannot branch further on Node k .
 - 2.1. We have reached to the last row of the matrix M .
 - 2.2. The selected variables along Node 0 to Node k together with \tilde{x}_j , for each $j \in J \setminus J_k$, satisfy all the equations, where $J_k = \{j \in J | x_j \text{ has been selected along the branches from Node 0 to Node } k\}$.
 - 2.3. We do not have any candidate for satisfying an equation with regard to modification 1.
3. If we cannot branch further on Node k under the conditions 2.1 and 2.2, then we assign \tilde{x}_j to x_j for each $j \in J \setminus J_k$.

Note that if we cannot branch further on Node k with the value of Z_k under the conditions 2.1 and 2.2, then Z_k represents the objective value of problem (7)–(9) for the obtained vector x along Node 0 to Node k . Then the total of Z_k along the branches from Node 0 to Node k can be calculated as follows:

$$\text{Total } Z_k = Z_k + \sum_{j \in J} c_j \tilde{x}_j, \quad (10)$$

where total Z_k represents the objective value of problem (1)–(2) for the obtained vector x .

In problem (7)–(9), each equation i , for $i \in I_1$, of its constraints can be satisfied by \tilde{x}_j , for $j \in J_i^-$, or \hat{x}_j , for $j \in J_i^+$. In this case, the i th equation may be satisfied without imposing any extra cost to the objective function. On the other hand, each equation i , for $i \in I_2$, can only be satisfied by \hat{x}_j , for $j \in J_i^+$, that is, we have to expend extra cost for satisfying each equation i , for $i \in I_2$. Since each equation i , for $i \in I_2$, is satisfied with extra cost, we start the modified branch-and-bound method from row(s) $i \in I_2$. In this case, the i th equation, for $i \in I_1$, may be satisfied with the expended cost for $i \in I_2$ or without imposing any extra cost to the objective function. Hence, we expect that the visited nodes of the modified branch-and-bound method can be decreased when we consider Remark 3.

We are now ready to design an algorithm to solve problem (1)–(2) based on the obtained results up to now.

4.2 An algorithm for the resolution of problem (1)–(2)

Algorithm 1. Consider the optimization problem (1)–(2).

Step 1. Compute the lower and upper bound of \tilde{x} and \hat{x} applying Lemma 3.

Step 2. If $b = 0$ and $\tilde{x}_j \leq \hat{x}_j$, for each $j \in J$, then $S(A^+, A^-, b) = \{x \mid \tilde{x} \leq x \leq \hat{x}\}$ and $x^* = \tilde{x}$ is an optimal solution of problem (1)–(2) with regard to Corollary 1 and stop!

Step 3. If $\tilde{x}_j < \hat{x}_j$, for each $j \in J$, and $b_i > 0$, for each $i \in I$, then go to Step 4. Otherwise, use Lemmas 4 and 6.

Step 4. Compute the matrices of Q^+ and Q^- , the index sets of I_j^+ and I_j^- , for each $j \in J$, and the index sets of J_i^+ and J_i^- , for each $i \in I$ using Definitions 1 and 2.

Step 5. Check the consistency of bipolar max- T_p FREs (2) using Theorem 2. If it is inconsistent, then stop! Otherwise, go to Step 6.

Step 6. Perform the process of problem reduction as follows:

6.1. Compute two index sets I_1 and I_2 using Definition 3.

6.2. If $I_2 = \emptyset$, then $x^* = \tilde{x}$ is an optimal solution of problem (1)–(2) with regard to Corollary 3 and stop!

6.3. If the conditions of Lemma 10 are satisfied, then the unique optimal solution x^* of problem (1)–(2) can be obtained by relation (4) and stop!

6.4. Check the conditions of Theorem 7. If the conditions are satisfied, then there exists an optimal solution x^* according to relation (4) and stop!

6.5. Check the conditions of Theorem 4. If the conditions are satisfied, then remove all the equations with numbers $i \in T_{i_1}$ from the matrices of Q^+ and Q^- with regard to Corollary 2.

6.6. If there exists a pair $i \in I$ and $k \in J$ such that $q_{ik}^+ = q_{ik}^- = 1$, then we can remove the row of i in the computation of the minimum objective value with regard to Theorem 6.

Step 7. If $Q^+ = Q^- = \emptyset$, then assign \tilde{x}_j to x_j^* and go to Step 10.

Step 8. Rearrange the rows of the matrices Q^+ and Q^- according to Remark 3. Also, generate the value matrix of M using Definition 4.

Step 9. Employ the modified branch-and-bound method with the jump-tracking technique on the matrix M to solve the optimization problem of (1)–(2).

Step 10. Produce the optimal solution and the optimal value of problem (1)–(2). End.

5 Numerical examples

We now illustrate Algorithm 1 by the following examples.

Example 1. Consider the following optimization problem:

$$\min \quad x_1 + 3x_2 + 2x_3 + 5x_4 + 8x_5 + 7x_6, \quad (11)$$

$$\text{s.t.} \quad A^+ \circ x \vee A^- \circ \neg x = b, \quad (12)$$

$$x \in [0, 1]^6,$$

where

$$A^+ = \begin{pmatrix} 0.4 & 0.1 & 0.25 & 0.29 & 0.18 & 0.07 \\ 0.32 & 0.16 & 0.23 & 0.1 & 0.2 & 0.48 \\ 0.08 & 0.05 & 0.02 & 0.1 & 0.03 & 0.06 \\ 0.15 & 0.3 & 0.17 & 0.2 & 0.14 & 0.05 \\ 0.09 & 0.2 & 0.03 & 0.04 & 0.15 & 0.24 \\ 0.49 & 0.58 & 0.6 & 0.37 & 0.75 & 0.54 \end{pmatrix},$$

$$A^- = \begin{pmatrix} 0.21 & 0.4 & 0.19 & 0.27 & 0.12 & 0.04 \\ 0.15 & 0.32 & 0.8 & 0.14 & 0.22 & 0.2 \\ 0.1 & 0.08 & 0.05 & 0.18 & 0.07 & 0.06 \\ 0.17 & 0.04 & 0.15 & 0.07 & 0.3 & 0.2 \\ 0.11 & 0.1 & 0.03 & 0.14 & 0.14 & 0.07 \\ 0.24 & 0.52 & 0.47 & 0.5 & 0.33 & 0.25 \end{pmatrix},$$

$b = (0.3, 0.24, 0.09, 0.18, 0.12, 0.6)^T$, and $x = (x_1, x_2, x_3, x_4, x_5, x_6)^T$. Now, we apply Algorithm 1 to solve the optimization problem of (11)–(12).

Step 1. The lower and upper bound of \tilde{x} and \hat{x} are as follows:

$\tilde{x} = (0.1, 0.25, 0.7, 0.5, 0.4, 0.1)^T$ and $\hat{x} = (0.75, 0.6, 1, 0.9, 0.8, 0.5)^T$.

Step 2. Since the conditions of Corollary 1 do not hold, we go to Step 3.

Step 3. In this example $\tilde{x}_j < \hat{x}_j$, for each $j \in J$ and $b_i > 0$, for each $i \in I$. Therefore, we go to Step 4.

Step 4. Applying Definition 1, the matrices of Q^+ and Q^- are obtained as follows:

$$Q^+ = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 & 5 & 6 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{matrix} & \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 \end{pmatrix} \end{matrix} \text{ and } Q^- = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 & 5 & 6 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{matrix} & \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{matrix}.$$

Also, the index sets of I_j^+ and I_j^- , for all $j \in J$ can be computed as follows:

$I_1^+ = \{1, 2\}$, $I_2^+ = \{4, 5\}$, $I_3^+ = \{6\}$, $I_4^+ = \{3, 4\}$, $I_5^+ = \{5, 6\}$, $I_6^+ = \{2, 5\}$, $I_1^- = \{3\}$, $I_2^- = \{1, 2\}$, $I_3^- = \{2\}$, $I_4^- = \{3\}$, $I_5^- = \{4\}$, and $I_6^- = \{4\}$.

Moreover, we can compute the index sets of J_i^+ and J_i^- , for all $i \in I$, as follows:

$J_1^+ = \{1\}$, $J_2^+ = \{1, 6\}$, $J_3^+ = \{4\}$, $J_4^+ = \{2, 4\}$, $J_5^+ = \{2, 5, 6\}$, $J_6^+ = \{3, 5\}$, $J_1^- = \{2\}$, $J_2^- = \{2, 3\}$, $J_3^- = \{1, 4\}$, $J_4^- = \{5, 6\}$, and $J_5^- = J_6^- = \emptyset$.

Step 5. The bipolar max- T_p FREs of $A^+ \circ x \vee A^- \circ \neg x = b$ is consistent according to Theorem 2. So, we go to Step 6.

Step 6. Perform the process of problem reduction as follows:

6.1. Two index sets I_1 and I_2 are as follows: $I_1 = \{1, 2, 3, 4\}$ and $I_2 = \{5, 6\}$.

Since the optimization problem of (11)–(12) cannot be reduced by Steps 6.2–6.4, we go to Step 6.5.

6.5. In this example, $T_1 = \{2\}$, that is, $J_1^+ \subseteq J_2^+$ and $J_1^- \subseteq J_2^-$. Applying Corollary 2, the second row of two matrices Q^+ and Q^- can be removed.

6.6. Since $q_{34}^+ = q_{34}^- = 1$, the third equation can be eliminated from our consideration with regard to Theorem 6.

The matrices of Q^+ and Q^- cannot be reduced further. So, we go to Step 7.

Step 7. Since $Q^+ \neq \emptyset$ and $Q^- \neq \emptyset$, we go to Step 8.

Step 8. With regard to Remark 3, the matrices of Q^+ and Q^- can be updated as follows:

$$Q^+ = \begin{matrix} & 1 & 2 & 3 & 4 & 5 & 6 \\ \begin{matrix} 5 \\ 6 \\ 1 \\ 4 \end{matrix} & \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \end{pmatrix} \end{matrix} \text{ and } Q^- = \begin{matrix} & 1 & 2 & 3 & 4 & 5 & 6 \\ \begin{matrix} 5 \\ 6 \\ 1 \\ 4 \end{matrix} & \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix} \end{matrix}$$

For the updated matrices of Q^+ and Q^- , the value matrix of M can be generated as follows:

$$M = \begin{matrix} & 1 & 2 & 3 & 4 & 5 & 6 \\ \begin{pmatrix} \infty & \infty & 1.05 & \infty & \infty & \infty & 3.2 & \infty & 2.8 & \infty \\ \infty & \infty & \infty & \infty & 0.6 & \infty & \infty & \infty & 3.2 & \infty & \infty & \infty \\ 0.65 & \infty & \infty & 0 & \infty & \infty & \infty & \infty & \infty & \infty & \infty & \infty \\ \infty & \infty & 1.05 & \infty & \infty & \infty & 2 & \infty & \infty & 0 & \infty & 0 \end{pmatrix} \end{matrix}.$$

Step 9. We are now ready to use the modified branch-and-bound method with the jump-tracking technique on the matrix M . We begin with the first equation, that is, $i = 1$. The set of $\{\hat{x}_2, \hat{x}_5, \hat{x}_6\}$ introduces three candidates to satisfy the first equation. Therefore, we have to branch from Node 0 in Figure 2. If we select \hat{x}_2 (Node 1), then the value of Z_1 is 1.05. Note that \hat{x}_2 cannot be used for further branching on Node 1. Also, if we select \hat{x}_5 (\hat{x}_6), then we have $Z_2 = 3.2$ ($Z_3 = 2.8$). Furthermore, we never use \hat{x}_5 (\hat{x}_6) to branch further on Node 2 (Node 3). Here, Node 1 is selected to branch further because of the least objective value.

Move to the second row of the matrix M . Since the set of $\{\hat{x}_3, \hat{x}_5\}$ contains two candidates to satisfy the second equation, we have two branches from Node 1 as it has been illustrated in Figure 2. If \hat{x}_3 (Node 4) is selected, then the value of Z_4 is 1.65. If \hat{x}_5 (Node 5) is considered, then we have $Z_5 = 4.25$. Now, we can branch further on four Nodes 2, 3, 4, and 5 but Node 4 is chosen with regard to the least objective value.

Move to the third row of the matrix M . The set of $\{\hat{x}_1, \hat{x}_2\}$ contains two candidates to satisfy the third equation, but \hat{x}_2 cannot be used to branch further on Node 4 because \hat{x}_2 has been chosen along Node 0 to Node 4 (modification 1). Therefore, the set of $\{\hat{x}_1\}$ contains the only candidate to

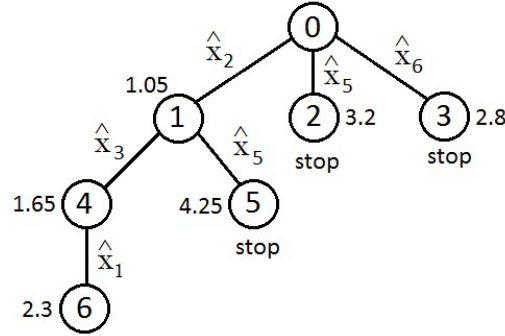


Figure 2: The modified branch-and-bound method

satisfy the third equation here. If \hat{x}_1 (Node 6) is selected, then it is concluded that $Z_6 = 2.3$.

Since \hat{x}_1 , \hat{x}_2 , and \hat{x}_3 together with \check{x}_4 , \check{x}_5 , and \check{x}_6 satisfy all the equations, we do not branch further on Node 6 with regard to modification 2.2. Therefore, considering modification 3, the vector $x = (\hat{x}_1, \hat{x}_2, \hat{x}_3, \check{x}_4, \check{x}_5, \check{x}_6)^T$ is a solution with the objective value of 2.3 for the equivalent problem. Since the value of Z_6 is less than Z_2 , Z_3 , and Z_5 , we can stop further branching on all Nodes 2, 3, and 5. Thus, the total of Z_6 can be computed as follows:

$$\text{Total } Z_6 = Z_6 + \sum_{j=1}^6 c_j \check{x}_j = 2.3 + 8.65 = 10.95. \text{ Every node is stopped further}$$

branching. Hence, the optimal solution can be obtained from Node 6, that is, $x_1^* = 0.75$, $x_2^* = 0.6$, $x_3^* = 1$, $x_4^* = 0.5$, $x_5^* = 0.4$, and $x_6^* = 0.1$ with the total of $Z_6 = 10.95$.

Step 10. The optimal objective value is 10.95 and the optimal solution is $x^* = (x_1^*, x_2^*, x_3^*, x_4^*, x_5^*, x_6^*)^T = (0.75, 0.6, 1, 0.5, 0.4, 0.1)^T$.

If we do not apply Remark 3 to solve this example, we need to consider 25 nodes. Considering Remark 3, we need only six nodes to solve this example.

Example 2. Consider the following optimization problem:

$$\min \quad 2x_1 + 5x_2 + 3x_3 + 4x_4 + x_5 + 6x_6, \quad (13)$$

$$\text{s.t.} \quad A^+ \circ x \vee A^- \circ \neg x = b, \quad (14)$$

$$x \in [0, 1]^6,$$

where

$$A^+ = \begin{pmatrix} 0.7 & 0.3 & 0.96 & 0.66 & 0.8 & 0.05 \\ 0.31 & 0.54 & 0.34 & 0.3 & 0.36 & 0.45 \\ 0.06 & 1 & 0.04 & 0.625 & 0.03 & 0.08 \\ 0.15 & 0.15 & 0.19 & 0.225 & 0.16 & 0.03 \\ 0.6 & 0.08 & 0.07 & 0.06 & 0.13 & 0.4 \\ 0.3 & 0.05 & 0.09 & 0.07 & 0.11 & 0.19 \end{pmatrix},$$

$$A^- = \begin{pmatrix} 0.11 & 0.4 & 0.19 & 0.27 & 0.12 & 0.04 \\ 0.14 & 0.24 & 0.09 & 0.14 & 0.22 & 0.2 \\ 0.1 & 0.08 & 0.05 & 0.18 & 0.8 & 0.8 \\ 0.2 & 0.04 & 0.15 & 0.3 & 0.11 & 0.05 \\ 0.11 & 0.4 & 0.32 & 0.14 & 0.14 & 0.07 \\ 0.03 & 0.2 & 0.06 & 0.02 & 0.04 & 0.07 \end{pmatrix},$$

$b = (0.6, 0.27, 0.5, 0.18, 0.24, 0.12)^T$, and $x = (x_1, x_2, x_3, x_4, x_5, x_6)^T$. Now, we apply Algorithm 1 to solve the optimization problem of (13)–(14).

Step 1. The lower and upper bound of \tilde{x} and \hat{x} are as follows:

$\tilde{x} = (0.1, 0.4, 0.25, 0.4, 0.375, 0.375)^T$ and $\hat{x} = (0.4, 0.5, 0.625, 0.8, 0.75, 0.6)^T$.

Step 2. Since the conditions of Corollary 1 do not hold, we go to Step 3.

Step 3. In this example, $\tilde{x}_j < \hat{x}_j$, for each $j \in J$ and $b_i > 0$, for each $i \in I$. Therefore, we go to Step 4.

Step 4. Applying Definition 1, the matrices of Q^+ and Q^- are obtained as follows:

$$Q^+ = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 & 5 & 6 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{matrix} & \begin{pmatrix} 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{matrix} \text{ and } Q^- = \begin{matrix} & \begin{matrix} 1 & 2 & 3 & 4 & 5 & 6 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{matrix} & \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix} \end{matrix}.$$

Also, the index sets of I_j^+ and I_j^- , for all $j \in J$ can be computed as follows:

$I_1^+ = \{5, 6\}$, $I_2^+ = \{2, 3\}$, $I_3^+ = \{1\}$, $I_4^+ = \{3, 4\}$, $I_5^+ = \{1, 2\}$, $I_6^+ = \{2, 5\}$, $I_1^- = \{4\}$, $I_2^- = \{5, 6\}$, $I_3^- = \{5\}$, $I_4^- = \{4\}$, $I_5^- = \{3\}$, and $I_6^- = \{3\}$.

Moreover, we can compute the index sets of J_i^+ and J_i^- , for all $i \in I$, as follows:

$J_1^+ = \{3, 5\}$, $J_2^+ = \{2, 5, 6\}$, $J_3^+ = \{2, 4\}$, $J_4^+ = \{4\}$, $J_5^+ = \{1, 6\}$, $J_6^+ = \{1\}$, $J_1^- = J_2^- = \emptyset$, $J_3^- = \{5, 6\}$, $J_4^- = \{1, 4\}$, $J_5^- = \{2, 3\}$, and $J_6^- = \{2\}$.

Step 5. The bipolar max- T_p FREs of $A^+ \circ x \vee A^- \circ \neg x = b$ is consistent according to Theorem 2. So, we go to Step 6.

Step 6. Perform the process of problem reduction as follows:

6.1. Two index sets I_1 and I_2 are as follows: $I_1 = \{3, 4, 5, 6\}$ and $I_2 = \{1, 2\}$.

6.2. The condition of Substep 2 is not satisfied for this problem.

6.3. The conditions of this substep hold for this problem. Since for each $j \in J - \bigcup_{i \in I_2} = \{1, \dots, 6\} - (J_1^+ \cup J_2^+) = \{1, 4\}$, we have $c_1 = 2 > 0$ and $c_4 = 4 > 0$. Also, for $k \in \bigcap_{i \in I_2} = \{5\}$, the following conditions are satisfied:

1. for all $j \in \bigcup_{i \in I_2} J_i^+ - \{5\} = \{2, 3, 6\}$, we have $c_5(\hat{x}_5 - \check{x}_5) = 0.225 < c_j(\hat{x}_j - \check{x}_j)$.
2. $I_5^- - I_5^+ = \{3\} \subseteq \bigcup_{j \in J, j \neq 5} I_j^- = \{3, 4, 5, 6\}$.

Therefore, according to Lemma 10, the unique optimal solution of the problem is as follows: $x^* = (0.1, 0.4, 0.25, 0.4, 0.75, 0.375)$ with the optimal objective value $Z^* = 7.55$

6 Comparison of the proposed algorithm with the methods in other papers

In this section, we compare the proposed algorithm with the methods in other papers [9, 20, 23, 3] to solve problem (1)–(2) with regard to the obtained results from Examples 1 and 2 in Section 5.

As it is seen in Figure 2, Algorithm 1 solves the problem of Example 1 only in six nodes by the branch-and-bound method. Also, Algorithm 1 solves Example 2 without using the branch-and-bound method. Its optimal solution is found in Substep 6.3 by Lemma 10.

Freson, De Baets, and De Meyer [9] discussed problem (1)–(2) with the max-min composition operator. They designed an algorithm to solve the problem. Its Step 1 checks the necessary condition (46) in [9]. Its computational cost is $O(mn)$. Step 2 constructs vectors g^+ and s^- taking supremum and infimum from m maximal solutions which its computational cost is $O(mn)$. Step 3 generates all elements of set B and keeps those that satisfy constraints (30) in [9]. To generate all the elements of B according to relation (48) in [9], we firstly need to produce the set $(\{0, 1\} \cup \{b_i, \bar{b}_i | i = 1, \dots, m\})^n$. Its computational cost is $O((2m+2)^n)$. Then we must check whether each its element belongs to $[s^-, g^+]$ or not? Its computational cost is $O(n(2m+2)^n)$. Now, we must check whether each element B satisfies constraints (30) in [9] or not? The computational cost of this work is $O(mn)$ for each element of the set B . If we check all the elements of set B , then its computational cost is $O(mn(2m+2)^n)$. So, the computational cost of Step 3 is $T_3 = O(n(2m+2)^n + mn(2m+2)^n) = O(mn(2m+2)^n)$. Step 4 selects the elements in $B \cap D$ with the highest value for the objective function. To do this, we must check the objective function for $|B \cap D|$ elements which its computational cost is $O(|B \cap D|) \leq O((2m+2)^n)$. Therefore, the computational complexity of the given algorithm in [9] is $T_F = O(mn(2m+2)^n)$. For an instance of the problem with the dimensions $m = n = 6$ like Examples 1 and 2, its computational complexity is $T_F = O(6 \times 6 \times (2 \times 6 + 2)^6) = 271063296 \times O(1)$. This point implies that $(2 \times 6 + 2)^6 = 7529536$ elements are produced and

its feasibility is checked in m equations of n -variable. Then the optimizer is found by computing the objective function values in the feasible vectors.

In [20], problem (1)–(2) is discussed with the max-Lukasiewicz t-norm composition. Li and Liu [20] directly converted the problem to a 0-1 integer linear optimization problem without its simplification and reduction. If we apply their method to solve the problem of Example 1, we should consider the following 0-1 integer programming problem:

$$\begin{aligned} Z = \min & \quad 8.65 + 0.65u_1 + 1.05u_2 + 0.6u_3 + 2u_4 + 3.2u_5 + 2.8u_6, \\ \text{s.t.} & \quad \begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ 1 & -1 & -1 & 0 & 0 & 1 \\ -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & -1 & -1 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \end{pmatrix} \geq \begin{pmatrix} 0 \\ -1 \\ -1 \\ -1 \\ 1 \\ 1 \end{pmatrix}, \\ & \quad u \in \{0, 1\}^6. \end{aligned}$$

If we apply Algorithm 1 of this paper for Example 1, we should solve the following problem with smaller dimensions:

$$\begin{aligned} Z = \min & \quad 8.65 + 0.65u_1 + 1.05u_2 + 0.6u_3 + 2u_4 + 3.2u_5 + 2.8u_6, \\ \text{s.t.} & \quad \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & -1 & -1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \end{pmatrix} \geq \begin{pmatrix} 1 \\ 1 \\ 0 \\ -1 \end{pmatrix}, \\ & \quad u \in \{0, 1\}^6. \end{aligned}$$

If we apply Li and Liu's method to solve the problem of Example 2, we should consider the following 0-1 integer programming problem:

$$\begin{aligned} Z = \min & \quad 7.175 + 0.6u_1 + 0.5u_2 + 1.125u_3 + 1.6u_4 + 0.375u_5 + 1.35u_6, \\ \text{s.t.} & \quad \begin{pmatrix} 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & -1 & -1 \\ -1 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & -1 & 0 & 0 & 1 \\ 1 & -1 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \end{pmatrix} \geq \begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \\ -1 \\ 0 \end{pmatrix}, \\ & \quad u \in \{0, 1\}^6. \end{aligned}$$

If we apply Algorithm 1 of this paper for Example 2, we directly obtain the unique optimal solution of the problem of Example 2 without using the branch-and-bound method and 0-1 integer programming problem.

In [3], problem (1)–(2) was discussed with the max-Hamacher t -norm composition with some rules to simplify the problem. The rules are completely different from the proposed procedures of this paper for simplification. The rules in [3] are not applicable for the problem in Example 1. Hence, if we use the given algorithm in [3] for Example 1, the branch-and-bound method should be applied on the matrix M as follows:

$$M = \begin{pmatrix} & 1 & 2 & 3 & 4 & 5 & 6 \\ 0.65 & \infty & 0 & \infty & \infty & \infty & \infty & \infty \\ 0.65 & \infty & \infty & 0 & \infty & \infty & \infty & 2.8 \\ \infty & 0 & \infty & \infty & \infty & 2 & 0 & \infty \\ \infty & \infty & 1.05 & \infty & \infty & 2 & \infty & \infty \\ \infty & \infty & 1.05 & \infty & \infty & \infty & 3.2 & \infty \\ \infty & \infty & \infty & \infty & 0.6 & \infty & 3.2 & \infty \end{pmatrix}. \quad (15)$$

If we use Algorithm 1 for Example 1, the branch-and-bound method should be applied on the matrix M as follows:

$$M = \begin{pmatrix} & 1 & 2 & 3 & 4 & 5 & 6 \\ \infty & \infty & 1.05 & \infty & \infty & \infty & 3.2 & \infty \\ \infty & \infty & \infty & \infty & 0.6 & \infty & \infty & \infty \\ 0.65 & \infty & \infty & 0 & \infty & \infty & \infty & \infty \\ \infty & \infty & 1.05 & \infty & \infty & \infty & 2 & \infty \end{pmatrix}. \quad (16)$$

Also, the rules in [3] are not applicable for the problem in Example 2. Hence, if we use the given algorithm in [3] for Example 2, the branch-and-bound method should be applied on the matrix M as follows:

$$M = \begin{pmatrix} & 1 & 2 & 3 & 4 & 5 & 6 \\ \infty & \infty & \infty & \infty & 1.125 & \infty & \infty & \infty \\ \infty & \infty & 0.5 & \infty & \infty & \infty & 0.375 & \infty \\ \infty & \infty & 0.5 & \infty & \infty & 1.6 & \infty & 0 \\ \infty & 0 & \infty & \infty & \infty & 1.6 & 0 & \infty \\ 0.6 & \infty & \infty & 0 & \infty & \infty & \infty & 1.35 \\ 0.6 & \infty & \infty & 0 & \infty & \infty & \infty & \infty \end{pmatrix}. \quad (17)$$

If we use Algorithm 1 for Example 2, its optimal solution is directly found in Substep 6.3 without using the branch-and-bound method and $M = \emptyset$.

In [23], problem (1)–(2) was discussed with the max-Lukasiewicz composition with some rules to simplify the problem. The rules are completely different from the proposed procedures of this paper for simplification. If we apply the method in [23] for the problem of Example 1 with the max-product composition, rule 3 in [23] can be used for its simplification. Applying the equivalent form of the rule 3 for problem (1)–(2), the forth row of the matrix M in the relation (15) is removed and the branch-and-bound method should be employed on the 0-1 integer programming equivalent to the following value

matrix:

$$M = \begin{pmatrix} & 1 & 2 & 3 & 4 & 5 & 6 \\ 0.65 & \infty & \infty & 0 & \infty & \infty & \infty & \infty & \infty \\ 0.65 & \infty & \infty & 0 & \infty & 0 & \infty & \infty & 2.8 & \infty \\ \infty & 0 & \infty & \infty & \infty & \infty & 2 & 0 & \infty & \infty & \infty \\ \infty & \infty & 1.05 & \infty & \infty & \infty & \infty & \infty & 3.2 & \infty & 2.8 & \infty \\ \infty & \infty & \infty & \infty & 0.6 & \infty & \infty & \infty & 3.2 & \infty & \infty & \infty \end{pmatrix}. \quad (18)$$

If we use Algorithm 1 for Example 1, the branch-and-bound method should be applied on the matrix M in the relation (16) and considering Remark 3, we need only six nodes to solve this example. If we apply the method in [23] for the problem of Example 2 with the max-product composition, rule 3 in [23] can be used for its simplification. Applying the equivalent form of rule 3 in [23], the third row of the matrix M in relation (17) is removed and the branch-and-bound method should be employed on the 0-1 integer programming equivalent to the following value matrix:

$$M = \begin{pmatrix} & 1 & 2 & 3 & 4 & 5 & 6 \\ \infty & \infty & \infty & \infty & 1.125 & \infty & \infty & \infty & 0.375 & \infty & \infty & \infty \\ \infty & \infty & 0.5 & \infty & \infty & \infty & \infty & \infty & 0.375 & \infty & 1.35 & \infty \\ \infty & 0 & \infty & \infty & \infty & \infty & 1.6 & 0 & \infty & \infty & \infty & \infty \\ 0.6 & \infty & \infty & 0 & \infty & 0 & \infty & \infty & \infty & \infty & 1.35 & \infty \\ 0.6 & \infty & \infty & 0 & \infty & \infty & \infty & \infty & \infty & \infty & \infty & \infty \end{pmatrix}. \quad (19)$$

If we use Algorithm 1 for Example 2, its optimal solution is directly found in Substep 6.3 without using the branch-and-bound method and $M = \emptyset$.

The other preferences of the proposed algorithm with respect to the presented algorithms in [9, 20, 23, 3] for the resolution of problem (1)–(2) are as follows. The proposed algorithm introduces two new classes of problem (1)–(2) with the max-product composition operator, which can directly be solved only by Substep 6.3 or Substep 6.4. The optimal solution of these classes of the problem satisfying conditions of Substep 6.3 or Substep 6.4 can be obtained by the relation (4) without applying the branch-and-bound method or using rules repeatedly. The classes have not been introduced in [9, 20, 23, 3]. Other proposed rules in Step 6 of Algorithm 1 are different from the given rules in [23, 3]. In [9, 20], the authors have not used the rules of simplification to reduce the original problem. Since the algorithms in [20, 23, 3] and the proposed algorithm are based on the branch-and-bound algorithm, the algorithms are convergent. The algorithm in [9] checks all the possible feasible solutions to find the optimal solution.

7 Conclusions and future works

The linear optimization problem with the bipolar max-product FREs was studied in this paper. The characterizations of its feasible domain were investigated. Some simplification operations were proposed to delete some equations. With regard to these operations, the size of the original problem was reduced. Then, some sufficient conditions were presented to determine one of the optimal solutions to the problem and its uniqueness. Moreover, a value matrix was defined based on the characteristic matrices of the feasible domain of the problem. Then, the branch-and-bound method was modified to solve the reduced problem with regard to the value matrix. An algorithm was finally designed to solve the problem and compared with other methods to show the efficiency of the proposed algorithm. In future work, the linear optimization problem will be developed by supposing fuzzy linear systems with bipolar fuzzy numbers based on references [10, 11].

Acknowledgements

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

References

1. Aliannezhadi, S. and Abbasi Molai, A. *Geometric programming with a single-term exponent subject to bipolar max-product fuzzy relation equation constraints*, Fuzzy Sets Syst. 397 (2020), 61–83.
2. Aliannezhadi, S. and Abbasi Molai, A. *A new algorithm for geometric optimization with a single-term exponent constrained by bipolar fuzzy relation equations*, Iran J. Fuzzy Syst. 18(1) (2021), 137–150.
3. Aliannezhadi, S., Abbasi Molai, A. and Hedayatfar, B. *Linear optimization with bipolar max-parametric Hamacher fuzzy relation equation constraints*, Kybernetika, 52(4) (2016), 531–557.
4. Chiu, Y.-L., Guu, S.-M., Yu, J. and Wu, Y.-K. *A single-variable method for solving min-max programming problem with addition-min fuzzy relational inequalities*, Fuzzy Optim. Decis. Ma., 18 (2019), 433–449.
5. Cornejo, M.E., Lobo, D. and Medina, J. *On the solvability of bipolar max-product fuzzy relation equations with the standard negation*, Fuzzy Sets Syst. 410 (2021), 1–18.

6. Cornejo, M.E., Lobo, D. and Medina, J. *On the solvability of bipolar max-product fuzzy relation equations with the product negation*, J. Comput. Appl. Math. 354 (2019), 520–532.
7. De Baets, B. *Analytical solution methods for fuzzy relational equations*, in: D. Dubois, H. Prade (Eds.), *Fundamentals of Fuzzy Sets*, The Handbooks of Fuzzy Sets Series. Kluwer Academic Publishers, Dordrecht, 291–340 (2000).
8. Fang, S.-C. and Li, G. *Solving fuzzy relation equations with a linear objective function*, Fuzzy Sets Syst. 103 (1999), 107–113 .
9. Freson, S., De Baets, B. and De Meyer, H. *Linear optimization with bipolar max-min constraints*, Inf. Sci. 234 (2013), 3–15.
10. Ghanbari, R., Ghorbani-Moghadam, Kh. and Mahdavi-Amiri, N. *Duality in bipolar fuzzy number linear programming problem*, Fuzzy Inf. Eng. 11 (2019) 175–185.
11. Ghanbari, R., Ghorbani-Moghadam, Kh. and Mahdavi-Amiri, N. *Duality in bipolar triangular fuzzy number quadratic programming problems*, 2017 International Conference on Intelligent Sustainable Systems (ICISS), 1236–1238 (2017) .
12. Guo, F.-F. and Shen, J. *A smoothing approach for minimizing a linear function subject to fuzzy relation inequalities with addition-min composition*, Int. J. Fuzzy Syst. 21(1) (2019), 281–290.
13. Guu, S.-M. and Wu, Y.-K. *Multiple objective optimization for systems with addition-min fuzzy relational inequalities*, Fuzzy Optim. Decis. Ma. 18 (2019), 529–544.
14. Guu, S.-M. and Wu, Y.-K. *Minimizing a linear objective function with fuzzy relation equation constraints*, Fuzzy Optim. Decis. Ma. 1 (2002), 347–360.
15. Hedayatfar, B., Abbasi Molai, A. and Aliannezhadi, S. *Separable programming problems with the max-product fuzzy relation equation constraints*, Iran J. Fuzzy Syst. 16(1) (2019), 1–15.
16. Li, M. and Wang, X.-P. *Remarks on minimal solutions of fuzzy relation inequalities with addition-min composition*, Fuzzy Sets Syst. 410 (2021), 19–26.
17. Li, P. and Fang, S.-C. *On the resolution and optimization of a system of fuzzy relational equations with sup-T composition*, Fuzzy Optim. Decis. Ma. 7 (2008), 169–214.
18. Li, P. and Jin, Q. *Fuzzy relational equations with min-biimplication composition*, Fuzzy Optim. Decis. Ma. 11 (2012), 227–240.

19. Li, P. and Jin, Q. *On the resolution of bipolar max-min equations*, Kybernetika, 52 (2016) 514–530.
20. Li, P. and Liu, Y. *Linear optimization with bipolar fuzzy relational equation constraints using the Lukasiewicz triangular norm*, Soft Comput. 18 (2014), 1399–1404.
21. Lichun, C. and Boxing, P. *The fuzzy relation equation with union or intersection preserving operator*, Fuzzy Sets Syst. 25 (1988), 191–204.
22. Lin, H. and Yang, X.-P. *Dichotomy algorithm for solving weighted min-max programming problem with addition-min fuzzy relation inequalities constraint*, Comput. Ind. Eng. 146 (2020), 106537.
23. Liu, C.-C., Lur, Y.-Y. and Wu, Y.-K. *Linear optimization of bipolar fuzzy relational equations with max-Lukasiewicz composition*, Inf. Sci. 360 (2016), 149–162.
24. Loetamonphong, J. and Fang, S.-C. *Optimization of fuzzy relation equations with max-product composition*, Fuzzy Sets Syst. 118 (2001), 509–517.
25. Luoh, L., Wang, W.-J. and Liaw, Y.-K. *New algorithms for solving fuzzy relation equations*, Math. Comput. Simul. 59 (2002), 329–333.
26. Peeva, K. *Composite fuzzy relational equations in decision making: chemistry*, In: B. Cheshankov, M. Todorov (eds) Proceedings of the 26th summer school applications of mathematics in engineering and economics, Sozopol 2000. Heron press, 260–264 (2001).
27. Peeva, K. *Universal algorithm for solving fuzzy relational equations*, Ital. j. pure appl. math. 19 (2006) 169–188.
28. Peeva, K. and Kyosev, Y. *Fuzzy relational calculus: theory, applications and software*, World Scientific, New Jersey (2004).
29. Peeva, K., Zahariev, ZL. and Atanasov, IV. *Optimization of linear objective function under max-product fuzzy relational constraint*, In: 9th WSEAS international conference on FUZZY SYSTEMS (FS'08) Sofia, Bulgaria, 132–137 (2008).
30. Peeva, K., Zahariev, ZL. and Atanasov, IV. *Software for optimization of linear objective function with fuzzy relational constraint*, In: Fourth international IEEE conference on intelligent systems, Verna (2008).
31. Sanchez, E. *Resolution of composite fuzzy relation equations*, Inf. Control. 30 (1976), 38–48 .
32. Vasantha Kandasamy, W.B. and Smarandache, F. *Fuzzy relational maps and neutrosophic relational maps*, hexis church rock (see chapters one and two) <http://mat.iitm.ac.in/~wbv/book13.htm> (2004).

33. Wu, Y.-K. and Guu, S.-M. *A note on fuzzy relation programming problems with max-strict-t-norm composition*, Fuzzy Optim. Decis. Ma. 3(3) (2004), 271–278.
34. Wu, Y.-K. and Guu, S.-M. *Minimizing a linear function under a fuzzy max-min relational equation constraint*, Fuzzy Sets Syst. 150, 147–162 (2005).
35. Wu, Y.-K., Guu, S.-M. and Liu, J.Y.-C. *An accelerated approach for solving fuzzy relation equations with a linear objective function*, IEEE Trans. Fuzzy Syst. 10(4) (2002), 552–558.
36. Yang, X., Qiu, J., Guo, H. and Yang, X.-P. *Fuzzy relation weighted min-max programming with addition-min composition*, Comput. Ind. Eng. 147 (2020), 106644.
37. Yang, X.-P. *Resolution of bipolar fuzzy relation equations with max-Lukasiewicz composition*, Fuzzy Sets Syst. 397 (2020), 41–60.
38. Yang, X.-P. *Solutions and strong solutions of min-product fuzzy relation inequalities with application in supply chain*, Fuzzy Sets Syst. 384 (2020), 54–74.
39. Yeh, C.-T. *On the minimal solutions of max-min fuzzy relational equations*, Fuzzy Sets Syst. 159 (2008), 23–39.
40. Zhong, Y.-B., Xiao, G. and Yang, X.-P. *Fuzzy relation lexicographic programming for modelling P2P file sharing system*, Soft Comput. 23 (2019), 3605–3614.
41. Zhou, J., Yu, Y., Liu, Y. and Zhang, Y. *Solving nonlinear optimization problems with bipolar fuzzy relational equation constraints*, J. Inequal. Appl. 126 (2016), 1–10.



Review of the strain-based formulation for analysis of plane structures

Part I: Formulation of basics and the existing elements

M. Rezaiee-Pajand*, N. Gharaei-Moghaddam and M. Ramezani

Abstract

Since the introduction of the finite element approach, as a numerical solution scheme for structural and solid mechanics applications, various formulation methodologies have been proposed. These ways offer different advantages and shortcomings. Among these techniques, the standard displacement-based approach has attracted more interest due to its straightforward scheme and generality. Investigators have proved that the other strategies, such as the force-based, hybrid, assumed stress, and assumed strain provides special advantages in comparison with the classic finite elements. For instance, the mentioned techniques are able to solve difficulties, like shear locking, shear parasitic error, mesh sensitivity, poor convergence, and rotational dependency. The main goal of this two-part study is to present a brief yet clear portrait of the basics and advantages of the direct strain-based method for development of high-performance plane finite elements. In this article, which is the first part of this study, assumptions and the basics of this method are introduced. Then, a detailed review of all the existing strain-based membrane elements is presented. Although the strain formulation is applicable for different types of structures, most of the existing elements pertain to the plane structures. The second part of this study deals with the application and performance of the reviewed elements in the analysis of plane stress/strain problems.

AMS subject classifications (2020): 74K15, 74G15.

*Corresponding author

Received 25 October 2020; revised 8 June 2021; accepted 9 June 2021

Mohammad Rezaiee-Pajand

Professor of Civil Engineering, School of Engineering, Ferdowsi University of Mashhad, Iran. e-mail: Rezaiee@um.ac.ir, Tel/fax: +98-51-38412912

Nima Gharaei-Moghaddam

PhD of Structural Engineering, School of Engineering, Ferdowsi University of Mashhad, Iran. e-mail: Nima.Gharaei@gmail.com, Tel: +98-915-1589342

Mohammadreza Ramezani

PhD Student of Structural Engineering, School of Engineering, Ferdowsi University of Mashhad, Iran. e-mail: Mohammadrezaramezani1994@gmail.com, Tel: +98-915-1076010

Keywords: Strain-based formulation; Higher-order strain field; Equilibrium condition; Numerical evaluation; Drilling degrees of freedom.

1 Introduction

Numerical methods are proved to be powerful and effective computational tools for analysis of complicated and practical engineering problems. According to different features of scientific activities and in demand of special requirements, many diverse numerical techniques are developed in the past decades, such as the finite element method, finite difference technique, boundary element method, and discrete element approach. Each of these various numerical methods have their own advantages and shortcomings, but the finite element methods gain more popularity due to their strong mathematical bases and inherent capabilities, which result in increasing application of this scheme in different fields of science and especially engineering fields [5, 6, 28, 39, 40, 42, 43, 78, 79]. Therefore, various formulation techniques are developed in the past decades and there are thousands of finite elements available for analysis of dissimilar types of problems and structures [8, 13, 21, 22, 41, 53].

Among the available approaches for finite element formulation, the most well-known and widely applicable one is the displacement-based technique. This method, which sometimes is called with different terms, such as, the classical or stiffness approach, is the first scheme that was used for the development of finite elements; see [80]. Clear and straightforward process and applicability to different types of problems and structures are the prominent advantages of the displacement-based formulation for structural and mechanical applications. However, this process has various shortcomings. For instance, inaccuracy and discontinuity of stresses, which are secondary parameters in stiffness approach, are a vital deficiency in structural applications, where stress is a decisive parameter in the design practice; see [33, 52]. It should be noted that this problem can partly be solved by using higher order formulations, which leads to the application of internal nodes, and increases computational costs, and reduces the numerical efficiency of the analysis [23, 36, 54, 55]. Another common problem of displacement-based finite elements in various locking phenomena, such as, shear and membrane locking, which necessitate special treating, which sometimes requires considerable time and effort and reduces the efficiency of the method [54, 55]. Moreover, in severely nonlinear problems, the displacement-based elements usually necessitate utilization of very fine meshes, which is inappropriate from the efficiency standpoint [46, 47, 74]. To remedy the mentioned and other shortcomings of the displacement approach, other finite element formulations, such as the force-based, hybrid or mixed, assumed stress, and assumed strain has been developed. Fortunately, these new procedures have

their own advantages and shortcomings. For example, the force-based formulation performs very well in the linear and nonlinear analysis of frame structures and also provides an appropriate platform for the development of advanced frame elements [3, 34, 56, 57, 73, 76]. It is reminded that the force formulation approach is limited to skeletal structures and that its application for continuous structures is very difficult if not impossible.

Although each method has different merits and limitations, some of the approaches have received more attention from researchers, while the others have remained less treated. One of these techniques that has received less attention despite its promising performance, is the strain-based or assumed strain approach [17, 25, 35, 38]. Therefore, the main purpose of this study is to introduce basics of the strain-based formulation. The strain formulation method can be classified in three distinct groups, namely free formulation, assumed natural strain method, and direct approach.

The free formulation method is based on the kinematic decomposition, in which the element displacements are decomposed in two basic and higher-order parts [25, 38]. The basic part represents the rigid body motions and constant strain state, which is necessary for the convergence criteria. On the other hand, the higher-order terms improve accuracy of the finite element by establishing proper rank of the stiffness matrix. The conditions of force and energy orthogonality result in the algebraic formulation of the stiffness matrix, which satisfies the individual element test. It is noteworthy that the strain functions in this method are dependent on displacements one; see [19]. The second method for the development of strain-based elements is the assumed natural deviatoric strain (ANDES) approach [27]. In this technique, the independent strain function is applied, which includes basic and higher-order parts. The deviation of strain from the constant strain is represented by the higher-order part of the strain function. In this way, the higher-order part is selected so that the integral of higher-order strain through the element equals zero. Therefore, *ANDES* technique satisfies the individual element test [27].

The third method is the direct formulation approach, in which Taylor series for the strain field is used to approximate the strain field [63]. The resulting elements from this approach lead free from the shear locking and parasitic shear error. The strain states in this approach can be divided into rigid body motions, constant strain, and higher-order strain states. The rigid body modes and constant strain states guarantee the convergence of resulting element. The higher-order terms have a parametric form that can be optimized to obtain efficient elements. The optimization is performed by enforcing different optimal conditions. Some of these optimal conditions will be discussed in the coming sections. It is important to note that the optimization process of the finite element templates is a relatively complicated task that requires innovation [63].

As mentioned, the strain-based formulation itself can be categorized in three different subdivisions. In this study, only the direct method is covered.

After the presentation of the formulation basics, a detailed review of the existing assumed-strain membrane elements is presented. This article is the first part of a two-part study. In the second part, numerical comparison of the reviewed elements is performed. This article is organized in the following order: Section 2 presents basis steps for the development of the plane elements with the assumed strain approach. Other important optimal criteria for assume strain elements are introduced in Section 3. The existing triangular membrane elements are reviewed in Section 4, and Section 5 discusses the available quadrilateral plane elements. In Section 6, some of the accessible strain-based elements for the other types of structures are briefly introduced. Finally, Section 7 presents the concluding remarks of the article.

2 Basics of the formulation

In the assumed strain formulation, the element strain field is approximated by an assumed mathematical function. As mentioned, like any other form of the finite element formulation, there are different types of assumed strain formulation [4, 20, 26, 37]. However, in this study, only the direct strain-based formulation will be discussed [63]. The following formulation steps pertain to development of plane finite elements. Needless to say, the presented process can be simply applied to the other types of elements with only slight modifications. The main reasons behind narrowing the scope of the present review are the availability of the relatively large number of existing strain-based elements developed based on the various types of strain-based formulation. These researches made the current study very lengthy. Moreover, the growing interest from the structural analysis community toward the direct method in recent years, and also the valuable experiences of the authors in this field, which should be exposed to the younger analysts.

In the case of plane problems, the strain field consists of three components, namely ε_x , ε_y , and γ_{xy} . Based on the Taylor expansion, each function can be approximated by a polynomial function of arbitrary order. Therefore, the strain components are approximated as follows:

$$\begin{cases} \varepsilon_x(x, y) = (\varepsilon_x)_o + (\varepsilon_{x,x})_o x + (\varepsilon_{x,y})_o y + (\varepsilon_{x,xx})_o \left(\frac{x^2}{2}\right) + (\varepsilon_{x,xy})_o (xy) + (\varepsilon_{x,yy})_o \left(\frac{y^2}{2}\right) + \dots, \\ \varepsilon_y(x, y) = (\varepsilon_y)_o + (\varepsilon_{y,x})_o x + (\varepsilon_{y,y})_o y + (\varepsilon_{y,xx})_o \left(\frac{x^2}{2}\right) + (\varepsilon_{y,xy})_o (xy) + (\varepsilon_{y,yy})_o \left(\frac{y^2}{2}\right) + \dots, \\ \gamma_{xy}(x, y) = (\gamma_{xy})_o + (\gamma_{xy,x})_o x + (\gamma_{xy,y})_o y + (\gamma_{xy,xx})_o \left(\frac{x^2}{2}\right) + (\gamma_{xy,xy})_o (xy) + (\gamma_{xy,yy})_o \left(\frac{y^2}{2}\right) + \dots \end{cases} \quad (1)$$

Here “,” indicates differentiating with respect to its following variable. Moreover, the subscript “o” indicates the value of the strain gradient at the origin of the coordinate system. The coefficients with the subscript “o” are called strain states. Selection of diverse polynomials with a different number of terms and various orders results in finite elements with a different number of degrees of freedom, as well as, the specific properties. Due to the importance of the constant strain states for the convergence of the resulting finite

element, the selection of the constant and linear terms for strain components is necessary. Selection of the higher-order terms would increase the accuracy and the convergence rate of the resulting finite element, but instead reduces its numerical efficiency. As it was mentioned previously, it is possible to apply different optimal criteria, such as pure plain bending test, for improving performance of the resulting finite element. However, such optimal conditions are optional, and the only necessity is to include the constant terms in the opted assumed strain field. However, similar to the classical formulation, it is suggested not to give any priority to each of the coordinates, (x or y). In addition, if the analytical estimation of the strain field is available, then the terms of the approximated field can be selected based on the known analytical solution.

Nevertheless, after choosing the desired terms, the assumed strain field can be optimized by applying any desired optimal condition [38, 63, 71]. The most common criteria are compatibility and equilibrium conditions [63, 71]. The compatibility of the strain field is achieved provided that the next relationship exists between the strain components [71]:

$$\frac{\partial^2 \varepsilon_x}{\partial y^2} + \frac{\partial^2 \varepsilon_y}{\partial x^2} = \frac{\partial^2 \gamma_{xy}}{\partial x \partial y}. \quad (2)$$

It is obvious that in the general case, for the satisfaction of the compatibility or any other criteria; it might be needed that some strain states be dependent on each other. Therefore, imposing the optimized condition results in the dependency of some strain states to each other and therefore, it reduces the number of independent strain states [27]. As it will be shown later, the number of independent strain states is equal to the number of required degrees of freedom for the element.

The other conventional optimal condition is the equilibrium. It was proved that if the equilibrium equation is satisfied within an element, then it can be included in the Trefftz formulation [27]. The equation of equilibrium for the plane problems is defined as follows:

$$\begin{cases} \frac{\partial \sigma_x}{\partial x} + \frac{\partial \tau_{xy}}{\partial y} + F_x = 0, \\ \frac{\partial \sigma_y}{\partial y} + \frac{\partial \tau_{xy}}{\partial x} + F_y = 0, \end{cases} \quad (3)$$

where F_x and F_y are the body forces in the x and y directions, respectively. Also, σ_x , σ_y and τ_{xy} are normal and shearing stresses, respectively. To rewrite the equilibrium equation in terms of strain, it is necessary to relate the stresses to the strains. For the plane problems, the coming relations connect stresses and strains:

$$\begin{cases} \sigma_x = 2G\varepsilon_x + \lambda(\varepsilon_x + \varepsilon_y), \\ \sigma_y = 2G\varepsilon_y + \lambda(\varepsilon_x + \varepsilon_y), \\ \tau_{xy} = G\gamma_{xy}. \end{cases} \quad (4)$$

In the previous equations, G and ν are the shear modulus and Poisson's ratio, respectively. Indeed λ is called the Lamé constant and is equal to $\frac{\nu E}{(1+\nu)(1-2\nu)}$ for the plane stress condition. In the case of plane strain, this constant is equal to $\frac{\nu E}{(1+\nu)(1-2\nu)}$. Here, E stands for modulus of elasticity. Substituting equations (4) in the equilibrium equation results in the following relations:

$$\begin{cases} (2G + \lambda) \frac{\partial \varepsilon_x}{\partial x} + \lambda \frac{\partial \varepsilon_y}{\partial x} + G \frac{\partial \gamma_{xy}}{\partial y} + F_x = 0, \\ \lambda \frac{\partial \varepsilon_x}{\partial y} + (2G + \lambda) \frac{\partial \varepsilon_y}{\partial y} + G \frac{\partial \gamma_{xy}}{\partial x} + F_y = 0. \end{cases} \quad (5)$$

It is noteworthy that imposing the optimized conditions, such as; compatibility and equilibrium, is not necessary steps for developing a plane element based on assumed strain approach, and as it will be shown in the coming section, there are finite elements, which do not consider these criteria [71, 72]. However, enforcing these conditions to the assumed strain field improves the performance of the resulting element. As mentioned previously, the inclusion of the optimized condition makes some of the strain states to be depended on the other ones. When the dependent strain states are determined, the assumed strain field is rewritten in terms of the independent ones. The next step is to calculate the associated displacement field. For this purpose, the strain-displacement formulas are utilized:

$$\begin{cases} \varepsilon_x = \frac{\partial u}{\partial x}, \\ \varepsilon_y = \frac{\partial v}{\partial y}, \\ \gamma_{xy} = \frac{1}{2} \left(\frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} \right). \end{cases} \quad (6)$$

In these relations, u and v are displacements in the x and y directions, respectively. Based on these qualities, the displacements in x and y directions are derived by integrating normal strain components with respect to their associated coordinates:

$$\begin{cases} u(x, y) = \int \varepsilon_x dx + f_1(y), \\ v(x, y) = \int \varepsilon_y dy + f_2(x). \end{cases} \quad (7)$$

Here, f_1 and f_2 are the results of integrating shear strain with respect to the coordinates and imposing the rigid body modes condition. In the case of plane problems, three rigid body modes exist in the displacement field, namely u_o , v_o , and r_o , signifying the rigid body displacements in x and y directions and the rigid body rotation, respectively. As it was mentioned, the existence of these terms is the necessary convergence condition. Accordingly, these modes are also counted among the independent strain states that can be arranged in a vector arrangement indicated by S . This vector is called strain state vector, which consists of the independent strain states. They are the coefficients of the strain field approximations in (1), as well as, the rigid body modes. The strain state vector is somehow equivalent to the

nodal displacement vector in the traditional displacement-based approach. By using the matrix notation, which is traditionally used in finite element formulation in structural engineering applications, it is possible to relate the displacement and strain fields to the strain state vector in the subsequent forms:

$$U = N_s S + \tilde{U}, \quad (8)$$

$$\varepsilon = B_s S + \tilde{\varepsilon}. \quad (9)$$

In these equations, N_s and B_s represent the displacement and strain interpolation matrices, respectively. The particular response of the displacement and strain fields, that is, \tilde{U} and $\tilde{\varepsilon}$, depend on the body forces. The next relation can be established between the vectors of nodal displacements and the strain states:

$$D = AS + \tilde{D} = \bar{D} + \tilde{D}. \quad (10)$$

Here, D and \tilde{D} are the nodal displacement vectors and the displacements due to body forces. Also, A is a geometric matrix including of the nodal displacement interpolation matrices. It is possible to construct the subsequent relations between the displacement and strains' fields of the element with the nodal displacement vector using (10) as follows:

$$U = N_s S + \tilde{U} = N_s (A^{-1} \bar{D}) + \tilde{U} = (N_s A^{-1}) \bar{D} + \tilde{U} = N \bar{D} + \tilde{U}, \quad (11)$$

$$\varepsilon = B_s S + \tilde{\varepsilon} = B_s (A^{-1} \bar{D}) + \tilde{\varepsilon} = (B_s A^{-1}) \bar{D} + \tilde{\varepsilon} = B \bar{D} + \tilde{\varepsilon}. \quad (12)$$

Assuming the body forces to be negligible in comparison with the applied external loads, the strains and displacements due to body forces, $\tilde{\varepsilon}$ and \tilde{U} , are neglected.

The final step of the formulation is to derive the element stiffness matrix and the nodal force vector. Among different approaches for this purpose, the minimization of the total potential energy is the most common approach. The total potential energy functional can be established as follows:

$$\Pi = \frac{1}{2} \int \sigma^T \varepsilon \, dV - \int U^T F \, dV - D^T P_{ext}. \quad (13)$$

In this equation, P_{ext} and F stand for the external nodal and the body forces, respectively. The element stiffness matrix and nodal force vector are derived by establishing the stationary of the following functional:

$$\frac{\partial \Pi}{\partial \bar{D}} = A^{-T} \left(\int B_s^T D_m B_s \, dv \right) A^{-1} \bar{D} - A^{-T} \left(\int N_s^T F \, dv \right) - P_{ext} = K \bar{D} - P = 0. \quad (14)$$

Therefore, the element stiffness matrix and the nodal force vector are derived in the following form:

$$K = A^{-T} \left(\int B_s^T D_m B_s dv \right) A^{-1} = A^{-T} K_0 A^{-1}, \quad (15)$$

$$P = P_{ext} + A^{-T} \left(\int N_s^T F dv \right), \quad (16)$$

where, D_m is the material matrix:

$$D_m = \frac{E}{1 - \nu^2} \begin{bmatrix} 1 & \nu & 0 \\ \nu & 1 & 0 \\ 0 & 0 & \frac{1-\nu}{2} \end{bmatrix}. \quad (17)$$

It is noteworthy that for plane problems, which is the main subject of this study, the volume integral in the above-mentioned equations simply transforms into an area integral considering the constant unit thickness for the plane structures.

3 Required optimal criteria

As mentioned previously, in order to achieve an optimal and efficient formulation, it is possible to impose different criteria on the assumed strain field. Two of these criteria, compatibility and equilibrium, are defined in the previous section. In this section, two other required criteria for the optimal performance of assumed strain elements are defined.

3.1 Pure bending test

To achieve optimal performance in flexural behavior, Felippa [18, 26] utilized the pure bending test. In this experiment, an Euler–Bernoulli beam is discretized by rectangular (or triangular) elements and loaded by the constant bending moments. To study in-plane bending along x and y axes, the two beams depicted in Figures 1 and 2 are considered.

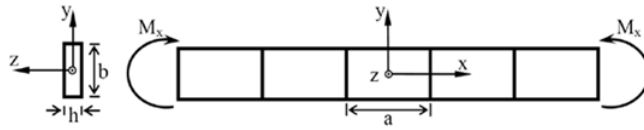
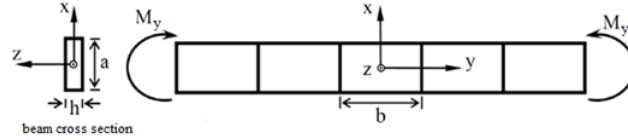


Figure 1: Pure bending test in the xy plane

The stored elastic energy due to the constant moments, M_x and M_y , in part of the beams meshed with one rectangular element (or two triangular elements) is derived according to the following relations:

$$U_x^{exact} = \frac{6aM_x^2}{Eb^3h}, \quad (18)$$

$$U_y^{exact} = \frac{6bM_y^2}{Ea^3h}. \quad (19)$$

Figure 2: Pure bending test in the zy plane

Based on this test, an element can present bending behavior exactly, provided that it can compute precise elastic stored energy. In other words, the ratio of the energy calculated according to the results of element analysis to the exact stored elastic energy should be equal to 1. For this purpose, the element stored energy, based on finite element responses, is computed using the coming equations:

$$U_x^{element} = \frac{1}{2} D_{bx}^T K D_{bx}, \quad (20)$$

$$U_y^{element} = \frac{1}{2} D_{by}^T K D_{by}. \quad (21)$$

In these relations, D_{bx} and D_{by} are the nodal displacement vector of the element and K is the element stiffness matrix. Therefore, the previously mentioned flexural energy ratios are derived as follows:

$$r_x = \frac{U_x^{element}}{U_x^{exact}}, \quad (22)$$

$$r_y = \frac{U_y^{element}}{U_y^{exact}}. \quad (23)$$

If r_x or r_y is equal to 1, the element passes the pure bending test and is able to represent exact bending behavior. If these ratios are greater or less than 1, then the element is over-stiff or over-flexible, respectively. Moreover, if the energy ratios are equal to 1 for any aspect ratio (a/b), then the element is optimal in bending behavior. Finally, the element suffers from shear locking on the condition that $a \gg b$ results in $r_x \gg 1$ or if $b \gg a$ leads to $r_y \gg 1$.

To study this test for the strain-based formulation, it is required to investigate the exact strain fields of the elements under pure bending. In the following, only the case of M_x will be discussed. The same reasoning goes for M_y , as well. The exact stress components of the beam subjected to the

pure bending M_x are as follows:

$$\begin{cases} \sigma_x = -\frac{12M_x y}{b^3 h}, \\ \sigma_y = 0, \\ \tau_{xy} = 0. \end{cases} \quad (24)$$

Therefore, the element stress field can be demonstrated by the linear function of y for σ_x in the subsequent form:

$$\begin{cases} \sigma_x = \alpha_1 + \alpha_2 y, \\ \sigma_y = 0, \\ \tau_{xy} = 0. \end{cases} \quad (25)$$

Based on the Hook's law, the corresponding strain field is as follows:

$$\begin{cases} \varepsilon_x = \frac{\alpha_1}{E} + \frac{\alpha_2}{E} y = \beta_1 + \beta_2 y, \\ \sigma_y = \frac{-\nu\alpha_1}{E} - \frac{\nu\alpha_2}{E} y = \beta_3 + \beta_4 y, \\ \tau_{xy} = 0. \end{cases} \quad (26)$$

From the previous relations, it can be concluded that a strain-based element would surely pass the pure bending test, provided that the constant and linear terms for the normal strains are included in the assumed strain field.

3.2 Rotational invariance

Because the finite elements may be rotated with the different angles and be placed in various locations in the structural meshes used for the analysis of the diverse problems, their characteristics must not change due to the rotation. Such an element is called a rotational invariant. Assuming that the coordinate system xy is rotated to a new form, which is indicated by $x'y'$. The displacements' components in this new coordinate system can be related to those of the initial coordinate system by using the following simple transformations:

$$u' = u \cos \theta + v \sin \theta, \quad (27)$$

$$v' = -u \sin \theta + v \cos \theta. \quad (28)$$

In this relation, θ is the rotation angle from xy system to the new $x'y'$ coordinates. Based on this relation, having the rigid body motions, u_0 and v_0 in the original coordinate, it is possible to calculate the correct value of u'_0 and v'_0 , in the $x'y'$ coordinates, by using equations (27) and (28). Regarding the rotational invariance property, it is required to consider the strain field with

a complete order. In other words, some incomplete interpolation polynomials produce strain states that are not invariant with rotation. Therefore, the rotational invariance can be guaranteed with the inclusion of all strain terms with a given order. For instance, the rotational mapping of a constant strain state is as follows:

$$\varepsilon'_x = \varepsilon_x \cos^2 \theta + \varepsilon_y \sin^2 \theta + \gamma_{xy} \sin \theta \cos \theta , \quad (29)$$

$$\varepsilon'_y = \varepsilon_x \sin^2 \theta + \varepsilon_y \cos^2 \theta - \gamma_{xy} \sin \theta \cos \theta , \quad (30)$$

$$\gamma'_{xy} = (\varepsilon_x - \varepsilon_y) \sin 2\theta + \gamma_{xy} \cos 2\theta . \quad (31)$$

Based on these relations, a strain-based element can represent the constant strains with respect to any system of the coordinates, only on the condition that its formulation takes into account all three cases of the constant strain states. Although, the completeness of the assumed strain field guarantees rotational invariance of the element, the elements with incomplete strain fields are not necessarily rotational dependent.

4 Triangular membrane elements

To the author's best knowledge, there are thirteen strain-based triangular elements formulated by using direct strain-based formulation. In this section, thirteen triangular membrane elements proposed based on the assumed-strain approach are briefly reviewed. These studies are arranged in historical order. Most of the available triangular element has similar geometry, a three-node triangle with three degrees of freedom at each node. However, in this configuration, the incorporation of the drilling degrees of freedom improves the performance of the element in bending analysis, but more recent works utilized different distribution of degrees of freedom provide better results. In addition, as it will be demonstrated, in the most of the existing elements, the equilibrium conditions are not imposed on the assumed strain fields. According to the outcomes of the more recent elements, considering the equilibrium conditions improve the accuracy and convergence rate of the suggested strain-based elements. More details about these elements are provided in the following descriptions. To facilitate understanding and reproducing of the elements by readers, the main details are presented in a table format.

4.1 Sabir (1985)

The first researcher that utilized the assumed strain approach to develop more powerful membrane elements is Sabir. In one of his early works, he [71] proposed a three-node triangular element with three-degrees of freedom at

each node. The assumed strain components satisfied the compatibility equation, but the equilibrium equations were not satisfied. The drilling degree of freedom for the element was defined by the subsequent relation:

$$\theta = \frac{1}{2} \left(\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} \right). \quad (32)$$

Details of the element formulation are presented in Table 1.

4.2 Sabir and Sfsndji (1995)

Sabir and Sfsndji [72] suggested a four-node triangular element by assumption of the linear normal strains and constant shear strain. The selected strain field satisfied the compatibility condition, but the equilibrium equations were not imposed on this strain field. Therefore, the strain state vector consisted of eight independent unknown coefficients. Three nodes were located at the vertexes of the triangle, and the fourth node was placed in the middle of one side. Each node had two translational degrees of freedom. Consequently, the element possessed eight degrees of freedom totally. This geometry made the element appropriate to be used as transitional element in finite element meshes. They compared the performance of their suggested element with standard displacement-based element by using few simple numerical problems [72]. Their attained results showed better performance of this triangular strain-based element. Table 2 presents details of this element.

4.3 Tayeh (2003)

In 2003, Tayeh [75] developed new strain-based elements for analysis of the plane structures. In contrast to the previous works by Sabir, he utilized higher-order terms and assumed an incomplete second-order field for the element (see Table 3). It is evident that some coefficients in this assumed strain field are common between different strain components. Tayeh provided no clear reason for this selection, but he stated indirectly that this strain field was selected in order to satisfy the compatibility condition. Similar to the element proposed by Sabir and Sfsndji [72], the assumed strain field did not satisfy equilibrium equations.

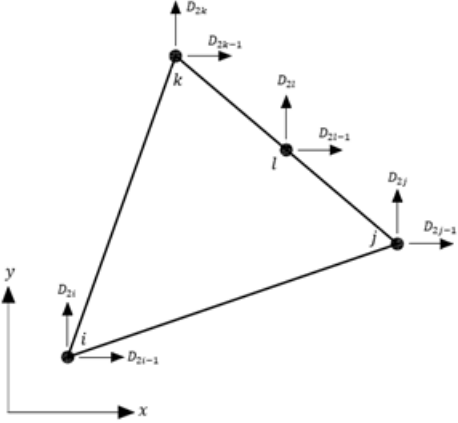
Table 1: Details of the triangular element proposed by Sabir [71]

Ref	Properties	Geometry
Sabir [71]	Geometry	
	Strain field	$\begin{cases} \varepsilon_x(x, y) = (\varepsilon_x)_o + (\varepsilon_{x,y})_o y + (\varepsilon_{y,x})_o x \\ \varepsilon_y(x, y) = (\varepsilon_y)_o + (\varepsilon_{y,x})_o x + (\varepsilon_{x,y})_o y \\ \gamma_{xy}(x, y) = (\gamma_{xy})_o + (\gamma_{xy,x})_o x + (\gamma_{xy,y})_o y \end{cases}$
	Strain state vector	$S = \{u_o \quad v_o \quad r_o \quad (\varepsilon_x)_o \quad (\varepsilon_y)_o \quad (\gamma_{xy})_o \quad (\varepsilon_{x,y})_o \quad (\varepsilon_{y,x})_o \quad (\gamma_{xy,x})_o\}^T$
	Nodal displacement vector	$D = \{D_{3i-2} \quad D_{3i-1} \quad D_{3i} \quad D_{3j-2} \quad D_{3j-1} \quad D_{3j} \quad D_{3k-2} \quad D_{3k-1} \quad D_{3k}\}^T$
	Strain interpolation matrix	$B_s = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & y & x & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & y & x & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & x+y \end{bmatrix}$
	Displacement interpolation matrix	$N_s = \begin{bmatrix} 1 & 0 & -y & x & 0 & \frac{y}{2} & xy & \frac{x^2-y^2}{2} & \frac{y^2}{2} \\ 0 & 1 & x & 0 & y & \frac{x}{2} & \frac{y^2-x^2}{2} & xy & \frac{x^2}{2} \\ 0 & 0 & 1 & 0 & 0 & 0 & -x & y & \frac{x-y}{2} \end{bmatrix}$
	Geometric matrix	$A = [N_{si} \quad N_{sj} \quad N_{sk}]^T$

4.4 Belarbi and Bourezane (2005-first triangular element)

As mentioned previously, most of the available triangular strain-based plane elements have the geometry similar to the one in Table 1. In fact, many of these elements attempted to improve the performance of element suggested

Table 2: Details of the triangular element proposed by Sabir and Sfindji [72]

Ref	Properties	Geometry
Sabir and Sfindji [72]	Geometry	
	Strain field	$\begin{cases} \varepsilon_x(x, y) = (\varepsilon_x)_o + (\varepsilon_{x,y})_o y \\ \varepsilon_y(x, y) = (\varepsilon_y)_o + (\varepsilon_{y,x})_o x \\ \gamma_{xy}(x, y) = (\gamma_{xy})_o \end{cases}$
	Strain state vector	$S = \{u_o \quad v_o \quad r_o \quad (\varepsilon_x)_o \quad (\varepsilon_y)_o \quad (\gamma_{xy})_o \quad (\varepsilon_{x,y})_o \quad (\varepsilon_{y,x})_o\}^T$
	Nodal displacement vector	$D = \{D_{2i-1} \quad D_{2i} \quad D_{2j-1} \quad D_{2j} \quad D_{2k-1} \quad D_{2k} \quad D_{2l-1} \quad D_{2l}\}^T$
	Strain interpolation matrix	$B_s = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & y & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & x \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$
	Displacement interpolation matrix	$N_s = \begin{bmatrix} 1 & 0 & -y & x & 0 & \frac{y}{2} & xy & -\frac{y^2}{2} \\ 0 & 1 & x & 0 & y & \frac{x}{2} & -\frac{x^2}{2} & xy \end{bmatrix}$
	Geometric matrix	$A = [N_{si} \quad N_{sj} \quad N_{sk} \quad N_{sl}]^T$

by Sabir [71]. In one of these research works, Belarbi and Bourezane [9] proposed a new element by incorporating Poisson's ratio in the assumed strain field. They suspected that unsatisfactory performance of the element proposed by Sabir might be due to the existence of coupling terms in the direct strains. Therefore, they included the Poisson's ratio in their assumed strain-field (see Table 4). Like the previous strain-based triangular elements, the utilized strain field in this study satisfied the compatibility condition, but the equilibrium equations were not considered in this formulation.

Table 3: Details of the triangular element proposed by Tayeh [75]

Ref	Properties	Geometry
Tayeh [75]	Geometry	Same as Table 1
	Strain field	$\begin{cases} \varepsilon_x(x, y) = (\varepsilon_x)_o + (\varepsilon_{x,y})_o y + (\varepsilon_{x,yy})_o \frac{y^2}{4} \\ \varepsilon_y(x, y) = (\varepsilon_y)_o + (\varepsilon_{y,x})_o x - (\varepsilon_{x,yy})_o \frac{x^2}{4} \\ \gamma_{xy}(x, y) = (\gamma_{xy})_o + (\varepsilon_{y,xx})_o x + (\varepsilon_{y,xx})_o y \\ - (\varepsilon_{x,y})_o \frac{x^2}{4} + (\varepsilon_{y,x})_o \frac{y^2}{4} \end{cases}$
	Strain state vector	$S = \{u_o \quad v_o \quad r_o \quad (\varepsilon_x)_o \quad (\varepsilon_y)_o \quad (\gamma_{xy})_o \quad (\varepsilon_{x,y})_o \quad (\varepsilon_{y,x})_o \quad (\varepsilon_{x,yy})_o\}^T$
	Nodal displacement vector	Same as Table 1
	Strain interpolation matrix	$B_s = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & y & 0 & \frac{y^2}{4} \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & x & -\frac{x^2}{4} \\ 0 & 0 & 0 & 0 & 0 & 1 & -\frac{x^2}{4} & \frac{y^2}{4} & x + y \end{bmatrix}$
	Displacement interpolation matrix	$N_s = \begin{bmatrix} 1 & 0 & -y & x & 0 & \frac{y}{2} & xy & \frac{y^3}{12} - \frac{y^2}{2} & \frac{xy^2}{4} + \frac{y^2}{2} \\ 0 & 1 & x & 0 & y & \frac{x}{2} & -\frac{x^3}{12} - \frac{x^2}{2} & xy & \frac{x^2}{2} - \frac{x^2 y}{4} \\ 0 & 0 & 1 & 0 & 0 & 0 & -\frac{x^2}{8} - x & -\frac{y^2}{8} + y & \frac{x}{2} - \frac{y}{2} - \frac{xy}{2} \end{bmatrix}$
	Geometric matrix	Same as Table 1

4.5 Belarbi and Bourezane (2005- second triangular element)

In 2005, Belarbi and Bourezane [10] performed another study and proposed a triangular strain-based element with the geometry similar to their previous work, but with a different strain field. They assumed linear variation of normal strain with respect to the perpendicular direction for this element, while the incomplete second-order field was assumed for shear strain. Once again, the three-node nine-degrees of freedom geometry was considered for this element. Further details of the element are provided in Table 5.

4.6 Rezaiee-Pajand and Yaghoobi (2014- first triangular element)

Rezaiee-Pajand and Yaghoobi [65] proposed a five-node triangular element with a complete linear strain field, which its geometry and details are presented in Table 6.

In the assumed strain field, there are nine strain states and three rigid body modes, which totally become twelve strain states. The selected strain

Table 4: Details of the first triangular element proposed by Belarbi and Bourezane [9]

Ref	Properties	Geometry
Belarbi and Bourezane [9]	Geometry	Same as Table 1
	Strain field	$\begin{cases} \varepsilon_x(x, y) = (\varepsilon_x)_o + (\varepsilon_{x,y})_o y - (\varepsilon_{x,x})_o x \frac{1-\nu}{2} \\ -(\varepsilon_{y,x})_o x \nu \\ \varepsilon_y(x, y) = (\varepsilon_y)_o + (\varepsilon_{y,x})_o x - (\varepsilon_{x,x})_o y \frac{1-\nu}{2} \\ -(\varepsilon_{x,y})_o y \nu \\ \gamma_{xy}(x, y) = (\gamma_{xy})_o + (\varepsilon_{x,x})_o x + (\varepsilon_{x,x})_o y \end{cases}$
	Strain state vector	$S = \{u_o \quad v_o \quad r_o \quad (\varepsilon_x)_o \quad (\varepsilon_y)_o \quad (\gamma_{xy})_o \quad (\varepsilon_{x,y})_o \quad (\varepsilon_{y,x})_o \quad (\varepsilon_{x,x})_o\}^T$
	Nodal displacement vector	Same as Table 1
	Strain interpolation matrix	$B_s = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & y & -\nu x & \frac{-\nu x}{2} \\ 0 & 0 & 0 & 0 & 1 & 0 & -\nu y & x & \frac{-\nu y}{2} \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & x + y \end{bmatrix}$
	Displacement interpolation matrix	$N_s = \begin{bmatrix} 1 & 0 & -y & x & 0 & \frac{y}{2} & xy & -\frac{y^2}{2} - \frac{\nu x^2}{2} & \frac{y^2}{2} - \frac{\nu x^2}{4} \\ 0 & 1 & x & 0 & y & \frac{x}{2} - \frac{x^2}{2} - \frac{\nu y^2}{2} & xy & \frac{x^2}{2} - \frac{\nu y^2}{4} \\ 0 & 0 & 1 & 0 & 0 & 0 & -x & y & \frac{x-y}{2} \end{bmatrix}$
	Geometric matrix	Same as Table 1

components satisfy the compatibility requirement. In addition, Rezaiee-Pajand and Yaghoobi enforced equilibrium conditions in this strain field. They found that the necessary condition for satisfaction of the equilibrium criteria is that some strain states be dependent to others. Therefore, they selected the two dependent strain states, and as a result; ten independent strain states remained. In agreement with the number of independent strain states, the resulting element needs ten degrees of freedom. Rezaiee-Pajand and Yaghoobi considered a triangular element with six nodes, as depicted in Table 6. It is evident, four of these nodes have two translational degrees of freedom, while the other two only have one translational degree of freedom perpendicular to the corresponding side of the element.

The displacements of mid-side nodes, which are perpendicular to the element sides, can be connected to the displacements of the nodes in x and y direction, by using the following relationship:

$$w = u \cdot \sin(\alpha) + v \cdot \cos(\alpha). \quad (33)$$

In which, α is the angle between the degree of freedom normal to the side and the x axis. In this case, due to the difference in the number of degrees of freedom at the nodes, the geometrical matrix is a bit dissimilar to the

Table 5: Details of the second triangular element proposed by Belarbi and Bourezane [10]

Ref	Properties	Geometry
Belarbi and Bourezane [10]	Geometry	Same as Table 1
	Strain field	$\begin{cases} \varepsilon_x(x, y) = (\varepsilon_x)_o + (\varepsilon_{x,y})_o y \\ \varepsilon_y(x, y) = (\varepsilon_y)_o + (\varepsilon_{y,x})_o x \\ \gamma_{xy}(x, y) = (\gamma_{xy})_o + (\gamma_{xy,xx})_o x^2 + (\gamma_{xy,xx})_o y^2 \end{cases}$
	Strain state vector	$S = \{u_o \quad v_o \quad r_o \quad (\varepsilon_x)_o \quad (\varepsilon_y)_o \quad (\gamma_{xy})_o \quad (\varepsilon_{x,y})_o \quad (\varepsilon_{y,x})_o \quad (\gamma_{xy,xx})_o\}^T$
	Nodal displacement vector	Same as Table 1
	Strain interpolation matrix	$B_s = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & y & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & x & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & x^2 + y^2 \end{bmatrix}$
	Displacement interpolation matrix	$N_s = \begin{bmatrix} 1 & 0 & -y & x & 0 & \frac{y}{2} & xy & -\frac{y^2}{2} & \frac{y^3}{3} \\ 0 & 1 & x & 0 & y & \frac{x}{2} & -\frac{x^2}{2} & xy & \frac{x^3}{3} \\ 0 & 0 & 1 & 0 & 0 & 0 & -x & y & \frac{x^2 - y^2}{2} \end{bmatrix}$
	Geometric matrix	Same as Table 1

previously reviewed finite elements (see Table 6). Here, the first sub-matrices are derived by replacing the node coordinates in the matrix N_s while for the last two sub-matrices, a different relation was utilized (see Table 6).

Despite its irregular degrees of freedom, the numerical evaluations showed that the performance of this element is better than many of the previous membrane elements. There are different reasons for the better performance of this element. First, all the components of its strain field had the complete term for a linear approximation. Moreover, the equilibrium equations are imposed on the assumed strain field, and the independent strain states are excluded from the strain state vector. The later property provides also the advantage of reducing the number of degrees of freedom in the element. These investigators realized that the geometry of the element and its node locations also had significant effects on its performance. This influence resulted in a new series of more accurate elements, which will be discussed in the coming sections [58, 59, 61, 62].

Table 6: Details of the first triangular element proposed by Rezaiee-Pajand and Yaghoobi [65]

Ref	Properties	Geometry
Rezaiee-Pajand and Yaghoobi [65]	Geometry	
	Strain field	$\begin{cases} \varepsilon_x(x, y) = (\varepsilon_x)_o + (\varepsilon_{x,x})_o x + (\varepsilon_{x,y})_o y \\ \varepsilon_y(x, y) = (\varepsilon_y)_o + (\varepsilon_{y,x})_o x + (\varepsilon_{y,y})_o y \\ \gamma_{xy}(x, y) = (\gamma_{xy})_o + (\gamma_{xy,x})_o x + (\gamma_{xy,y})_o y \end{cases}$
	Strain state vector	$S = \{u_o \quad v_o \quad r_o \quad (\varepsilon_x)_o \quad (\varepsilon_y)_o \quad (\gamma_{xy})_o \quad (\varepsilon_{x,x})_o \quad (\varepsilon_{x,y})_o \quad (\varepsilon_{y,x})_o \quad (\varepsilon_{y,y})_o\}^T$
	Nodal displacement vector	$D = \{D_{2i-1} \quad D_{2i} \quad D_{2j-1} \quad D_{2j} \quad D_{2k-1} \quad D_{2k} \quad D_{2l-1} \quad D_{2l} \quad D_m \quad D_n\}^T$
	Strain interpolation matrix	$B_s = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & x & y & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & x & y \\ 0 & 0 & 0 & 0 & 0 & 1 & -\frac{(2G+\lambda)y}{G} & -\frac{\lambda}{G}x & -\frac{\lambda}{G}y & -\frac{(2G+\lambda)x}{G} \end{bmatrix}$
	Displacement interpolation matrix	$N_s = \begin{bmatrix} 1 & 0 & -y & x & 0 & \frac{y}{2} & \frac{x^2}{2} - \frac{(2G+\lambda)y^2}{2G} & xy & -y^2(\frac{\lambda}{2G} + \frac{1}{2}) & 0 \\ 0 & 1 & x & 0 & y & \frac{x}{2} & 0 & -x^2(\frac{\lambda}{2G} + \frac{1}{2}) & xy & \frac{y^2}{2} - \frac{(2G+\lambda)x^2}{2G} \end{bmatrix}$
	Geometric matrix	$A = [N_{si} \quad N_{sj} \quad N_{sk} \quad N_{sl} \quad N_{sm} \quad N_{sn}]^T$ $N_{s\beta} = \begin{bmatrix} \sin(\alpha) & \cos(\alpha) & x_\beta \cos(\alpha) & -y_\beta \sin(\alpha) \\ x_\beta \sin(\alpha) & y_\beta \cos(\alpha) & \frac{x_\beta \cos(\alpha)}{2} - \frac{y_\beta \sin(\alpha)}{2} \\ \left(\frac{x_\beta^2}{2} - \frac{(2G+\lambda)y_\beta^2}{2G}\right) \sin(\alpha) \\ (x_\beta y_\beta) \sin(\alpha) - \left(x_\beta^2 \left(\frac{\lambda}{2G} + \frac{1}{2}\right)\right) \cos(\alpha) \\ (x_\beta y_\beta) \cos(\alpha) - \left(y_\beta^2 \left(\frac{\lambda}{2G} + \frac{1}{2}\right)\right) \sin(\alpha) \\ \left(\frac{y_\beta^2}{2} - \frac{(2G+\lambda)x_\beta^2}{2G}\right) \cos(\alpha) \end{bmatrix} \quad \beta = m, n$

4.7 Rezaiee-Pajand and Yaghoobi (2014- second triangular element)

In another study, Rezaiee-Pajand and Yaghoobi [66] utilized the complete linear strain field of the previous study, but with a different element geometry. In this work, they proposed a seven-node triangular element, which is depicted in Table 7. Six of the ten required degrees of freedom were allocated at the vertex nodes, which had two translational degrees of freedom each. Three mid-side nodes had only one translational degrees of freedom perpendicular to their sides, and the last degree of freedom was a drilling one at the center node. The first three sub-matrices for nodes i , j , and k of the geometric matrix, A , are derived by replacing the coordinates of these nodes in the displacement interpolation matrix. For the three mid-side nodes, the respective sub-matrix is computed by using the equation given in Table 6 for $N_{s\beta}$. Finally, the last sub-matrix, N_{so} , is derived from the next equality presented in the last row of Table 7, after the geometric matrix. The authors developed this formulation for geometrical nonlinear analysis of plane structures. For this purpose, they took advantage of the co-rotational formulation [54, 3]. For this purpose, a local coordinate system, which its origin was located at the center of the element, was considered. This coordinate system translates and rotates with the element. Then, the element was formulated in this new system. Since the nonlinear analysis was not in the scope of this study, for further information about the co-rotational formulation, one can refer to references [54, 66].

4.8 Rebiai (2018)

In a more recent attempt to propose three-node nine-degree of freedom triangular element, Rebiai [48] suggested a new strain-based element with incomplete second-order strain field. This strain field satisfies the compatibility condition, but the equilibrium conditions were not fulfilled within the element. In fact, Rebiai did not intend to impose the equilibrium equations, (Please see Table 8).

4.9 Rezaiee-Pajand, Gharaei-Moghaddam, and Ramezani (2019- first triangular element)

In one of the most-recent studies, Rezaiee-Pajand, Gharaei-Moghaddam, and Ramezani [58] suggested new triangular elements. They utilized the complete linear strain field, which was the same as the previous studies by Rezaiee-

Table 7: Details of the second triangular element proposed by Rezaiee-Pajand and Yaghoobi [66]

Ref	Properties	Geometry
Rezaiee-Pajand and Yaghoobi [66]	Geometry	
	Strain field	Same as Table 6
	Strain state vector	Same as Table 6
	Nodal displacement vector	$D = \{D_{2i-1} \ D_{2i} \ D_{2j-1} \ D_{2j} \ D_{2k-1} \ D_{2k} \ D_l \ D_m \ D_n \ D_p\}^T$
	Strain interpolation matrix	Same as Table 6
	Displacement interpolation matrix	Same as Table 6
	Geometric matrix	$A = [N_{si} \ N_{sj} \ N_{sk} \ N_{sl} \ N_{sm} \ N_{sn} \ N_{sp}]^T$ $N_{sp} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 & \frac{(2G+\lambda)y_p}{2G} \\ -\frac{(2G+\lambda)x_p}{2G} & \frac{(2G+\lambda)y_p}{2G} & -\frac{(2G+\lambda)x_p}{2G} \end{bmatrix}$

Pajand and Yaghoobi [65, 66]. Moreover, they also imposed the equilibrium condition to specify the dependent strain states. The differences between these new elements and previous ones are in the geometry of the elements. In their first element, the authors assumed a five-node triangular element, with three vertex and two mid-side nodes (see Table 9). Each node has two translational degrees of freedom.

As mentioned previously, it has been proved through extensive investigations, that configuration of the nodes and distribution of the degrees of freedom have been influential in the performance of the resulting elements [65, 66]. Numerical studies showed that the last presented element provided

Table 8: Details of the triangular element proposed by Rebiai [48]

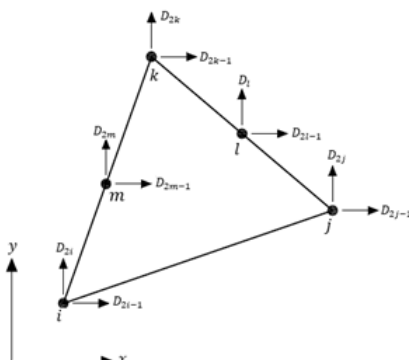
Ref	Properties	Geometry
Rebiai [48]	Geometry	Same as Table 1
	Strain field	$\begin{cases} \varepsilon_x(x, y) = (\varepsilon_x)_o + (\varepsilon_{x,y})_o y + (\varepsilon_{x,yy})_o y^2 \\ \varepsilon_y(x, y) = (\varepsilon_y)_o + (\varepsilon_{y,x})_o x + (\varepsilon_{x,yy})_o x^2 \\ \gamma_{xy}(x, y) = (\gamma_{xy})_o + 2(\varepsilon_{x,yy})_o x^2 + 2(\varepsilon_{x,yy})_o y^2 + 4(\varepsilon_{x,yy})_o xy \end{cases}$
	Strain state vector	$S = \{u_o \quad v_o \quad r_o \quad (\varepsilon_x)_o \quad (\varepsilon_y)_o \quad (\gamma_{xy})_o \quad (\varepsilon_{x,y})_o \quad (\varepsilon_{y,x})_o \quad (\varepsilon_{x,yy})_o\}^T$
	Nodal displacement vector	Same as Table 1
	Strain interpolation matrix	$B_s = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & y & 0 & y^2 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & x & x^2 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 2(x+y)^2 \end{bmatrix}$
	Displacement interpolation matrix	$N_s = \begin{bmatrix} 1 & 0 & -y & x & 0 & \frac{y}{2} & xy & -\frac{y^2}{2} & \frac{2y^3}{3} + xy^2 \\ 0 & 1 & x & 0 & y & \frac{x}{2} & -\frac{x^2}{2} & xy & \frac{2x^3}{3} + yx^2 \\ 0 & 0 & 1 & 0 & 0 & 0 & -x & y & x^2 - y^2 \end{bmatrix}$
	Geometric matrix	Same as Table 1

very accurate responses, especially for the elements under the bending loads. Something that might be questionable about the geometry of this element was the arbitrariness in the selection of the sides with mid nodes. To show that this selection did not have considerable effect on the performance of the element, the authors solved a numerical example with different locations of the mid-side nodes, and the results were identical. Moreover, the main idea behind this selection was to produce a powerful finite element for transitional purposes. To demonstrate this fact, Rezaiee-Pajand, Gharaei-Moghaddam, and Ramezani [58] investigated the performance of this element as a transitional element in numerical example and compared its behavior with standard displacement-based transitional elements. The results showed superiority of the mentioned element.

4.10 Rezaiee-Pajand, Gharaei-Moghaddam, and Ramezani (2019- second triangular element)

After presenting the previous element, Rezaiee-Pajand, Gharaei-Moghaddam, and Ramezani [58] considered another configuration for the complete linear strain field. In this element, the authors considered the well-known three node nine-degree of freedom triangular element and added an internal node with one translational degree of freedom in an arbitrary direction. This element geometry is demonstrated in Table 10. By using (33), the displacement of the

Table 9: Details of the first triangular element proposed by Rezaiee-Pajand, Gharaei-Moghaddam, and Ramezani [58]

Ref	Properties	Geometry
Rezaiee-Pajand and Gharaei-Moghaddam [58]	Geometry	
	Strain field	Same as Table 6
	Strain state vector	Same as Table 6
	Nodal displacement vector	$D = \{D_{2i-1} \ D_{2i} \ D_{2j-1} \ D_{2j} \ D_{2k-1} \ D_{2k} \ D_{2l-1} \ D_{2l}\}$ $D_{2m-1} \ D_{2m}\}^T$
	Strain interpolation matrix	$B_s = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & x & y & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & x & y \\ 0 & 0 & 0 & 0 & 0 & 1 & -\frac{(2G+\lambda)y}{G} & -\frac{\lambda}{G}x & -\frac{\lambda}{G}y & -\frac{(2G+\lambda)x}{G} \end{bmatrix}$
	Displacement interpolation matrix	$N_s = \begin{bmatrix} 1 & 0 & -y & x & 0 & \frac{y}{2} & \frac{x^2}{2} - \frac{(2G+\lambda)y^2}{2G} & xy & -y^2(\frac{\lambda}{2G} + \frac{1}{2}) & 0 \\ 0 & 1 & x & 0 & y & \frac{x}{2} & 0 & -x^2(\frac{\lambda}{2G} + \frac{1}{2}) & xy & \frac{y^2}{2} - \frac{(2G+\lambda)x^2}{2G} \end{bmatrix}$
	Geometric matrix	$A = [N_{si} \ N_{sj} \ N_{sk} \ N_{sl} \ N_{sm}]^T$

internal node can be connected to the displacements in x and y directions. The internal node of the element can be removed by the static condensation approach, and therefore, the element becomes a three-node nine-degrees of freedom triangular element, which is a common element in general finite element programs. This element provided surprisingly accurate results under different types of loading and especially shear loading, in which many of the available elements failed to provide the exact response in the case of distorted mesh [58]. In addition to the fast convergence and high accuracy, this element was highly insensitive to the mesh distortion. Numerical examinations demonstrated that the direction of internal degree of freedom did not have any noticeable effect on the accuracy and performance of the element [58].

Table 10: Details of the second triangular element proposed by Rezaiee-Pajand, Gharaei-Moghaddam, and Ramezani [58]

Ref	Properties	Geometry
Rezaiee-Pajand and Gharaei-Moghaddam [58]	Geometry	
	Strain field	Same as Table 6
	Strain state vector	Same as Table 6
	Nodal displacement vector	$D = \{D_{2i-1} \ D_{2i} \ D_{2j-1} \ D_{2j} \ D_{2k-1} \ D_{2k} \ D_{2l-1} \ D_{2l}\}$ $D_{2m-1} \ D_{2m}\}^T$
	Strain interpolation matrix	$B_s = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & x & y & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & x & y \\ 0 & 0 & 0 & 0 & 0 & 1 & -\frac{(2G+\lambda)y}{2G} & -\frac{\lambda}{G}x & -\frac{\lambda}{G}y & -\frac{(2G+\lambda)x}{G} \end{bmatrix}$
	Displacement interpolation matrix	$N_{s\beta}$ $= \begin{bmatrix} 1 & 0 & -y_\beta & x_\beta & 0 & \frac{y_\beta}{2} & \frac{x_\beta^2}{2} - \frac{(2G+\lambda)y_\beta^2}{2G} & x_\beta y_\beta & -y_\beta^2(\frac{\lambda}{2G} + \frac{1}{2}) & 0 \\ 0 & 1 & x_\beta & 0 & y_\beta & \frac{x_\beta}{2} & 0 & -x_\beta^2(\frac{\lambda}{2G} + \frac{1}{2}) & x_\beta y_\beta & \frac{x_\beta^2}{2} - \frac{(2G+\lambda)y_\beta^2}{2G} \\ 0 & 0 & 1 & 0 & 0 & 0 & \frac{(2G+\lambda)y_\beta}{2G} & -\frac{(2G+\lambda)x_\beta}{2G} & \frac{(2G+\lambda)y_\beta}{2G} & -\frac{(2G+\lambda)x_\beta}{2G} \end{bmatrix}$ $\beta = i, j, k$
	Geometric matrix	$A = [N_{si} \ N_{sj} \ N_{sk} \ N_{sl}]^T$ $N_{sl} = \begin{bmatrix} \sin(\alpha) & \cos(\alpha) & x_l \cos(\alpha) - y_l \sin(\alpha) & x_l \sin(\alpha) & y_l \cos(\alpha) \\ \frac{x_l \cos(\alpha)}{2} - \frac{y_l \sin(\alpha)}{2} & \left(\frac{x_l^2}{2} - \frac{(2G+\lambda)y_l^2}{2G}\right) \sin(\alpha) \\ (x_l y_l) \sin(\alpha) - \left(x_l^2 \left(\frac{\lambda}{2G} + \frac{1}{2}\right)\right) \cos(\alpha) \\ (x_l y_l) \cos(\alpha) - \left(y_l^2 \left(\frac{\lambda}{2G} + \frac{1}{2}\right)\right) \sin(\alpha) \\ \left(\frac{y_l^2}{2} - \frac{(2G+\lambda)x_l^2}{2G}\right) \cos(\alpha) \end{bmatrix}$

4.11 Rezaiee-Pajand, Gharaei-Moghaddam, and Ramezani (2019- third triangular element)

In another study, Rezaiee-Pajand, Gharaei-Moghaddam, and Ramezani [59] formulated a higher-order strain-based assuming complete second-order normal strains and linear shear strains (see Table 11). After imposing the equilibrium and compatibility conditions on the assumed strain field, seven de-

pendent strain states were excluded and eleven independent strain states remained.

Table 11: Details of the third triangular element proposed by Rezaiee-Pajand, Gharaei-Moghaddam, and Ramezani [59]

Ref	Properties	Geometry
Rezaiee-Pajand et al. [59]	Geometry	
	Strain field	$\begin{cases} \varepsilon_x(x, y) = (\varepsilon_x)_o + (\varepsilon_{x,x})_o x + (\varepsilon_{x,y})_o y + (\varepsilon_{x,xx})_o \left(\frac{x^2}{2}\right) \\ \quad + (\varepsilon_{x,xy})_o (xy) + (\varepsilon_{x,yy})_o \left(\frac{y^2}{2}\right) \\ \varepsilon_y(x, y) = (\varepsilon_y)_o + (\varepsilon_{y,x})_o x + (\varepsilon_{y,y})_o y + (\varepsilon_{y,xx})_o \left(\frac{x^2}{2}\right) \\ \quad + (\varepsilon_{y,xy})_o (xy) + (\varepsilon_{y,yy})_o \left(\frac{y^2}{2}\right) \\ \gamma_{xy}(x, y) = (\gamma_{xy})_o + (\gamma_{xy,x})_o x + (\gamma_{xy,y})_o y \end{cases}$
	Strain state vector	$S = \{u_o \quad v_o \quad r_o \quad (\varepsilon_x)_o \quad (\varepsilon_y)_o \quad (\gamma_{xy})_o \quad (\varepsilon_{x,x})_o \quad (\varepsilon_{x,y})_o \quad (\varepsilon_{y,x})_o \quad (\varepsilon_{y,y})_o \quad (\varepsilon_{x,yy})_o\}^T$
	Nodal displacement vector	$D = \{D_{2i-1} \quad D_{2i} \quad D_{2j-1} \quad D_{2j} \quad D_{2k-1} \quad D_{2k} \quad D_l \quad D_m \quad D_n \quad D_{2p-1} \quad D_{2p}\}^T$
	Strain interpolation matrix	$B_s = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & x & y & 0 & 0 & \frac{y^2}{2} + \frac{\lambda x^2}{2(2G+\lambda)} \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & x & y & -\frac{x^2}{2} + \frac{\lambda y^2}{2(2G+\lambda)} \\ 0 & 0 & 0 & 0 & 0 & 1 & -\frac{(2G+\lambda)y}{G} & -\frac{\lambda}{G}x & -\frac{\lambda}{G}y & -\frac{(2G+\lambda)x}{G} & 0 \end{bmatrix}$
	Displacement interpolation matrix	$N_s = \begin{bmatrix} 1 & 0 & -y & x & 0 & \frac{y}{2} & \frac{x^2}{2} - \frac{(2G+\lambda)y^2}{2G} & xy & -y^2 \frac{(G+\lambda)}{2G} & 0 & -\frac{x^3}{6} \frac{\lambda}{(2G+\lambda)} + \frac{xy^2}{2} \\ 0 & 1 & x & 0 & y & \frac{x}{2} & 0 & -x^2 \frac{(G+\lambda)}{2G} & xy & y^2 - \frac{(2G+\lambda)x^2}{2G} & -\frac{y^3}{6} \frac{\lambda}{(2G+\lambda)} + \frac{yx^2}{2} \end{bmatrix}$
	Geometric matrix	$A = [N_{si} \quad N_{sj} \quad N_{sk} \quad T_{sl} \quad T_{sm} \quad T_{sn} \quad N_{so}]^T$ $T_s = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 & \frac{(2G+\lambda)y}{2G} & -\frac{(2G+\lambda)x}{2G} & \frac{(2G+\lambda)y}{2G} & -\frac{(2G+\lambda)x}{2G} & -xy \end{bmatrix}$

The selected geometry of the element is demonstrated in Table 11. As it can be seen, the element had seven nodes and eleven degrees of freedom in agreement with the independent strain states.

4.12 Rezaiee-Pajand, Gharaei-Moghaddam, and Ramezani (2020- fourth triangular element)

Rezaiee-Pajand, Gharaei-Moghaddam, and Ramezani [61] utilized the assumed strain of their previous study [59] to formulate another higher-order strain-based element. The difference of this element with the previous one is the geometry of the element, which is demonstrated in Table 12.

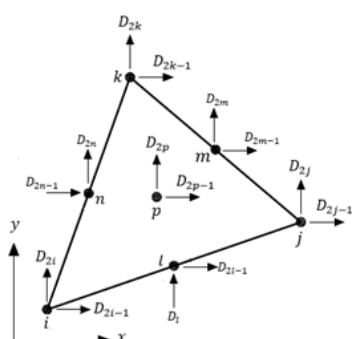
Table 12: Details of the fourth triangular element proposed by Rezaiee-Pajand, Gharaei-Moghaddam, and Ramezani [61]

Ref	Properties	
Rezaiee-Pajand et al. [61]	Geometry	
	Strain field	Same as Table 11
	Strain state vector	Same as Table 11
	Nodal displacement vector	$D = \{D_{3i-2} \ D_{3i-1} \ D_{3i} \ D_{3j-2} \ D_{3j-1} \ D_{3j} \ D_{3k-2} \ D_{3k-1} \ D_{3k} \ D_{2l-1} \ D_{2l}\}^T$
	Strain interpolation matrix	Same as Table 11
	Displacement interpolation matrix	Same as Table 11
	Geometric matrix	$A_G = [N_{si} \ T_{si} \ N_{sj} \ T_{sj} \ N_{sk} \ T_{sk} \ N_{sl} \ N_{sm} \ N_{sn} \ N_{so}]^T$

4.13 Rezaiee-Pajand, Ramezani, and Gharaei-Moghaddam (2020- fifth triangular element)

Since it was expected that using higher-order strain fields improves the accuracy of the elements, Rezaiee-Pajand, Ramezani, and Gharaei-Moghaddam [62] formulated a seven-node triangular element using complete second-order strain field (see Table 13). In this element formulation, the compatibility and equilibrium conditions were again satisfied.

Table 13: Details of the fifth triangular element proposed by Rezaiee-Pajand, Ramezani, and Gharaei-Moghaddam [62]

Ref	Properties
Rezaiee-Pajand et al. [62]	<div>Geometry</div> 
	<div>Strain field</div> $\begin{cases} \varepsilon_x(x, y) = (\varepsilon_x)_o + (\varepsilon_{x,x})_o x + (\varepsilon_{x,y})_o y + (\varepsilon_{x,xx})_o \left(\frac{x^2}{2}\right) \\ \quad + (\varepsilon_{x,xy})_o (xy) + (\varepsilon_{x,yy})_o \left(\frac{y^2}{2}\right) \\ \varepsilon_y(x, y) = (\varepsilon_y)_o + (\varepsilon_{y,x})_o x + (\varepsilon_{y,y})_o y + (\varepsilon_{y,xx})_o \left(\frac{x^2}{2}\right) \\ \quad + (\varepsilon_{y,xy})_o (xy) + (\varepsilon_{y,yy})_o \left(\frac{y^2}{2}\right) \\ \gamma_{xy}(x, y) = (\gamma_{xy})_o + (\gamma_{xy,x})_o x + (\gamma_{xy,y})_o y + (\gamma_{xy,xx})_o \left(\frac{x^2}{2}\right) \\ \quad + (\gamma_{xy,xy})_o (xy) + (\gamma_{xy,yy})_o \left(\frac{y^2}{2}\right) \end{cases}$
	<div>Nodal displacement vector</div> $D = \{D_{2i-1} \ D_{2i} \ D_{2j-1} \ D_{2j} \ D_{2k-1} \ D_{2k} \ D_{2l-1} \ D_{2l} \ D_{2m-1} \ D_{2m} \ D_{2n-1} \ D_{2n} \ D_{2o-1} \ D_{2o}\}^T$
	<div>Strain interpolation matrix</div> $B_s = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & x & y & 0 & 0 & \frac{y^2}{2} - \frac{(G+\lambda)^2 x^2}{2G} - \frac{\lambda x^2}{2G} & xy & -\frac{(G+\lambda)x^2}{2K} & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & x & y & -\frac{(G+\lambda)y^2}{2G} & 0 & \frac{x^2}{2} - \frac{Gy^2}{2K} & xy \\ 0 & 0 & 0 & 0 & 0 & 1 & -\frac{Ky}{G} - \frac{\lambda}{G}x - \frac{\lambda}{G}y - \frac{Kx}{G} & xy & -\frac{Ky^2}{2G} - \frac{\lambda x^2}{2G} & xy & -\frac{Kx^2}{2G} - \frac{\lambda y^2}{2G} \end{bmatrix}$
	<div>Displacement interpolation matrix</div> $N_s = \begin{bmatrix} 1 & 0 & -y & x & 0 & \frac{y^2}{2} - \frac{(2G+\lambda)y^2}{2G} & xy & -y^2 \left(\frac{G+\lambda}{2G}\right) & 0 & -\frac{y^2}{2} - \frac{Gx^2}{G(2G+\lambda)} & \frac{y^2}{2} - \frac{(2G+\lambda)y^2}{2G} & -x^2 \left(\frac{G+\lambda}{2G}\right) & -y^2 \left(\frac{G+\lambda}{2G}\right) \\ 0 & 1 & x & 0 & y & \frac{x^2}{2} & 0 & -x^2 \left(\frac{G+\lambda}{2G}\right) & xy & \frac{x^2}{2} - \frac{(2G+\lambda)x^2}{2G} & -y^2 \left(\frac{G+\lambda}{2G}\right) & -x^2 \left(\frac{G+\lambda}{2G}\right) & \frac{y^2}{2} - \frac{Gx^2}{G(2G+\lambda)} \end{bmatrix}$
	<div>Geometric matrix</div> $A_G = [N_{si} \ N_{sj} \ N_{sk} \ N_{sl} \ N_{sm} \ N_{sn} \ N_{so}]^T$

5 Quadrilateral membrane elements

The existing quadrilateral strain-based membrane elements are reviewed in this section. In contrast with the triangular elements, which are mostly developed by the assumption of linear or second-order strain-field, higher-order strain fields were utilized in the formulation of quadrilateral elements. In the quadrilateral elements, there are two common configurations. The first one is a four-node twelve-degrees of freedom quadrilateral element, and the second common geometry is a five-node ten-degrees of freedom element. More details are provided in the following.

5.1 Sabir and Sfindji (1995)

Like the triangular elements, the first quadrilateral element is proposed by Sabir [72]. In 1995, Sabir and Sfindji suggested a rectangular strain-based finite element with linear strain field.

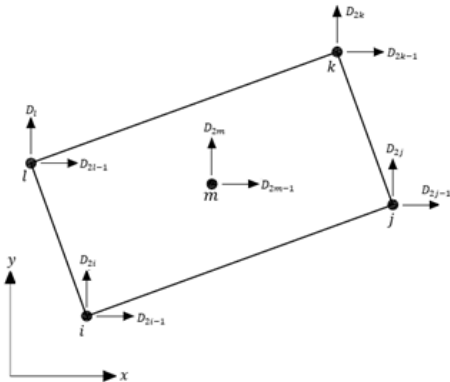
Once again, they assumed strain components satisfy compatibility condition, but the equilibrium equations are not fulfilled. Due to the existence of ten strain states, ten degrees of freedom are required. Therefore, Sabir and Sfindji considered the five-node rectangular element depicted in Table 15. Each node of this element has two translational degrees of freedom.

5.2 Tayeh (2003)

Another rectangular element was proposed by Tayeh [75]. He utilized an incomplete fourth-order approximation for normal strains and incomplete second-order assumption for the shear strain (see Table 15).

In agreement with these twelve strain states, a four-node rectangular element with three degrees of freedom at each node was considered. It should be noted that this element and the previous one proposed by Sabir and Sfindji had rectangular geometry, according to the original articles [72, 75], but based on the basics of strain- formulation, the effect of the geometry only entered in geometric matrix, A . So, by using the nodes of a general quadrilateral shape in construction of this matrix; the mentioned elements can have a general quadrilateral geometry.

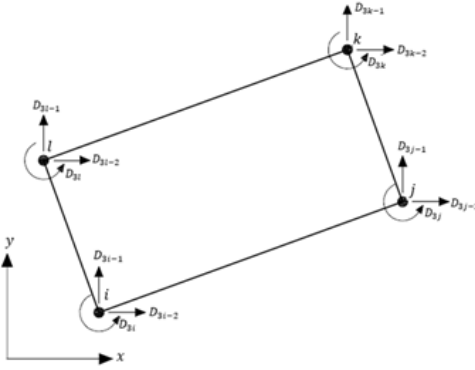
Table 14: Details of the Rectangular element proposed by Sabir and Sfindji [72]

Ref	Properties	
Sabir and Sfindji [72]	Geometry	
	Strain field	Same as Table 2
	Strain state vector	$S = \{u_o \quad v_o \quad r_o \quad (\varepsilon_x)_o \quad (\varepsilon_y)_o \quad (\gamma_{xy})_o \quad (\varepsilon_{x,y})_o \quad (\varepsilon_{y,x})_o \quad (\gamma_{xy,x})_o \quad (\gamma_{xy,y})_o\}^T$
	Nodal displacement vector	$D = \{D_{2i-1} \quad D_{2i} \quad D_{2j-1} \quad D_{2j} \quad D_{2k-1} \quad D_{2k} \quad D_{2l-1} \quad D_{2l} \quad D_{2m-1} \quad D_{2m}\}^T$
	Strain interpolation matrix	$B_s = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & y & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & x & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & x & y \end{bmatrix}$
	Displacement interpolation matrix	$N_s = \begin{bmatrix} 1 & 0 & -y & x & 0 & \frac{y}{2} & xy & \frac{-y^2}{2} & 0 & \frac{y^2}{2} \\ 0 & 1 & x & 0 & y & \frac{x}{2} & -\frac{x^2}{2} & xy & \frac{xyx^2}{2} & 0 \end{bmatrix}$
	Geometric matrix	$A = [N_{si} \quad N_{sj} \quad N_{sk} \quad N_{sl} \quad N_{sm}]^T$

5.3 Belarbi and Maalem (2005)

In 2005, Belarbi and Maalem [12] suggested an improved strain-based rectangular element by the assumption of linear normal strains and constant shear strain. The element geometry and distribution of nodes and degrees of freedom are the one proposed by Sabir and Sfindji.

Table 15: Details of the Rectangular element proposed by Tayeh [75]

Ref	Properties
Tayeh [75]	Geometry 
	Strain field $\begin{cases} \varepsilon_x(x, y) = (\varepsilon_x)_o + (\varepsilon_{x,y})_o y + (\varepsilon_{x,yy})_o y^2 + (\varepsilon_{x,yyy})_o xy^3 \\ \varepsilon_y(x, y) = (\varepsilon_y)_o + (\varepsilon_{y,x})_o x - (\varepsilon_{x,yy})_o x^2 - (\varepsilon_{x,yyy})_o yx^3 \\ \gamma_{xy}(x, y) = (\gamma_{xy})_o + (\gamma_{xy,x})_o x + (\gamma_{xy,y})_o y - (\varepsilon_{x,y})_o \frac{x^2}{2} - (\varepsilon_{y,x})_o \frac{y^2}{2} \end{cases}$
	Strain state vector $S = \{u_o \quad v_o \quad r_o \quad (\varepsilon_x)_o \quad (\varepsilon_y)_o \quad (\gamma_{xy})_o \quad (\varepsilon_{x,y})_o \quad (\varepsilon_{y,x})_o \quad (AE_{xy,x})_o \quad (\gamma_{xy,y})_o \quad (\varepsilon_{x,yy})_o \quad (\varepsilon_{x,yyy})_o\}^T$
	Nodal displacement vector $D = \{D_{3i-2} \quad D_{3i-1} \quad D_{3i} \quad D_{3i+1} \quad D_{3j-2} \quad D_{3j-1} \quad D_{3j} \quad D_{3j+1} \quad D_{3k-2} \quad D_{3k-1} \quad D_{3k} \quad D_{3k+1} \quad D_{3l-2} \quad D_{3l-1} \quad D_{3l} \quad D_{3l+1}\}^T$
	Strain interpolation matrix $B_s = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & y & 0 & 0 & 0 & y^2 & xy^3 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & x & 0 & 0 & -x^2 & -x^3 y \\ 0 & 0 & 0 & 0 & 0 & 1 & -\frac{x^2}{2} & -\frac{y^2}{2} & x & y & 0 & 0 \end{bmatrix}$
	Displacement interpolation matrix $N_s = \begin{bmatrix} 1 & 0 & -y & x & 0 & \frac{y}{2} & xy & -\frac{y^3}{6} - \frac{y^2}{2} & 0 & \frac{y^2}{2} & xy^2 & \frac{x^2 y^3}{2} \\ 0 & 1 & x & 0 & y & \frac{x}{2} & -\frac{x^3}{6} - \frac{x^2}{2} & xy & \frac{x^2}{2} & 0 & -x^2 y & -\frac{x^3 y^2}{2} \\ 0 & 0 & 1 & 0 & 0 & 0 & -\frac{x^2}{4} - x & \frac{y^2}{4} + y & \frac{x}{2} & -\frac{y}{2} & -2xy & -\frac{3x^2 y^2}{2} \end{bmatrix}$
	Geometric matrix $A = [N_{si} \quad N_{sj} \quad N_{sk} \quad N_{sl}]^T$

5.4 Rezaiee-Pajand and Yaghoobi (2012- first quadrilateral element)

The first strain-based generalized quadrilateral plane element, which its strain field satisfies both compatibility and equilibrium conditions, is proposed by Rezaiee-Pajand and Yaghoobi [63]. They [65, 66] took advantage of the linear strain field that they had used for development of triangular elements (see Table 6). After imposing the equilibrium criteria and determination of dependent strain states, ten independent strain states remain for the re-

Table 16: Details of the Rectangular element proposed by Belarbi and Maalem [12]

Ref	Properties
Belarbi and Maalem [12]	Geometry
	Strain field
	Strain state vector
	Nodal displacement vector
	Strain interpolation matrix
	Displacement interpolation matrix
	Geometric matrix

quired ten degrees of freedom. Rezaiee-Pajand and Yaghoobi considered the generalized five-node quadrilateral element, which is depicted in Table 17.

In another study, Rezaiee-Pajand and Yaghoobi [64] investigated the performance of two special rectangular variants of the generalized quadrilateral element. In the first element, the fifth node was located in its center, while this node was moved to the middle of one side in the next element. In this study, they showed that these strain-based elements were less sensitive to mesh distortion in comparison with their displacement-based counterparts. Moreover, they displayed that strain-based elements were completely free from shear parasitic errors.

5.5 Rebiai and Belounar (2013- first quadrilateral element)

A four-node rectangular strain-based element with incomplete fourth-order normal strains was proposed by Rebiai and Belounar [49]. The authors assumed linear shear strains for this element. This strain field, had twelve independent strain states. Regarding this field, it seems that there was no clear and rational basis behind this selection. Therefore, this field was probably the result of the trial-and-error process to attain the best performance of the resulting element. Rebiai and Belounar developed this element for axisymmetric, as well as, elastoplastic analysis of structures. For the nonlinear

Table 17: Details of the first quadrilateral element proposed by Rezaiee-Pajand and Yaghoobi [63]

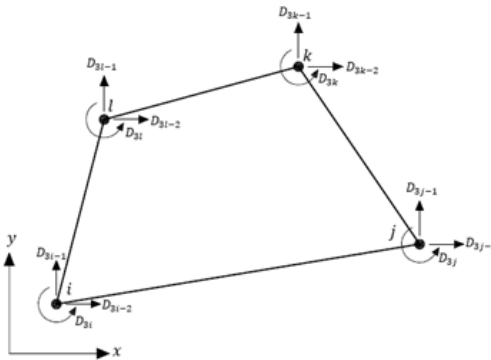
Ref	Properties	Geometry
Rezaiee-Pajand and Yaghoobi [63]	Geometry	
	Strain field	Same as Table 6
	Strain state vector	Same as Table 6
	Nodal displacement vector	$D = \{D_{2i-1} \ D_{2i} \ D_{2j-1} \ D_{2j} \ D_{2k-1} \ D_{2k} \ D_{2l-1} \ D_{2l} \ D_{2m-1} \ D_{2m}\}^T$
	Strain interpolation matrix	Same as Table 6
	Displacement interpolation matrix	Same as Table 6
	Geometric matrix	$A = \{N_{si} \ N_{sj} \ N_{sk} \ N_{sl} \ N_{sm}\}^T$

analysis, they took advantage of the Mohr–Coulomb yield criteria and the initial stress and strain methods [49]. They showed that the strain-based elements provided reliable and accurate results in nonlinear problems.

5.6 Rebiai and Belounar (2014- second quadrilateral element)

In another research work, Rebiai and Belounar suggested another variant of their previous element [50]. In this new element, they considered the strain field of their previous study [50], but added new linear term to the shear strain and changed the dependent term of the normal strains.

Table 18: Details of the first quadrilateral element proposed by Rebiai and Belounar [49]

Ref	Properties	
Rebiai and Belounar [49]	Geometry	
	Strain field	$\begin{cases} \varepsilon_x(x, y) = (\varepsilon_x)_o + (\varepsilon_{x,y})_o y + (\varepsilon_{x,x})_o x + (\varepsilon_{x,yy})_o y^2 + 2(\varepsilon_{x,xyy})_o xy^3 \\ \varepsilon_y(x, y) = (\varepsilon_y)_o + (\varepsilon_{x,x})_o x + 2(\varepsilon_{y,y})_o y - (\varepsilon_{x,yy})_o x^2 - 2(\varepsilon_{x,xyy})_o yx^3 \\ \gamma_{xy}(x, y) = 2(\gamma_{xy})_o + 2(\varepsilon_{x,y})_o x + 2(\varepsilon_{x,x})_o y + 2(\varepsilon_{y,y})_o y + 2(\varepsilon_{x,yy})_o y - 2(\varepsilon_{x,yy})_o x + 2(\gamma_{xy,x})_o x \end{cases}$
	Strain state vector	$S = \{u_o \quad v_o \quad r_o \quad (\varepsilon_x)_o \quad (\varepsilon_y)_o \quad (\gamma_{xy})_o \quad (\varepsilon_{x,y})_o \quad (\varepsilon_{x,x})_o \quad (\varepsilon_{y,y})_o \quad (\gamma_{xy,x})_o \quad (\varepsilon_{x,yy})_o \quad (\varepsilon_{x,xyy})_o\}^T$
	Nodal displacement vector	$D = \{D_{3i-2} \quad D_{3i-1} \quad D_{3i} \quad D_{3j-2} \quad D_{3j-1} \quad D_{3j} \quad D_{3k-2} \quad D_{3k-1} \quad D_{3k} \quad D_{3l-2} \quad D_{3l-1} \quad D_{3l}\}^T$
	Strain interpolation matrix	$B_s = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & y & x & 0 & 0 & y^2 & 2xy^3 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & x & 2y & 0 & -x^2 & -2x^3y \\ 0 & 0 & 0 & 0 & 0 & 2 & 2x & 2y & 2y & 2x & 2y - 2x & 0 \end{bmatrix}$
	Displacement interpolation matrix	$N_s = \begin{bmatrix} 1 & 0 & -y & x & 0 & y & xy & \frac{x^2+y^2}{2} & y^2 & 0 & xy^2 + y^2 & x^2y^3 \\ 0 & 1 & x & 0 & y & x & \frac{x^2}{2} & xy & y^2 & x^2 & -x^2y - x^2 & -x^3y^2 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & -y & x & -x - y - 2xy & -3x^2y^2 \end{bmatrix}$
	Geometric matrix	$A = [N_{si} \quad N_{sj} \quad N_{sk} \quad N_{sl}]^T$

Similar to their previous study, Rebiai and Belounar utilized this element for both linear and materially nonlinear analysis. In their numerical evaluations, they included different yield criteria, such as, Tresca, Von Mises and Mohr-Coulomb. Once more, the attained results demonstrated the superiority of strain formulation in comparison with classical displacement-based elements.

Table 19: Details of the second quadrilateral element proposed by Rebiai and Belounar [50]

Ref	Properties	
Rebiai and Belounar [50]	Geometry	Same as Table 18
	Strain field	$\begin{cases} \varepsilon_x(x, y) = (\varepsilon_x)_o + (\varepsilon_{x,y})_o y + (\varepsilon_y)_o x + (\varepsilon_{x,yy})_o y^2 \\ \quad + 2(\varepsilon_{x,yyy})_o xy^3 \\ \varepsilon_y(x, y) = (\varepsilon_y)_o + (\varepsilon_{y,x})_o x + (\varepsilon_{y,y})_o y - (\varepsilon_{x,yy})_o x^2 \\ \quad - 2(\varepsilon_{x,yyy})_o xy^3 \\ \gamma_{xy}(x, y) = 2(\gamma_{xy})_o + [2(\gamma_{xy})_o + 2(\varepsilon_{y,x})_o + (\varepsilon_{y,y})_o]y \\ \quad + 2[(\varepsilon_{x,y})_o + (\varepsilon_y)_o + (\gamma_{xy,x})_o]x \end{cases}$
	Strain state vector	$S = \{u_o \quad v_o \quad r_o \quad (\varepsilon_x)_o \quad (\varepsilon_y)_o \quad (\gamma_{xy})_o \quad (\varepsilon_{x,y})_o \quad (\varepsilon_{y,x})_o \quad (\varepsilon_{x,yy})_o \quad (\gamma_{xy,x})_o \quad (\varepsilon_{x,yyy})_o\}^T$
	Nodal displacement vector	Same as Table 18
	Strain interpolation matrix	$B_s = \begin{bmatrix} 0 & 0 & 0 & 1 & x & 0 & y & 0 & 0 & 0 & y^2 & 2xy^3 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & x & y & 0 & -x^2 & -2x^3y \\ 0 & 0 & 0 & 0 & 2x & 2y + 2 & 2x & 2y & y & 2x & 0 & 0 \end{bmatrix}$
	Displacement interpolation matrix	$N_s = \begin{bmatrix} 1 & 0 & -y & x & \frac{x^2}{2} & y^2 + y & xy & \frac{y^2}{2} & \frac{y^2}{2} & 0 & xy^2 & x^2y^3 \\ 0 & 1 & x & 0 & x^2 + y & x & \frac{x^2}{2} & xy & \frac{y^2}{2} & x^2 & -x^2y & -x^3y^2 \\ 0 & 0 & 1 & 0 & x & -y & 0 & 0 & -\frac{y}{2} & x & -2xy & -3x^2y^2 \end{bmatrix}$
	Geometric matrix	Same as Table 18

5.7 Rebiai, Saidani, and Bahloul (2015- Third quadrilateral element)

In the following of his previous researches, Rebiai, Saidani, and Bahloul [51] suggested another strain-based quadrilateral element for linear dynamic analysis of the plane problems. They utilized the general form of the strain field, which was used in their previous studies, but with slight modifications (see Table 20). The element geometry and its strain state and nodal displacement vectors were similar to the previous element proposed by Rebiai and Belounar [50]. The difference between these two membrane elements was in their strain and displacement interpolation matrices. Application of this new element for dynamic analysis of the plane problems proved acceptable accuracy of the assumed strain method for the dynamic problems. They utilized the lumped mass matrix for the element.

Table 20: Details of the quadrilateral element proposed by Rebiai, Saidani, and Bahloul [51]

Ref	Properties	
Rebiai et al. [51]	Geometry	Same as Table 18
	Strain field	$\begin{cases} \varepsilon_x(x, y) = (\varepsilon_x)_o + (\varepsilon_y)_o + (\varepsilon_{x,y})_o y + (\varepsilon_y)_o x + (\varepsilon_{x,yy})_o y^2 \\ \quad + 2(\varepsilon_{x,yyy})_o xy^3 \\ \varepsilon_y(x, y) = (\varepsilon_y)_o + (\varepsilon_{y,x})_o x + (\varepsilon_{y,y})_o y - (\varepsilon_{x,yy})_o x^2 \\ \quad - 2(\varepsilon_{x,yyy})_o yx^3 \\ \gamma_{xy}(x, y) = 2(\gamma_{xy})_o + [2(\gamma_{xy})_o + 2(\varepsilon_{y,x})_o + (\varepsilon_{y,y})_o] y \\ \quad + 2[(\varepsilon_{x,y})_o + (\varepsilon_y)_o + (\gamma_{xy,x})_o] x \end{cases}$
	Strain state vector	$S = \{u_o \quad v_o \quad r_o \quad (\varepsilon_x)_o \quad (\varepsilon_y)_o \quad (\gamma_{xy})_o \quad (\varepsilon_{x,y})_o \quad (\varepsilon_{y,x})_o \quad (\varepsilon_{x,yy})_o \quad (\gamma_{xy,x})_o \quad (\varepsilon_{x,yyy})_o \quad (\varepsilon_{x,yyy})_o\}^T$
	Nodal displacement vector	Same as Table 18
	Strain interpolation matrix	$B_s = \begin{bmatrix} 0 & 0 & 0 & 1 & x+1 & 0 & y & 0 & 0 & 0 & y^2 & 2xy^3 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & x & y & 0 & -x^2 & -2x^3y \\ 0 & 0 & 0 & 0 & 2x & 2y+2 & 2x & 2y & y & 2x & 0 & 0 \end{bmatrix}$
	Displacement interpolation matrix	$N_s = \begin{bmatrix} 1 & 0 & -y & x \frac{x^2}{2} + x & y^2 + y & xy & \frac{y^2}{2} & \frac{y^2}{2} & 0 & xy^2 & x^2y^3 \\ 0 & 1 & x & 0 & x^2 + y & x & \frac{x^2}{2} & xy & \frac{y^2}{2} & x^2 & -x^2y & -x^3y^2 \\ 0 & 0 & 1 & 0 & x & -y & 0 & 0 & -\frac{y}{2} & x & -2xy & -3x^2y^2 \end{bmatrix}$
	Geometric matrix	Same as Table 18

5.8 Rezaiee-Pajand and Yaghoobi (2015- fourth quadrilateral element)

In most of these formulations, there is no rational basis for the selection of the assumed strain field. However, Rezaiee-Pajand and his colleagues proposed application of the concept of Taylor expansion for this purpose. In an attempt to develop second order strain-based elements, Rezaiee-Pajand and Yaghoobi [67] suggested the complete second-order strain field (see Table 21). Excluding these dependent coefficients from the total strain states, fourteen independent strain states were remained. Accordingly, in this element, the four vertex nodes had three degrees of freedom, namely two displacement and one in-plane rotation, and the internal node had two translational degrees of freedom.

5.9 Rezaiee-Pajand and Yaghoobi (2015- fifth quadrilateral element)

Rezaiee-Pajand and Yaghoobi also proposed another second-order element, in which the assumed strain field was the same as the one presented in Table 21. In addition, the equilibrium criteria were imposed only on the linear terms of the strain components [67]. Therefore, only two strain states were characterized as dependent ones, and eighteen independent strain states remained. To generate an element with eighteen degrees of freedom, they considered a nine-node generalized quadrilateral element, which is demonstrated in Table 22. Each node of this element had two translational degrees of freedom.

5.10 Hamadi, Ayoub, and Maalem (2016)

In 2016, Hamadi, Ayoub, and Maalem [31] independently proposed a new quadrilateral finite element. This element is indeed the same as the element previously proposed by Rezaiee-Pajand and Yaghoobi [63], restricted to rectangular shapes.

5.11 Rezaiee-Pajand and Yaghoobi (2018- sixth quadrilateral elements)

In order to analyze geometrically nonlinear plane structures, Rezaiee-Pajand and Yaghoobi [63] modified their five-node quadrilateral element by the corotational approach [70]. Their findings showed that the strain-based formulation can provide accurate results for geometrically nonlinear analysis of structures. Besides, it did not show any sensitivity to the aspect ratio and mesh distortion. To sum up, the summary of the reviewed elements is presented in Table 23.

6 Other types of strain-based elements

Since the focus of the present study is on the membrane elements, only the available strain-based elements were reviewed in the previous sections. However, the advantages of strain formulation persuade researchers to have taken advantage of this approach in the development of finite elements for other types of structures. In order to provide a brief introduction to the application of assumed strain technique for the proposition of the different finite

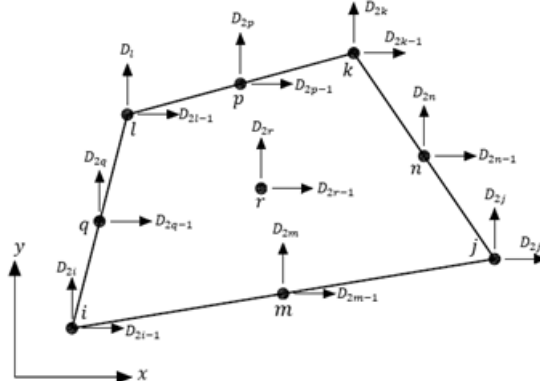
Table 21: Details of the fourth quadrilateral element proposed by Rezaiee-Pajand and Yaghoobi [67]

Ref	Properties	
Rezaiee-Pajand and Yaghoobi [67]	Geometry	
	Strain field	$\begin{cases} \varepsilon_x(x, y) = (\varepsilon_x)_o + (\varepsilon_{x,x})_o x + (\varepsilon_{x,y})_o y + (\varepsilon_{x,xx})_o \frac{x^2}{2} \\ \quad + (\varepsilon_{x,xy})_o xy + (\varepsilon_{x,yy})_o \frac{y^2}{2} \\ \varepsilon_y(x, y) = (\varepsilon_y)_o + (\varepsilon_{y,x})_o x + (\varepsilon_{y,y})_o y + (\varepsilon_{y,xx})_o \frac{x^2}{2} \\ \quad + (\varepsilon_{y,xy})_o xy + (\varepsilon_{y,yy})_o \frac{y^2}{2} \\ \gamma_{xy}(x, y) = (\gamma_{xy})_o + (\gamma_{xy,x})_o x + (\gamma_{xy,y})_o y + (\gamma_{xy,xx})_o \frac{x^2}{2} \\ \quad + (\varepsilon_{x,xy} + \varepsilon_{y,xx})_o xy + (\gamma_{xy,yy})_o \frac{y^2}{2} \end{cases}$
	Strain state vector	$S = \{u_o \quad v_o \quad r_o \quad (\varepsilon_x)_o \quad (\varepsilon_y)_o \quad (\gamma_{xy})_o \quad (\varepsilon_{x,x})_o \quad (\varepsilon_{x,y})_o \quad (\varepsilon_{y,x})_o \quad (\varepsilon_{y,y})_o \quad (\varepsilon_{x,xx})_o \quad (\varepsilon_{x,xy})_o \quad (\varepsilon_{x,yy})_o \quad (\varepsilon_{y,xx})_o \quad (\varepsilon_{y,xy})_o \quad (\varepsilon_{y,yy})_o \quad (\gamma_{xy,xx})_o \quad (\gamma_{xy,xy})_o \quad (\gamma_{xy,yy})_o\}^T$
	Nodal displacement vector	$D = \{D_{3i-2} \quad D_{3i-1} \quad D_{3i} \quad D_{3j-2} \quad D_{3j-1} \quad D_{3j} \quad D_{3k-2} \quad D_{3k-1} \quad D_{3k} \quad D_{3l-2} \quad D_{3l-1} \quad D_{3l} \quad D_{2m-1} \quad D_{2m}\}^T$
	Strain interpolation matrix	$B_s = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & x & y & 0 & 0 & xy & \frac{x^2}{2} - \frac{G}{2(G+\lambda)}x^2 & 0 & -\frac{G+\lambda}{2(G+\lambda)}x^2 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & x & y & 0 & -\frac{G+\lambda}{2(G+\lambda)}y^2 & xy & \frac{y^2}{2} - \frac{G}{2(G+\lambda)}y^2 \\ 0 & 0 & 0 & 0 & 0 & 1 & -\frac{2G+\lambda}{G}y - \frac{\lambda}{G}x - \frac{2G+\lambda}{G}x - \frac{\lambda}{2G}x^2 - \frac{2G+\lambda}{2G}y^2 & xy & -\frac{\lambda}{2G}y^2 - \frac{2G+\lambda}{2G}x^2 & xy \end{bmatrix}$
	Displacement interpolation matrix	$N_s = \begin{bmatrix} 1 & 0 & -y & x & 0 & \frac{x^2}{2} - \frac{2G+\lambda}{2G}y^2 & xy & -\frac{2G+\lambda}{2G}y^2 & 0 & \frac{x^2}{2} - \frac{2G+\lambda}{2G}y^2 & \frac{y^2}{2} - \frac{G}{2(G+\lambda)}x^2 & -\frac{G+\lambda}{2(G+\lambda)}y^2 & -\frac{G+\lambda}{12(G+\lambda)}x^3 & -\frac{G+\lambda}{12(G+\lambda)}y^3 \\ 0 & 1 & x & 0 & y & \frac{y^2}{2} & 0 & -\frac{2G+\lambda}{2G}x^2 & xy & \frac{x^2}{2} - \frac{2G+\lambda}{2G}y^2 & -\frac{G+\lambda}{2(G+\lambda)}x^2 & -\frac{G+\lambda}{12(G+\lambda)}y^3 & \frac{G+\lambda}{12(G+\lambda)}x^3 & \frac{G+\lambda}{12(G+\lambda)}y^3 \end{bmatrix}$
	Geometric matrix	$A = [N_{si} \quad T_{si} \quad N_{sj} \quad T_{sj} \quad N_{sk} \quad T_{sk} \quad N_{sl} \quad T_{sl} \quad N_{sm}]^T$ $T_s = [0 \quad 0 \quad 1 \quad 0 \quad 0 \quad \frac{2G+\lambda}{2G}y - \frac{\lambda}{2G}x - \frac{2G+\lambda}{2G}x - \frac{2G+\lambda}{2G}y^2 - \frac{\lambda}{2G}x^2 - \frac{2G+\lambda}{2G}(y^2 - x^2) - \frac{xy}{2} - \frac{2G+\lambda}{4G}(x^2 - y^2) \frac{xy}{2}]$

elements, a short review of the other works is presented in this section. Needless to say, for more details, it is necessary to refer to the original references, which are cited in the following paragraphs.

In the case of plate bending analysis, Belounar and Guenfoud [15] proposed a four-node rectangular plate element by assumption of linear curvature and second-order shear strains. The assumed strain field for the element only satisfies the compatibility condition. They showed that this element is more efficient than the corresponding displacement-based element. In 2014, Hamadi et al. [30] developed a rectangular element for plate bending analysis, based on the Kirchhoff theory, and compared its performance with the

Table 22: Details of the fifth element proposed by Rezaiee-Pajand and Yaghoobi [67]

Ref	Properties	
Rezaiee-Pajand and Yaghoobi [67]	Geometry	
	Strain field	Same as Table 21
	Strain state vector	$S = \{u_o \ v_o \ r_o \ (\varepsilon_x)_o \ (\varepsilon_y)_o \ (\gamma_{xy})_o \ (\varepsilon_{x,x})_o \ (\varepsilon_{x,y})_o \ (\varepsilon_{y,x})_o \ (\varepsilon_{y,y})_o \ (\varepsilon_{x,xx})_o \ (\varepsilon_{x,xy})_o \ (\varepsilon_{x,yy})_o \ (\varepsilon_{y,xx})_o \ (\varepsilon_{y,xy})_o \ (\varepsilon_{y,yy})_o \ (\gamma_{xy,xx})_o \ (\gamma_{xy,yy})_o\}^T$
	Nodal displacement vector	$D = \{D_{2i-1} \ D_{2i} \ D_{2j-1} \ D_{2j} \ D_{2k-1} \ D_{2k} \ D_{2l-1} \ D_{2l} \ D_{2m-1} \ D_{2m} \ D_{2n-1} \ D_{2n} \ D_{2p-1} \ D_{2p} \ D_{2q-1} \ D_{2q} \ D_{2r-1} \ D_{2r}\}^T$
	Strain interpolation matrix	$B_s = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & x & y & 0 & 0 & \frac{x^2}{2} & xy & \frac{y^2}{2} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & x & y & 0 & 0 & 0 & \frac{x^2}{2} & xy & \frac{y^2}{2} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -\frac{2G+\lambda}{G}y & -\frac{\lambda}{G}x & -\frac{\lambda}{G}y & -\frac{2G+\lambda}{G}x & 0 & 0 & xy & xy & 0 & 0 & \frac{x^2}{2} & \frac{y^2}{2} \end{bmatrix}$
	Displacement interpolation matrix	$N_s = \begin{bmatrix} 1 & 0 & -y & x & 0 & \frac{y}{2} & \frac{x^2}{2} - \frac{2G+\lambda}{2G}y^2 & xy & 0 & -\frac{G+\lambda}{2G}y^2 \\ 0 & 1 & x & 0 & y & \frac{x}{2} & 0 & -\frac{G+\lambda}{2G}x^2 & xy & 0 \\ 0 & 0 & 0 & \frac{x^3}{6} & \frac{x^2y}{2} & \frac{y^2x}{2} & 0 & -\frac{y^3}{6} & 0 & 0 & \frac{y^3}{6} \\ \frac{y^2}{2} - \frac{2G+\lambda}{2G}x^2 & 0 & -\frac{x^3}{6} & 0 & \frac{x^2y}{2} & \frac{y^2x}{2} & \frac{y^3}{6} & \frac{x^3}{6} & 0 & 0 \end{bmatrix}$
	Geometric matrix	$A = [N_{s_i} \ N_{s_j} \ N_{s_k} \ N_{s_l} \ N_{s_m} \ N_{s_n} \ N_{s_p} \ N_{s_q} \ N_{s_r}]^T$

displacement-based formulation. They showed the superiority of strain-based formulation in removing the shear locking problem. Another thin plate element, based on assumed strain approach, is proposed by Abderrahmani, Maalam, and Hamadi [1]. In this study, they utilized higher-order strain field in comparison with the previous elements. In another study, Abderrahmani et al. [2] formulated a new strain-based sector element for linear analysis of circular thin plates in 2017. Many years before this study, another sector strain-based element was proposed by Belarbi and Charif [11]. In 2018, the first triangular plate element, based on assumed strain approach, is proposed by Belounar, Benmebarek, and Belounar [14]. In this study, the

Table 23: Summary of the existing strain-based plane elements reviewed in this article

No.	Element	Geometry	Number of nodes	Number of dof	Assumed strain field			Drilling	Other features
					Normal strain	Shear strain	Optimal criteria		
1	Sabir [71]	Triangular	3	9	Complete linear	Complete linear	Compatibility	yes	-
2	Sabir and Sfindji [72]	Triangular	4	8	Incomplete linear	constant	Compatibility	no	Can be used as transitional element
3	Tayeh [75]	Triangular	3	9	Incomplete second-order	Incomplete second-order	Compatibility	yes	-
4	Belarbi and Bourezane [9]	Triangular	3	9	Complete linear	Incomplete linear	Compatibility	yes	Poisson's ratio is included in the element
5	Belarbi and Bourezane [10]	Triangular	3	9	Incomplete linear	Incomplete second-order	Compatibility	yes	-
6	Rezaiee-Pajand and Yaghoobi [65]	Triangular	6	10	Complete linear	Complete linear	Compatibility Equilibrium	no	Can be used as transitional element
7	Rezaiee-Pajand and Yaghoobi [66]	Triangular	7	10	Complete linear	Complete linear	Compatibility Equilibrium	yes	-
8	Rebiai [48]	Triangular	3	9	Incomplete second-order	Complete second-order	Compatibility	yes	-
9	Rezaiee-Pajand, Gharai-Moghaddam, and Ramezani [58]	Triangular	5	10	Complete linear	Complete linear	Compatibility Equilibrium	no	Can be used as transitional element
10	Rezaiee-Pajand, Gharai-Moghaddam, and Ramezani [58]	Triangular	4	10	Complete linear	Complete linear	Compatibility Equilibrium	yes	-
11	Rezaiee-Pajand, Gharai-Moghaddam, and Ramezani [59]	Triangular	7	11	Complete second-order	Complete linear	Compatibility Equilibrium	yes	-
12	Rezaiee-Pajand, Gharai-Moghaddam, and Ramezani [61]	Triangular	4	11	Complete second-order	Complete linear	Compatibility Equilibrium	yes	-
13	Rezaiee-Pajand, Ramezani, and Gharai-Moghaddam [62]	Triangular	7	14	Complete second-order	Complete second-order	Compatibility Equilibrium	yes	-
14	Sabir and Sfindji [72]	Rectangular	5	10	Complete linear	Complete linear	-	no	-
15	Tayeh [75]	Rectangular	4	12	Incomplete fourth-order	Incomplete second-order	Compatibility	yes	-
16	Belarbi and Maalem [12]	Rectangular	5	10	Complete linear	Constant	Compatibility	no	-
17	Rezaiee-Pajand and Yaghoobi [63]	Quadrilateral	5	10	Complete linear	Complete linear	Compatibility Equilibrium	no	-
18	Rezaiee-Pajand and Yaghoobi [64]	Rectangular	5	10	Complete linear	Complete linear	Compatibility Equilibrium	no	-
19	Rezaiee-Pajand and Yaghoobi [64]	Rectangular	5	10	Complete linear	Complete linear	Compatibility Equilibrium	no	Can be used as transitional element
20	Rebiai and Beloumar [49]	Quadrilateral	4	12	Incomplete fourth-order	Complete linear	Compatibility	yes	-
21	Rebiai and Beloumar [50]	Quadrilateral	4	12	Incomplete fourth-order	Complete linear	Compatibility	yes	-
22	Rebiai, Saidani, and Bahloul [51]	Quadrilateral	4	12	Incomplete fourth-order	Complete linear	Compatibility	yes	-
23	Rezaiee-Pajand and Yaghoobi [67]	Quadrilateral	5	14	Complete second-order	Complete second-order	Compatibility Equilibrium	yes	-
24	Rezaiee-Pajand and Yaghoobi [67]	Quadrilateral	9	18	Complete second-order	Complete second-order	Compatibility	yes	-
25	Hamadi, Ayoub, and Maalem [31]	Rectangular	5	10	Complete linear	Complete linear	Compatibility Equilibrium	no	-

authors took an advantage of linear strain components and evaluated the performance of this element in analysis of both thin and thick planes.

One of the first applications of the strain formulation in finite element analysis dates back to 1972 when Ashwell and Sabir [7] proposed a rectangular cylindrical shell element. Later, in 2004, Djoudi and Bahai [24] developed a strain-based shell element, which included openings and cut-outs. They utilized this element to perform vibration analysis of shell structures, and concluded that the strain-based element is more economic than the conventional displacement-based elements. In a comparative study, Hamadi et al. [32] investigated the performance of strain-based shell elements in comparison with the displacement-based elements. This study can be considered an extension of their previous research about plate elements [30]. In 2015, Mousa and Djoudi [45] performed vibration analysis on circular cylindrical shells with oblique ends, by using a new strain-based triangular shell element. For this element, they assumed linear curvature and third-order normal in-plane strains. The most-recent strain-based shell element is a flat triangular hybrid element proposed by Rezaiee-Pajand and Yaghoobi [69]. Trefftz functional was used in this element to formulate independent internal and boundary fields. It must be noted that this work is not the only available hybrid strain-based element and there are other similar studies in the literature. For instance, To and Liu [77] also proposed a triangular shell element, based on the Hellinger-Reissner hybrid strain formulation. Moreover, the hybrid formulation was also utilized for other types of structures. For example, in 2017, Rezaiee-Pajand and Yaghoobi [68] developed a hybrid plane element with assumed strain field.

In addition to the plate and shell element, the assumed strain approach was also utilized for developing three-dimensional finite elements. Belounar and Guerraiche [16] formulated a 3D eight-node brick element by assumption of linear strains. In this formulation, only compatibility criterion was imposed to the assumed strain field. Guerraiche, Belounar, and Bouzidi [29] proposed another variant of this element by using a different assumed strain field. In this new element, they included Poisson's ratio of the material in the strain field of the element. In the most-recent study, Messai, Belounar, and Merzouki [44] suggested a nine-node brick element with linear assumed strain field. In one of the most recent studies in this field, Rezaiee-Pajand, Gharaei-Moghaddam, and Ramezani [60] utilized strain-based formulation and developed a cracked plain element for analysis of cracked structures with open stable cracks.

7 Discussion and conclusion

In this article, which is the first part of a two-part study, the basic formulation steps of the assumed strain approach for developing plane elements were presented. In addition, twenty-five of the available strain-based membrane elements were reviewed, using the unified notations. The reviewed elements

were categorized into two groups of triangular and quadrilateral elements. According to the available literature, most of the accessible elements were formulated by using strain fields, which do not satisfy the equilibrium equations. Moreover, in many of these elements, the assumed strain field seems to be resulted from the trial-and-error process. Because no clear justification was provided by the authors for their selections. Despite this fact, in some cases, especially the elements proposed by Rezaiee-Pajand and his colleagues, the concept of Taylor expansion was considered in the selection of assumed strain components, and in many of these elements, both compatibility and equilibrium criteria were imposed. It is known for the researchers that utilizing incomplete polynomials for strain field might lead to incapability of the element to include Poisson's effect of strain states. Therefore, in some cases the authors tried to consider those strain fields with complete polynomial approximation. It is worth mentioning that the other elements, which do not consider this limitation, can also provide accurate results.

According to the presented review and in some cases, the same elements were proposed by different authors, independently. Therefore, this is one of the main motivations of the present study to provide a comprehensive reference to be used by authors, which helps prevent such duplicate element formulation. In addition, it seems that most of the researchers who worked in this field only followed their own research line and did not take advantage of the other's experiences. Accordingly, there are many unanswered questions about the performance of the strain-based elements that remain to be answered. For instance, it is widely known that some patterns for distribution of the element degrees of freedom resulted in the singular geometric matrix, which was shown by A . In addition, it was concluded that distribution of the degrees of freedom between the element nodes had a considerable effect on the element performance. Despite these indications, it is not clear how to prevent this problem and choose an optimal configuration. Due to these warnings, the application of the higher-order strain fields for membrane elements requires further investigation. According to these points, in the second part of this study, extensive numerical investigations will compare the performance of the reviewed membrane elements.

References

1. Abderrahmani, S., Maalam, T. and Hamadi, D. *On improved thin plate bending rectangular finite element based on the strain approach*, International Journal of Engineering Research in Africa(27) 76–86, Trans Tech Publications, 2016.
2. Abderrahmani, S., Maalem, T., Zatar, A. and Hamadi, D. *A new strain based sector finite element for plate bending problems*, International Jour-

- nal of Engineering Research in Africa (31) 1–13, Trans Tech Publications, 2017.
3. Alsafadie, R., Hjiiaj, M. and Battini, J.M. *Three-dimensional formulation of a mixed corotational thin-walled beam element incorporating shear and warping deformation*, Thin-Walled Struct. 49(4) (2011), 523–533.
 4. Andelfinger U. and Ramm E. *EAS-elements for two-dimensional, three-dimensional, plate and shell structures and their equivalence to HR-elements*, Int. J. Num. Meth. Eng. 36(8) (1993), 1311–1337.
 5. Ansari, S.U., Hussain, M., Mazhar, S., Manzoor, T., Siddiqui, K.J., Abid, M. and Jamal, H. *Mesh partitioning and efficient equation solving techniques by distributed finite element methods: A survey*, Arch. Comput. Methods Eng. 26(1) (2019), 1–16.
 6. Arregui-Mena, J.D., Worth, R.N., Hall, G., Edmondson, P.D., Giorla, A.B. and Burchell, T.D. *A review of finite element method models for nuclear graphite applications*, Arch. Comput. Methods Eng. 27(1) (2020), 331–350.
 7. Ashwell, D.G. and Sabir, A.B. *A new cylindrical shell finite element based on simple independent strain functions*, Int. J. Mech. Sci. 14(3) (1972), 171–183.
 8. Awrejcewicz, J., Krysko, A.V., Zhigalov, M.V. and Krysko, V.A. *Size-Dependent Theories of Beams, Plates and Shells*, Mathematical Modelling and Numerical Analysis of Size-Dependent Structural Members in Temperature Fields 142 (2021) 25–78. Springer, Cham.
 9. Belarbi, M.T. and Bourezane, M., *On improved Sabir triangular element with drilling rotation*, Rev. eur. génie civ., 9(9-10) (2005), 1151–1175.
 10. Belarbi, M.T. and Bourezane, M. *An assumed strain based on triangular element with drilling rotation*, Courier de Savoir, 6 (2005), 117–123.
 11. Belarbi, M.T. and Charif, A. *Nouvel élément secteur basé sur le modèle de déformation avec rotation dans le plan*. Revue Européenne des Éléments Finis, 7(4) (1998), 439–458 (In French).
 12. Belarbi, M.T. and Maalem, T. *On improved rectangular finite element for plane linear elasticity analysis*, Revue Européenne des Éléments Finis, 14(8) (2005), 985–997.
 13. Belarbi, M.O., Zenkour, A.M., Tati, A., Salami, S.J., Khechai, A. and Houari, M.S.A. *An efficient eight- node quadrilateral element for free vibration analysis of multilayer sandwich plates*, Int. J. Num. Meth. Eng. 122(9) (2021), 2360–2387.

14. Belounar, A., Benmebarek, S. and Belounar, L. *Strain based triangular finite element for plate bending analysis*, Mech. Adv. Mater. Struct. 27 (8) (2018), 1–13.
15. Belounar, L. and Guenfoud, M. *A new rectangular finite element based on the strain approach for plate bending*, Thin-Walled Struct., 43(1) (2005), 47–63.
16. Belounar, L. and Guerraiche, K. *A new strain-based brick element for plate bending*, Alex. Eng. J. 53(1) (2014), 95–105.
17. Belytschko, T. and Bindeman, L.P. *Assumed strain stabilization of the eight node hexahedral element*, Comput. Methods Appl. Mech. Eng. 105(2) (1993), 225–260.
18. Bergan, P.G. and Felippa, C. *A triangular membrane element with rotational degrees of freedom*, Comput. Methods Appl. Mech. Eng. 50(1) (1985), 25–69.
19. Bergan, P.G. and Nygård, M.K. *Finite elements with increased freedom in choosing shape functions*, Int. J. Num. Meth. Eng. 20(4) (1984), 643–663.
20. Boutagougua D. *A new enhanced assumed strain quadrilateral membrane element with drilling degree of freedom and modified shape functions*, Int. J. Num. Meth. Eng. 110(6) (2017), 573–600.
21. Chyzy, T. and Mackiewicz, M. *Special finite elements with adaptive strain field on the example of one-dimensional elements*, Appl. Sci. 11(2) (2021), 609.
22. Cinefra, M., de Miguel, A.G., Filippi, M., Houriet, C., Pagani, A. and Carrera, E. *Homogenization and free-vibration analysis of elastic meta-material plates by Carrera Unified Formulation finite elements*, Mech. Adv. Mater. Struct. 28(5) (2021), 476–485.
23. De Souza, R.M. *Force-based finite element for large displacement inelastic analysis of frames* Doctoral dissertation, University of California, Berkeley, 2000.
24. Djoudi, M.S. and Bahai, H. *Strain based finite element for the vibration of cylindrical panels with openings*, Thin-Walled Struct. 42(4) (2004), 575–588.
25. Dow, J.O., Cabiness, H.D. and Ho, T.H. *Linear strain element with curved edges*, J. Struct. Eng. 112(4) (1986), 692–708.
26. Felippa C.A. *A study of optimal membrane triangles with drilling freedoms*, Comput. Methods Appl. Mech. Eng. **192**(16-18) (2003), 2125–2168.

27. Felippa C.A. and Milotello C. *Membrane triangles with corner drilling freedomsII. The ANDES element*, Finite Elem. Anal. Des. **12**(3-4) (1992), 189-201.
28. Gal, E. and Levy, R. *Geometrically nonlinear analysis of shell structures using a flat triangular shell finite element*, Arch. Comput. Methods Eng. **13**(3) (2006), 331-388.
29. Guerraiche, K., Belounar, L. and Bouzidi, L. *A new eight nodes brick finite element based on the strain approach*, J. Solid Mech. **10**(1) (2018), 186-199.
30. Hamadi, D., Abderrahmani, S., Maalem, T. and Temami, O. *Efficiency of the Strain Based Approach Formulation for Plate Bending Analysis*, World Academy of Science, Engineering and Technology, International Journal of Mechanical, Aerospace, Industrial, Mechatronic and Manufacturing Engineering, **8**(8) (2014), 1408-1412.
31. Hamadi, D., Ayoub, A. and Maalem, T. *A new strain-based finite element for plane elasticity problems*, Engineering Computations, **33**(2) (2016), 562-579.
32. Hamadi, D., Temami, O., Zatar, A. and Abderrahmani, S. *A Comparative Study between Displacement and Strain Based Formulated Finite Elements Applied to the Analysis of Thin Shell Structures*, World Academy of Science, Engineering and Technology, International Journal of Civil, Environmental, Structural, Construction and Architectural Engineering, **8**(8) (2014), 875-880.
33. Hughes, T.J., Taylor, R.L. and Kanoknukulchai, W. *A simple and efficient finite element for plate bending*, Int. J. Num. Meth. Eng. **11**(10) (1977), 1529-1543.
34. Jafari, V., Vahdani, S.H. and Rahimian, M. *Derivation of the consistent flexibility matrix for geometrically nonlinear Timoshenko frame finite element*, Finite Elem. Anal. Des. **46**(12) (2010), 1077-1085.
35. Jang, J. and Pinsky, P.M. *An assumed covariant strain based 9-node shell element*, Int. J. Num. Meth. Eng. **24**(12) (1987), 2389-2411.
36. Khorsandnia, N., Valipour, H., Foster, S. and Crews, K. *A force-based frame finite element formulation for analysis of two-and three-layered composite beams with material non-linearity*, Int. J. NonLinear Mech. **62** (2014), 12-22.
37. Korelc J. and Wriggers, P. *Improved enhanced strain four-node element with Taylor expansion of the shape functions*, Int. J. Num. Meth. Eng. **40**(3) (1997), 407-421.

38. Kwan, A.K.H. *Analysis of buildings using strain-based element with rotational DOFs*, J. Struct. Eng. 118(5) (1992), 1191–1212.
39. Li, L.X., Chen, Y.L. and Lu, Z.C. *Generalization of the multi-scale finite element method to plane elasticity problems*, Appl. Math. Model. 39(2) (2015), 642–653.
40. Logg, A. *Automating the finite element method*, Arch. Comput. Methods Eng. 14(2) (2007), 93–138.
41. Manta, D., Gonçalves, R. and Camotim, D. *Combining shell and GBT-based finite elements: Plastic analysis with adaptive mesh refinement*, Thin-Walled Struct. 158 (2021), 107205.
42. Marras, S., Kelly, J.F., Moragues, M., Müller, A., Kopera, M.A., Vázquez, M., Giraldo, F.X., Houzeaux, G. and Jorba, O. *A review of element-based Galerkin methods for numerical weather prediction: Finite elements, spectral elements, and discontinuous Galerkin*, Arch. Computat. Methods Eng. 23 (4) (2016), 673–722.
43. Meier, C., Popp, A. and Wall, W.A. *Geometrically exact finite element formulations for slender beams: Kirchhoff–Love theory versus Simo–Reissner theory*, Arch. Comput. Methods Eng. 26(1) (2019), 163–243.
44. Messai, A., Belounar, L. and Merzouki, T. *Static and free vibration of plates with a strain-based brick element*, Eur. J. Comput. Mech. (2019), 1–21.
45. Mousa, A. and Djoudi, M. *New strain based triangular finite element for the vibration of circular cylindrical shell with oblique ends*, Int. J. Civ. Environ. Eng. 15(5) (2015), 6–11.
46. Neuenhofer, A. and Filippou, F.C. *Evaluation of nonlinear frame finite-element models*, J. Struct. Eng. 123(7) (1997), 958–966.
47. Neuenhofer, A. and Filippou, F.C. *Geometrically nonlinear flexibility-based frame finite element*, J. Struct. Eng. 124(6) (1998), 704–711.
48. Rebiai, C. *Finite element analysis of 2-D structures by new strain based triangular element*, J. Mech. (2018), 1–9.
49. Rebiai, C. and Belounar, L. *A new strain based rectangular finite element with drilling rotation for linear and nonlinear analysis*, Archives of civil and mechanical engineering, 13(1) (2013), 72–81.
50. Rebiai, C. and Belounar, L. *An effective quadrilateral membrane finite element based on the strain approach*, Measurement, 50 (2014), 263–269.
51. Rebiai, C., Saidani, N. and Bahloul, E. *A New Finite Element Based on the Strain Approach for Linear and Dynamic Analysis*, Research Journal of Applied Sciences, Engineering and Technology, 11(6) (2015), 639–644.

52. Reddy, J.N. *An introduction to the finite element method*, (Vol. 2, No. 2.2). New York: McGraw-hill, 1993.
53. Rezaiee-Pajand, M., Arabi, E. and Moradi, A.H. *Static and dynamic analysis of FG plates using a locking free 3D plate bending element*, J. Braz. Soc. Mech. Sci. Eng. 43(1) (2021), 1–12.
54. Rezaiee-Pajand, M. and Gharaei-Moghaddam, N. *Analysis of 3D Timoshenko frames having geometrical and material nonlinearities*, Int. J. Mech. Sci. 94 (2015), 140–155.
55. Rezaiee-Pajand, M. and Gharaei-Moghaddam, N. *Frame nonlinear analysis by force method*, Int. J. Steel Struct. 17(2) (2017), 609–629.
56. Rezaiee-Pajand, M. and Gharaei-Moghaddam, N. *Vibration and static analysis of cracked and non-cracked non-prismatic frames by force formulation*, Eng. Struct. 185 (2019), 106–121.
57. Rezaiee-Pajand, M. and Gharaei-Moghaddam, N. *Force-based curved beam elements with open radial edge cracks*, Mech. Adv. Mater. Struct. 27(2) (2020), 128–140.
58. Rezaiee-Pajand, M., Gharaei-Moghaddam, N. and Ramezani, M. *Two triangular membrane element based on strain*, Int. J. Appl. Mech. 11(1) (2019), 1950010.
59. Rezaiee-Pajand, M., Gharaei-Moghaddam, N. and Ramezani, M.R. *A new higher-order strain-based plane element*, Scientia Iranica. Transaction A, Civil Engineering, 26(4) (2019), 2258–2275.
60. Rezaiee-Pajand, M., Gharaei-Moghaddam, N. and Ramezani, M., *Strain-based plane element for fracture mechanics' problems*, Theor. Appl. Fract. Mech. 108 (2020), 102569.
61. Rezaiee-Pajand, M., Gharaei-Moghaddam, N. and Ramezani, M., *Higher-order assumed strain plane element immune to mesh distortion*, Eng. Comput. 37(9) (2020), 2957–2981.
62. Rezaiee-Pajand, M., Ramezani, M. and Gharaei-Moghaddam, N. *Using higher-order strain interpolation function to improve the accuracy of structural responses*, Int. J. Appl. Mech. 12(3) (2020), 2050026.
63. Rezaiee-Pajand, M. and Yaghoobi, M. *Formulating an effective generalized four-sided element*, Eur. J. Mech. A Solids, 36 (2012), 141–155.
64. Rezaiee-Pajand, M. and Yaghoobi, M. *A free of parasitic shear strain formulation for plane element*, Research in Civil and Environmental Engineering, 1 (2013) 1–27.

65. Rezaiee-Pajand, M. and Yaghoobi, M. *A robust triangular membrane element*, Lat. Am. J. Solids Struct. 11(14) (2014), 2648–2671.
66. Rezaiee-Pajand, M. and Yaghoobi, M. *An efficient formulation for linear and geometric non-linear membrane elements*, Lat. Am. J. Solids Struct. 11(6) (2014), 1012–1035.
67. Rezaiee-Pajand, M. and Yaghoobi, M. *Two new quadrilateral elements based on strain states*, Civ. Eng. Infrastruct. J., 48(1) (2015), 133–156.
68. Rezaiee-Pajand, M. and Yaghoobi, M. *A hybrid stress plane element with strain field*, Civ. Eng. Infrastruct. J. 50(2) (2017), 255–275.
69. Rezaiee-Pajand, M. and Yaghoobi, M. *An efficient flat shell element*, Meccanica, 53(4-5) (2018), 1015–1035.
70. Rezaiee-Pajand, M. and Yaghoobi, M. *Geometrical nonlinear analysis by plane quadrilateral element*, Scientia Iranica, 25(5) (2018), 2488–2500.
71. Sabir, A.B. *A rectangular and triangular plane elasticity element with drilling degrees of freedom*, In Proceedings of the Second International Conference on Variational Methods in Engineering, Brebbia CA ed., Southampton University (1985), 17–25.
72. Sabir, A.B. and Sfindji, A. *Triangular and rectangular plane elasticity finite elements*, Thin-Walled Struct. 21(3) (1995), 225–232.
73. Saritas, A. and Filippou, F.C. *Inelastic axial-flexure-shear coupling in a mixed formulation beam finite element*, Int. J. Non Linear Mech. 44(8) (2009), 913–922.
74. Spacone, E., Ciampi, V. and Filippou, F.C. *Mixed formulation of nonlinear beam finite element*, Comput. Struct. 58 (1) (1996), 71–83.
75. Tayeh, S.M. *New strain-based triangular and rectangular finite elements for plane elasticity problems*, Thesis, The Islamic University, Gaza, 2003.
76. Taylor, R.L., Filippou, F.C., Saritas, A. and Auricchio, F. *A mixed finite element method for beam and frame problems*, Comput. Mech. 31(1) (2003), 192–203.
77. To, C.W.S. and Liu, M.L. *Hybrid strain based three-node flat triangular shell elements*, Finite Elem. Anal. Des., 17(3) (1994), 169–203.
78. Xu, M., Gitman, I.M. and Askes, H. *A gradient-enriched continuum model for magneto-elastic coupling: Formulation, finite element implementation and in-plane problems*, Comput. Struct. 212 (2019), 275–288.
79. Yang, H.T., Saigal, S., Masud, A. and Kapania, R.K. *A survey of recent shell finite elements*, Int. J. Num. Meth. Eng. 47(1a3) (2000), 101–127.

80. Zienkiewicz O.C. and Taylor R.L. *The finite element method for solid and structural mechanics*, Elsevier, 2005.



Review of the strain-based formulation for analysis of plane structures

Part II: Evaluation of the numerical performance

M. Rezaiee-Pajand*, N. Gharaei-Moghaddam and M. Ramezani

Abstract

In this part of the study, several benchmark problems are solved to evaluate the performance of the existing strain-based membrane elements, which were reviewed in the first part. This numerical evaluation provides a basis for comparison between these elements. Detailed discussions are offered after each benchmark problem. Based on the attained results, it is concluded that inclusion of drilling degrees of freedom and also utilization of higher-order assumed strain field result in higher accuracy of the elements. Moreover, it is evident that imposing the optimal criteria such as equilibrium and compatibility on the assumed strain field, in addition to reducing the number of degrees of freedom of the element, increases the convergence speed of the resulting strain-based finite elements.

AMS subject classifications (2020): 74K15, 74G15.

Keywords: Strain-based formulation; Higher-order strain field; Equilibrium condition; Numerical evaluation; Drilling degrees of freedom.

*Corresponding author

Received 25 October 2020; revised 8 June 2021; accepted 9 June 2021

Mohammad Rezaiee-Pajand

Professor of Civil Engineering, School of Engineering, Ferdowsi University of Mashhad, Iran. e-mail: Rezaiee@um.ac.ir, Tel/fax: +98-51-38412912

Nima Gharaei-Moghaddam

PhD of Structural Engineering, School of Engineering, Ferdowsi University of Mashhad, Iran. e-mail: Nima.Gharaei@gmail.com, Tel: +98-915-1589342

Mohammadreza Ramezani

PhD Student of Structural Engineering, School of Engineering, Ferdowsi University of Mashhad, Iran. e-mail: Mohammadrezaramezani1994@gmail.com, Tel: +98-915-1076010

1 Introduction

Among different formulation methods for the development of membrane finite elements, the assumed strain approach is proved to be very effective in removing problems such as shear parasitic error, mesh sensitivity, and different locking phenomena [28]. Therefore, various authors utilized this scheme to develop strain-based plane elements [6]. These finite elements were reviewed in the first part of this study. The main objective of the second part is to evaluate the numerical performance of the reviewed elements and study the effect of different assumptions and criteria on the performance of assumed strain formulation. For this purpose, the results attained by the reviewed elements for a series of benchmark problems are presented. Based on the obtained results by the reviewed membrane elements, a short discussion is provided after each problem. Moreover, according to the overall outcomes, the existing strain-based plane elements are ranked according to their different advantages and shortcomings. This ranking can be used to detect the most suitable assumptions and configurations to achieve a robust plane finite element. It should be noted that in the present paper, only the performance of the strain-based membrane elements in the analysis of linear problems is investigated. This is mainly because even the finite elements developed for nonlinear applications first should pass the upcoming benchmark tests to be considered as robust and powerful elements. It is also reminded that most of the reviewed research works evaluated the performance of their suggested elements in the analysis of linear problems. However, it is obvious that the reviewed element can also be used for the analysis of nonlinear problems and some of the previously published pursued this issue. The interested readers can refer to references [14, 26] for further information in this regard.

Tables 1 and 2 present a list of the elements used for comparison.

As it can be seen, an abbreviation is used for each element, which is selected based on the following order. The first part of the abbreviation is taken from the authors' names. The second part of the abbreviation starts with a letter that indicates the geometric shape of the element. Accordingly, "T", "Q", and "R" stand for triangular, quadrilateral, and rectangular, respectively. This letter is followed by a number that indicates the number of degrees of freedom. If the drilling degrees of freedom are used in the formulation of an element, then the letter "D" comes after the previously mentioned number. Finally, if two or more elements with the same geometry and number of nodes are proposed by the same authors, then roman numerals distinct those elements. For instance, based on this abbreviation method, the triangular element proposed by Belarbi and Bourezane, which has nine degrees of freedom and includes drilling degrees of freedom is called "BB-T9D", and since two different elements with the same abbreviation in this convention exist, they are distinguished from each other by roman numerals as "BB-T9D-I" and "BB-T9D-II".

Table 1: List of triangular plane elements used for comparison

No.	Abbreviation	Description of the element	Reference
Triangular Elements			
1	S-T9D	Three-node nine-degree of freedom triangular element with drilling proposed by Sabir	[28]
2	SS-T8	Four-node eight-degree of freedom triangular element proposed by Sabir and Sfindji	[29]
3	T-T9D	Three-node nine-degree- of freedom triangular element with drilling proposed by Tayeh	[30]
4	BB-T9D-I	First three-node nine-degree of freedom triangular element with drilling proposed by Belarbi and Bourezane	[2]
5	BB-T9D-II	Second three-node nine-degree of freedom triangular element with drilling proposed by Belarbi and Bourezane	[3]
6	RY-T10	Six-node ten-degree of freedom triangular element proposed by Rezaiee-Pajand and Yaghoobi	[25]
7	RY-T10D	Seven-node ten-degree of freedom triangular element with drilling proposed by Rezaiee-Pajand and Yaghoobi	[26]
8	R-T9D	Three-node nine-degree of freedom triangular element with drilling proposed by Rebiai	[13]
9	RGR-T10	Five-node ten-degree of freedom triangular element proposed by Rezaiee-Pajand et al.	[17]
10	RGR-T10D	Four-node ten-degree of freedom triangular element with drilling proposed by Rezaiee-Pajand et al.	[17]
11	RGR-T11D-I	Seven-node eleven-degree of freedom triangular element with drilling proposed by Rezaiee-Pajand et al.	[18]
12	RGR-T11D-II	Four-node eleven-degree of freedom triangular element with drilling proposed by Rezaiee-Pajand et al.	[19]
13	RGR-T14	Seven-node fourteen-degree of freedom triangular element proposed by Rezaiee-Pajand et al.	[22]

In addition to the reviewed membrane elements, which are formulated by the assumed strain approach, results of three common displacement-based elements namely four-node and eight-node isoparametric quadrilateral elements (Q4 and Q8) and linear strain triangular element (LST) are provided in some problems to compare the performance of the strain-based formulation with them.

Table 2: List of quadrilateral plane elements used for comparison

No.	Abbreviation	Description of the element	Reference
Quadrilateral Elements			
1	SS-R10	Five-node ten-degree of freedom rectangular element proposed by Sabir and Sfindji	[29]
2	T-R12D	Four-node twelve-degree of freedom rectangular element with drilling proposed by Tayeh	[30]
3	BM-R10	Five-node ten-degree of freedom rectangular element proposed by Belarbi and Maalem	[4]
4	RY-Q10	Five-node ten-degree of freedom quadrilateral element proposed by Rezaiee-Pajand and Yaghoobi	[23]
5	RY-R10-I	First five-node ten-degree of freedom rectangular element proposed by Rezaiee-Pajand and Yaghoobi	[24]
6	RY-R10-II	Second five-node ten-degree of freedom rectangular element proposed by Rezaiee-Pajand and Yaghoobi	[24]
7	RB-R12D	Four-node twelve-degree of freedom rectangular element with drilling proposed by Rebiai and Belounar	[14]
8	RB-Q12D	Four-node twelve-degree of freedom quadrilateral element with drilling proposed by Rebiai and Belounar	[15]
9	RSB-Q12D	Four-node twelve-degree of freedom quadrilateral element with drilling proposed by Rebiai et al.	[16]
10	RY-Q14D	Five-node fourteen-degree of freedom quadrilateral element with drilling proposed by Rezaiee-Pajand and Yaghoobi	[27]
11	RY-Q18	Nine-node eighteen-degree of freedom quadrilateral element proposed by Rezaiee-Pajand and Yaghoobi	[8]

2 Numerical evaluation

In this section, several benchmark problems are solved to evaluate the performance of the strain-based elements, which were reviewed in the first part of this study. It should also be noted that in the following benchmark problems, consistent units are used for various quantities. Accordingly, the problems are presented in a dimensionless format. Moreover, it should be noted that except for the elements proposed by the authors themselves, the results of the other elements are taken from the related references, and many of the reviewed references did not report the results for some of the following problem. Therefore, in some problems, the results of some elements are not reported.

2.1 Cantilever beam with distorted mesh

One of the available tests to examine the performance of the membrane elements in the coarse distorted meshes, under both bending and shear loadings, is the cantilever beam, which is depicted in Figure 1 [6, 26].

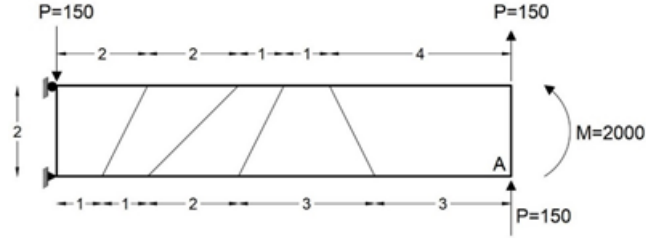


Figure 1: Cantilever beam with distorted quadrilateral mesh

This figure illustrates the geometric characteristics, loading, and utilized meshes for quadrilateral elements. The modulus of elasticity and Poisson's ratio of this beam are 1500 and 0.25, respectively, and its thickness is equal to 1. The utilized mesh for analysis using triangular elements is demonstrated in Figure 2. As it is evident, each quadrilateral element is divided by a dashed line into two triangular elements.



Figure 2: Triangular mesh for analysis Cantilever beam with distorted mesh

The analytical vertical displacements at point A under the shear and bending loadings are equal to 102.60 and 100, respectively. The attained results by Q4, Q8, and other strain-based elements are listed in Table 3. In fact, this test measures the performance of different elements for the analysis of structures with distorted meshes under bending and shear loading conditions. According to the results, almost all the strain-based elements, except T-T9D, provide acceptable accuracy. The most accurate quadrilateral element is RY-Q10 and BM-R10. Among the triangular elements, RGR-T10D, RGR-T11D-I, and RGR-T11D-II show the highest accuracy. An interesting finding is that, in general, the accuracy of the strain-based elements under flexural loading is higher. However, there are exceptions such as RY-Q18. Another important finding is the unexpectedly poor performance of T-T9D, which is the second weakest element after Q4. The attained results by RGR-T10 and RGR-T10D, which have the same assumed strain field and their differences are only in distribution and type of degrees of freedom, verify this

conjecture that inclusion of drilling degrees of freedom in the plane elements, improves their accuracy under in-plane bending.

Table 3: Deflection of point *A* of the cantilever beam with distorted mesh

	Element	Load P		Load M	
		Displacement	Relative Error (%)	Displacement	Relative Error (%)
Quadrilateral elements	Q4	50.70	50.58	45.70	54.30
	Q8	101.50	1.07	99.70	0.30
	SS-R10	97.91	4.51	98.57	1.43
	T-R12D	93.28	9.08	96.11	3.89
	BM-R10	101.77	0.81	99.93	0.07
	RY-Q10	102.79	0.18	100.00	0.00
	RB-R12D	98.83	3.67	97.30	2.70
	RB-Q12D	99.35	3.17	99.19	0.81
	RY-Q14D	104.16	1.52	101.66	1.66
	RY-Q18	103.52	0.89	101.48	1.48
Triangular elements	LST	101.05	1.51	98.30	1.70
	S-T9D	100.08	2.45	97.82	2.18
	SS-T8	100.89	1.67	98.36	1.64
	T-T9D	79.87	22.15	83.05	16.95
	RY-T10D	100.58	1.96	100.00	0.00
	R-T9D	100.98	1.57	99.86	0.14
	RGR-T10	103.65	1.02	98.50	1.50
	RGR-T10D	101.83	0.75	100.00	0.00
	RGR-T11D_I	103.92	1.29	100.70	0.70
	RGR-T11D_II	101.58	0.99	100.99	0.99
	RGR-T14	103.72	1.09	101.03	1.03
	Analytical Solution	102.60	-	100.00	-

2.2 Cantilever beam under parabolic shear loading

To investigate the performance of the elements in the analysis of structures under distributed surface traction, the cantilever beam demonstrated in Figure 3 is analyzed [5, 12, 26].

This beam is made of an elastic material with the modulus of elasticity and Poisson's ratio equal to 3000 and 0.25, respectively and its thickness is taken one unit. The beam is loaded by the parabolic distributed traction at its free end, which is equal to 40 units. This benchmark problem also evaluates the efficiency of elements in the analysis of structures using coarse meshes. As it is evident in Figure 3, the beam is discretized by four quadrilateral elements.

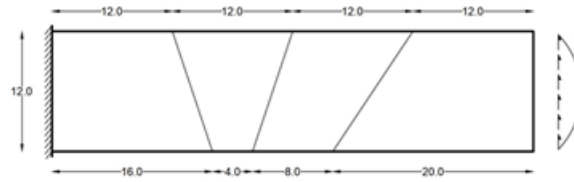


Figure 3: Cantilever beam under parabolic shear loading

In the case of triangular elements, eight elements are used, which the utilized mesh is demonstrated in Figure 4. However, results of some of the reviewed elements are reported for the regular mesh.

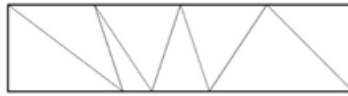


Figure 4: Triangular mesh for analysis Cantilever beam under parabolic shear loading

Table 4 presents the obtained responses by the mentioned membrane elements for deflections at the tip of the beam. Felippa reported the near-exact tip deflection of the beam equal to 0.35601 [7].

Based on the reported results, RGR-T10 and RY-Q14D are the most accurate elements in this problem with only 0.03 percent error in their estimations. Similar to the previous problem, T-T9D has the worst performance with 25 percent error, and again the RY-Q10 is among the most accurate quadrilateral elements. As it can be seen, RGR-T14 is among the most accurate elements. This was expected, since as mentioned in the respective reference, an important feature of the complete second-order assumed strain-fied is its ability in providing accurate responses for the problems with distributed loading [22].

2.3 Cook's skew beam

Cook trapezoidal beam is one of the most fundamental tests for checking shear displacements in non-rectangular geometry [6]. Figure 5 demonstrates this beam under uniformly distributed tip loading. This beam has a unit thickness and is made of a material whose Young's modulus and Poisson's ratio are 1 and $\frac{1}{3}$, respectively.

Many researchers also implement this benchmark to challenge the convergence of their elements. Here, four different meshes, namely 2×2 , 4×4 , 8×8 , and 16×16 , are used. These meshes are demonstrated in Figure 6. The

Table 4: Tip deflection of cantilever beam under parabolic shear

	Element	Vertical Dis- placement	Relative Error (%)
Quadrilateral elements	Q4	0.21290	40.20
	Q8	0.34790	2.28
	SS-R10	0.34070*	4.30
	T-R12D	0.31328	12.00
	BM-R10	0.34604*	2.80
	RY-Q10	0.35280	0.90
	RY-R10-I	0.32724*	8.08
	RY-R10-II	0.33027*	7.23
	RB-R12D	0.34120*	4.16
	RSB-Q12D	0.33470*	5.99
	RY-Q14D	0.35590	0.03
	RY-Q18	0.35230	1.04
Triangular elements	LST	0.34770	2.33
	T-T9D	0.26701	25.00
	BB-T9D-I	0.27822*	21.85
	RY-T10	0.35031*	1.60
	RY-T10D	0.34680	2.59
	RGR-T10	0.35610	0.03
	RGR-T10D	0.34680	2.59
	RGR-T11D-I	0.35850	0.70
	RGR-T11D-II	0.35713	0.31
	RGR-T14	0.35555	0.13
Near-exact solution		0.35601	

* The results are attained from a regular mesh

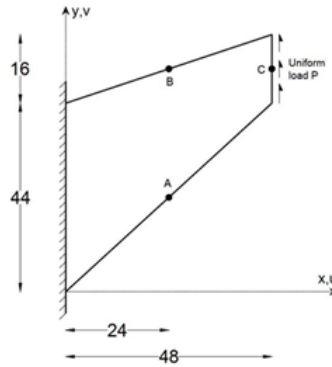


Figure 5: Cook's skew beam

results of the point C deflection are presented in Table 5. It should be noted that the near-exact solution for this problem is reported equal to 23.96 [21].

Outcomes of this problem are again in complete agreement with the findings of previous numerical examples, and once more, the RGR-T11D-I and RGR-T11D-II are among the best-performing elements. The other elements, which provide accurate estimations, are RY-Q10, RGR-T10, and R-T9D. It is somehow unexpected that R-T9D can compute a very accurate response by a coarse 4×4 mesh. One of the elements that have relatively fast convergence is RGR-T14. As it is evident, the convergence trend of different elements is not similar. While most of the elements converge to the exact response asymptotically from below, the RGR-T11D-I element approaches the accurate response from above. Also, there are elements, such as RY-T10 and RY-T10D, which show non-uniform convergence behavior, and even RY-Q14D goes beyond the response. Nevertheless, most of the strain-based elements demonstrate reasonable accuracy and convergence in this benchmark problem.

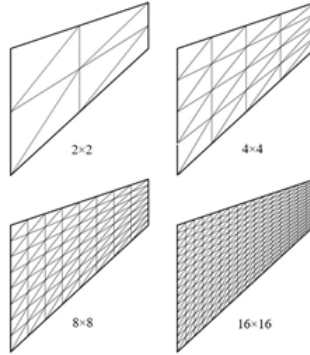


Figure 6: Utilized meshes for analysis of Cook's skew beam

2.4 Thick curved beam

To appraise the ability of finite elements, especially triangular ones, in the analysis of structures with curved geometry, many of the previous researchers have evaluated the performance of their proposed element in solving the curved beams, which is demonstrated in Figure 7 [5, 26, 32]. This beam is loaded by the shear load $P = 600$ at its tip.

The module of elasticity, poisson's ratio, and thickness of this beam are 1000, 0, and 1, respectively. As depicted in Figure 7, four quadrilateral elements are used to mesh this structure. In the case of triangular elements, eight elements are used as demonstrated in Figure 8.

Table 5: Deflection of point C of the Cook's beam

		Mesh			
	Element	2×2	4×4	8×8	16×16
Quadrilateral elements	Q4	11.80	18.29	22.08	23.43
	SS-R10	17.06	30.64	30.64	30.65
	T-R12D	14.85	17.25	19.88	21.80
	RY-Q10	25.65	24.27	24.01	23.96
	RB-Q12D	17.87	23.37	23.38	23.50
	RY-Q14D	27.61	30.48	31.85	32.44
	RY-Q18	23.45	23.70	23.86	23.92
Triangular elements	S-T9D	18.25	20.32	22.18	22.18
	SS-T8	17.86	20.15	21.21	21.46
	T-T9D	12.45	15.09	18.44	20.13
	BB-T9D-I	18.52	21.36	22.45	23.69
	BB-T9D-II	18.58	23.88	23.88	23.88
	RY-T10	20.94	23.84	24.18	24.13
	RY-T10D	25.82	27.19	27.23	27.09
	R-T9D	18.78	23.94	23.94	23.94
	RGR-T10	21.18	23.03	23.69	23.95
	RGR-T10D	19.06	22.85	23.14	23.87
	RGR-T11D-I	26.00	24.39	24.01	23.97
	RGR-T11D-II	23.37	23.42	23.93	23.97
	RGR-T14	23.64	23.73	23.85	23.96
Near-exact Solution		23.96			

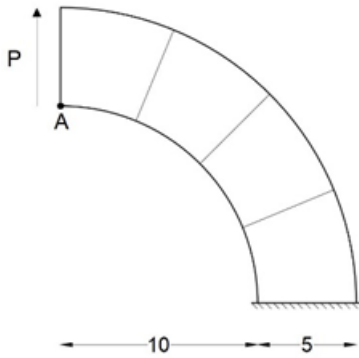


Figure 7: Thick curved beam with quadrilateral mesh

The exact vertical displacement of point A under the applied load is equal to 90.10. The attained results by different elements are presented in Table

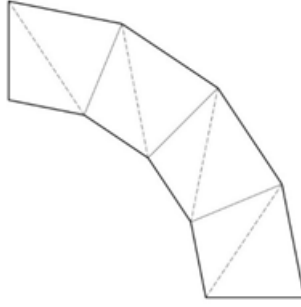


Figure 8: The triangular mesh for analysis of thick curved beam

6. It is evident that the RGR-T11D-I element provides the most accurate estimation with only 0.24% error. After this element, RGR-T10 with the relative error of 0.79% is in the second place. It is interesting to note that among the quadrilateral elements, the performance of Q8 is better than the strain-based elements. Nonetheless, the error of most of the strain-based elements is less than 5 percent, which for the utilized coarse mesh is negligible by any set of standards. This problem shows that the elements formulated by the assumed strain approach are a suitable option for efficient analysis of curved structures, and can compete with isoparametric elements in terms of accuracy and convergence.

Table 6: Deflection of point A of thick curved beam

		Load P		
	Element	Vertical placement	Dis- placement	Relative Error (%)
Quadrilateral elements	Q8	88.60		1.66
	SS-R10	98.71		9.56
	RY-Q10	86.92		3.53
	RY-Q14D	87.00		3.44
	RY-Q18	86.45		4.05
Triangular elements	RY-T10	87.15		3.27
	RY-T10D	87.47		2.92
	RGR-T10	89.39		0.79
	RGR-T10D	84.62		6.08
	RGR-T11D-I	89.88		0.24
	RGR-T11D-II	88.30		2.00
	RGR-T14	83.79		7.00
Analytical Solution		90.10		

2.5 Thin curved beam

To investigate the effect of the shear lock-in curved structures and also the convergence rate to achieve the precise response, a thin curved beam test is available. The modulus of elasticity, Poisson's ratio, and thickness of this structure, which is demonstrated in Figure 9 are 10^7 , 0.25, and 0.1, respectively [31, 32]. This beam is loaded by a unit vertical force at its tip.

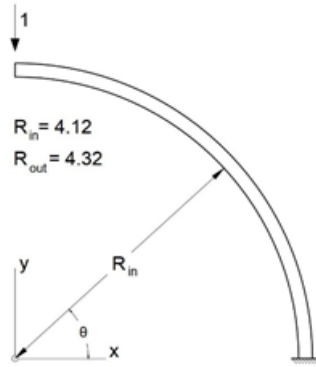


Figure 9: Thin curved beam

Three different meshes are used to analyze this structure, namely 1×6 , 2×12 , and 4×24 . These meshes are named based on the number of quadrilateral elements used in them. Needless to say, for analysis using triangular elements, each quadrilateral element is divided into two triangular elements. For instance, 1×6 is demonstrated in Figure 10.

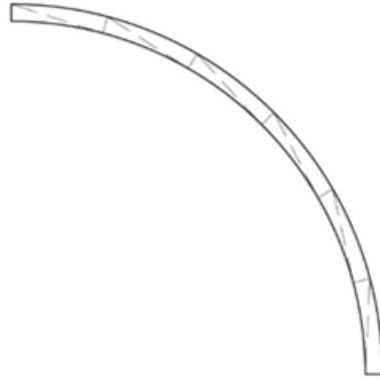


Figure 10: The used 1×6 mesh for analysis of thin curved beam

The main purpose of solving this problem is to compute the tip deflection of the beam under applied load and therefore, investigate the effect of the

locking problem on the performance of the strain-based elements. The exact vertical displacement at the tip is reported to be equal to 0.08734[23]. Table 7 presents the obtained results by some of the strain-based elements.

Table 7: Deflection of point A of thin curved beam

Element		Mesh					
		1×6		2×12		4×24	
		Deflection	Relative Error (%)	Deflection	Relative Error (%)	Deflection	Relative Error (%)
Quadrilateral elements	RY-Q10	-0.08901	1.91	-0.08844	1.26	-0.08846	1.28
	RY-Q14D	-0.08748	0.16	-0.08898	1.87	-0.08925	2.19
	RY-Q18	-0.08745	0.12	-0.08840	1.21	-0.08850	1.33
Triangular elements	RY-T10	0.05634	35.49	0.08491	2.78	0.08815	0.93
	RGR-T10	-0.06305	27.81	-0.08493	2.76	-0.08609	1.43
	RGR-T10D	-0.06486	25.74	-0.08501	2.67	-0.08650	0.96
	RGR-T11D	-0.08291	5.07	-0.08434	3.43	-0.08691	0.49
	RGR-T11D	-0.08265	5.36	-0.08656	0.89	-0.08622	1.28
	RGR-T11D	-0.08712	0.25	-0.08713	0.24	-0.08728	0.07
Analytical Solution		-0.08734					

It is evident that the mentioned triangular elements, except the RGR-T11D-I and II and RGR-T14, face the locking problem in the coarsest mesh and behave too stiffly. In contrast, these elements provide an acceptable response. In the coarsest mesh, these elements do not lock and have a maximum error of 5.36%. This error reduces to 0.07% in the finest mesh. It should be noted that the quadrilateral elements provide more accurate estimations in the coarse mesh. However, in the case of the finest utilized mesh, they tend to become a bit more flexible and therefore, predict responses higher than the exact values.

2.6 McNeal's beam

McNeal and Harder proposed this benchmark to examine the sensitivity of the elements to the mesh distortion and the trapezoidal locking phenomenon [9]. The geometry of this beam and the rectangular, parallelogram, and trapezoidal meshes used for analysis by quadrilateral elements are depicted in Figure 11. The utilized meshes for triangular meshes are demonstrated in Figure 12.

Modulus of elasticity, Poisson's ratio, and thickness of the structure are 10^7 , 0.3, and 0.1, respectively. Two modes of loading are assumed, as depicted in Figure 10. The derived responses by the strain-based elements are listed in Table 8. This test is a difficult problem for many of the displacement-based membrane elements since they demonstrate high sensitivity to the trapezoidal meshes. For example, the powerful Q8 element with all of its capabilities

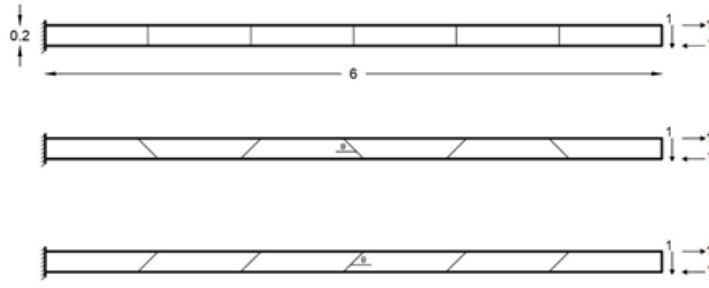


Figure 11: McNeal's beam and utilized quadrilateral meshes

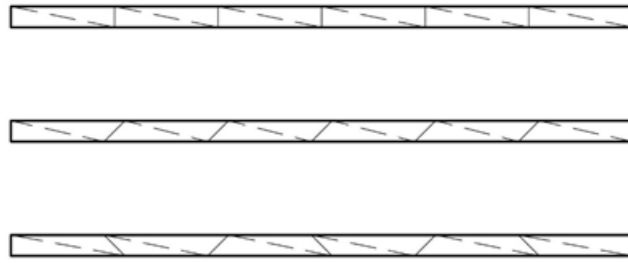


Figure 12: The utilized triangular meshes for analysis of McNeal's beam

faces fatal error for both modes of loading in trapezoidal mesh. However, as it is evident from the results presented in Table 7, most of the strain-based elements have no problem in this case. Although SS-R10 and S-T9D are exceptions, they suffer from trapezoidal locking severely. It is interesting to note that the RGR-T11D-II provides very accurate estimations for the shear loading without any problem due to locking, while most of the other elements face the trapezoidal locking under shear loading. In the flexural loading, RGR-T14 can capture the exact response in all the utilized meshes.

2.7 Higher-order patch test

The beam, which is demonstrated in Figure 13, is the next numerical example that evaluates the performance of plane strain-based elements.

This beam, which has a geometric ratio of 10, is made of the elastic material with a modulus of elasticity and Poisson's ratio equal to 100 and 0, respectively. The thickness of the beam is taken as 1. Two different types of meshes, namely regular and distorted, which are demonstrated in Figure 14, are used.

Table 8: Normalized tip deflection of the McNeal's beam

	Element	Load P			Load M		
		Rectang	Parallelo	Trapezoi	Rectang	Parallelo	Trapezoi
		ular	gram	dal	ular	gram	dal
		mesh	mesh	mesh	mesh	mesh	mesh
Quadrilatera elements	Q4	9.30	3.58	3.06	9.34	3.14	2.21
	Q8	95.12	91.94	85.43	100.00	75.94	9.32
	SS-R10	4.62	3.61	0.00	11.77	10.07	0.37
	RY-Q10	99.30	99.42	99.42	100.00	100.00	100.00
	RB-Q12D	99.26	98.69	98.78	99.63	99.26	99.26
	RSB-Q12D	100.00	97.59	97.78	100.00	98.89	98.89
	RY-Q14D	98.33	98.74	98.79	98.88	99.11	99.19
	RY-Q18	100.00	100.00	100.00	100.00	100.00	100.00
Triangular elements	LST	98.3	97.05	96.12	99.34	99.40	99.22
	S-T9D	4.75	3.63	0.05	11.82	10.13	0.04
	BB-T9D-I	94.42	87.40	83.35	94.83	94.42	95.21
	BB-T9D-II	96.40	95.04	98.82	98.94	98.79	98.81
	RY-T10	99.44	94.30	92.11	100.00	100.00	100.01
	RY-T10D	99.43	94.94	92.31	100.00	100.00	100.00
	R-T9D	99.63	97.87	97.87	99.62	99.25	99.25
	RGR-T10	99.41	99.52	99.92	100.00	99.95	100.00
	RGR-T10D	99.33	94.12	90.56	100.00	99.98	100.00
	RGR-T11D-I	104.34	102.48	104.99	100.79	100.56	100.94
	RGR-T11D-II	100.00	100.00	100.30	107.40	108.80	106.90
	RGR-T14	0.994	0.995	0.995	100.00	100.00	100.00
Analytical Solutions			0.1081			0.0054	

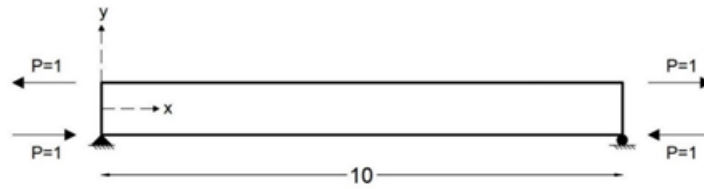


Figure 13: Higher-order patch test

This test examines the performance of the elements under pure bending and considering the simple support conditions. The attained results by the strain-based elements are listed in Table 9. It is evident that all of the elements can compute the exact response regardless of the utilized mesh.

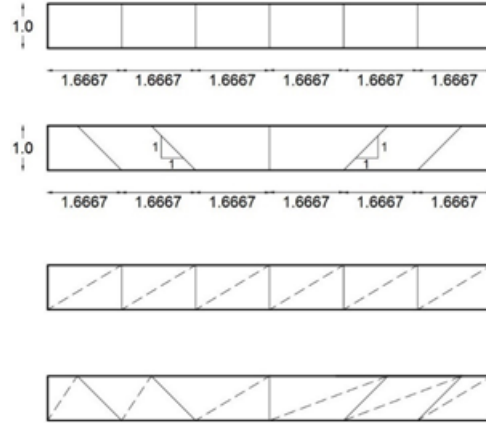


Figure 14: Utilized regular and distorted meshes

Table 9: Maximum displacements of the higher-order patch test

		Regular mesh		Distorted mesh	
Element		Max U	Max V	Max U	Max V
Quadrilateral elements	RY-Q10	-0.600	1.500	-0.600	1.500
	RY-R10-I	-0.600	1.500	-0.600	1.500
	RB-R12D	-0.600	1.500	-0.600	1.500
	RB-Q12D	-0.594	1.493	-0.592	1.484
	RSB-Q12D	-0.590	1.500	-0.590	1.490
	RY-Q14D	-0.600	1.500	-0.600	1.500
	RY-Q18	-0.600	1.500	-0.600	1.500
Triangular elements	RY-T10D	-0.600	1.500	-0.600	1.500
	RGR-T10	-0.600	1.500	-0.600	1.500
	RGR-T10D	-0.600	1.500	-0.600	1.500
	RGR-T11D-I	-0.600	1.500	-0.600	1.500
	RGR-T11D-II	-0.600	1.500	-0.600	1.500
	RGR-T14	-0.600	1.500	-0.600	1.500
Analytical Solution		-0.600	1.500	-0.600	1.500

2.8 Thick-walled cylinder

The cylindrical plane strain test of the thick wall under uniform internal pressure is the eighth problem, which investigates the effect of the Poisson's locking on the performance of strain-based elements [1]. Due to symmetry, only a quarter of this cylinder will be analyzed. This structure and utilized mesh are depicted in Figure 15.

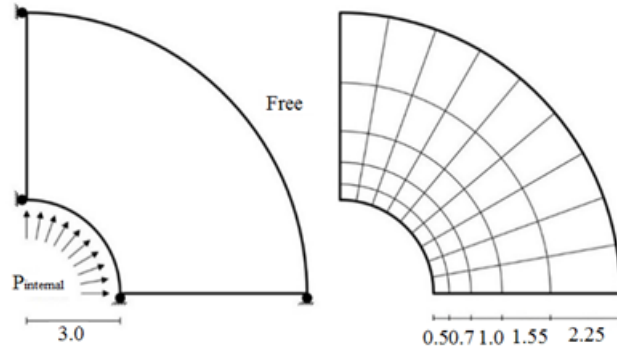


Figure 15: Thick-walled cylinder and used mesh

The elastic modulus of the material is 1000, and it is solved for different values of Poisson's ratio varying from 0.3 to 0.4999. The derived results by different elements are presented in Table 10. According to the outcomes, the assumed strain approach results in elements free from the Poisson's locking.

Table 10: Normalized radial displacement of the thick-walled cylinder at the inner radius

Element		Poisson's ratio			
		0.3	0.49	0.499	0.4999
Quadrilateral elements	RY-Q10	0.9799	0.9789	0.9790	0.9794
	RY-Q14D	1.1805	1.1839	1.1841	1.1846
	RY-Q18	0.9360	0.9576	0.9593	0.9599
Triangular elements	BB-T9D-I	0.9743	-	-	-
	RGR-T11D-I	1.01869	1.0356	1.0361	1.0365
	RGR-T11D-II	1.02838	1.04484	1.04545	1.04604
	RGR-T14	1.07564	1.07724	1.07726	1.07527
Analytical Solution [12]		0.00506	0.00506	0.00504	0.00458

2.9 Theoretical slender beam

The beam depicted in Figure 16, with a length of 100 is made of an elastic material with Young's modulus and Poisson's ratio of 10^6 and 0.3, respectively. This structure is used to investigate the shear effect on the slender plane problems. This structure is analyzed using two different meshes. The obtained results for tip displacements of the beam are listed in Table 11. RGR-T10 has the best performance among the reported elements. It is evi-

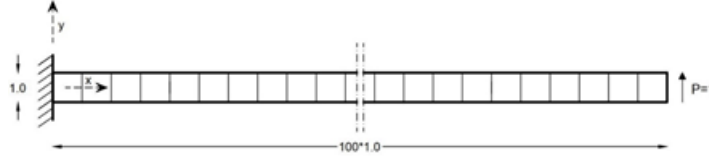


Figure 16: Extremely slender cantilever beam

dent that Q4 suffers from the locking problem and therefore, cannot compute the exact response even using a fine mesh.

Table 11: Tip displacements of slender cantilever beam

Element		Mesh	Displacements	
			$U_x \times 100$	U_y
Quadrilateral elements	Q4	1×100	2.0222	2.6965
		2×200	2.1280	2.8371
	RY-Q10	1×100	3.0046	4.0067
		2×200	2.9991	3.9982
	RY-R10-I	1×100	3.0046	4.0067
		2×200	2.9991	3.9982
	RY-R10-II	1×100	3.0000	4.0002
		2×200	2.9987	3.9976
	RY-Q14D	1×100	3.0000	4.0067
		2×200	3.193	4.2581
Triangular elements	RY-Q18	1×100	2.9983	3.9967
		2×200	2.9989	3.9980
	RY-T10	1×100	3.0000	4.0001
		2×200	2.9992	3.9986
	RGR-T10	1×100	3.0000	4.0000
		2×200	3.0000	4.0000
	RGR-T10D	1×100	2.9845	3.9767
		2×200	2.9944	3.9975
	RGR-T11D-I	1×100	3.0001	4.0003
		2×200	3.0001	4.0001
RGR-T11D-II	1×100	3.0002	4.0002	
	2×200	3.0001	4.0000	
	RGR-T14	1×100	3.0012	4.0131
2×200		3.0007	4.0043	
Analytical Solution			3	4

2.10 Cantilever beam with distortion parameter

A distorted mesh is a finite element mesh that some of its elements deviate vastly from the equilateral triangle and symmetric quadrilateral shapes. To study the influence of the distortion on the behavior of the strain-based elements and prove their superiority in comparison with displacement-based elements, the beam showed in Figure 17 is analyzed by using two quadrilateral or four triangular elements [6].

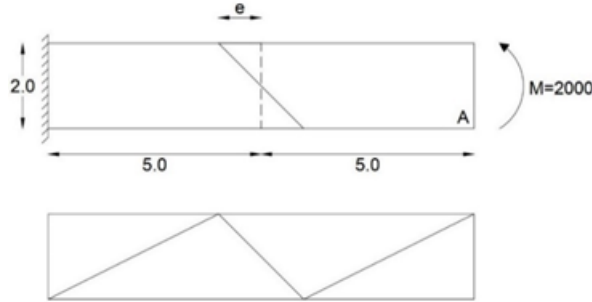


Figure 17: Cantilever beam with distortion parameter and utilized meshes

Table 12: Tip deflection of the cantilever beam with distortion parameter

Element		E						
		0	0.5	1	2	3	4	4.9
Quadrilateral elements	Q4	28.00	21.00	14.10	9.70	8.30	7.20	6.20
	Q8	100.00	99.90	99.30	89.39	59.70	32.01	-
	RY-Q10	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	RY-Q14D	99.80	100.00	100.10	100.70	101.20	102.8	-
	RY-Q18	96.60	97.60	98.50	100.4	105.30	116.8	-
Triangular elements	S-T9D	45.08	45.33	45.84	47.96	49.15	49.47	-
	BB-T9D-I	96.02	96.60	97.04	97.40	97.26	96.90	-
	BB-T9D-II	96.02	96.60	97.04	97.40	97.26	96.90	-
	RY-T10D	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	R-T9D	100.00	97.72	98.15	98.64	99.20	98.76	-
	RGR-T10	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	RGR-T10D	100.00	100.00	100.00	100.00	100.00	100.00	100.00
	RGR-T11D-I	99.96	99.98	99.94	99.96	99.95	99.89	99.91
	RGR-T11D-II	100.00	100.00	100.00	100.00	99.95	99.91	99.89
	RGR-T14	100.00	100.00	100.00	100.00	104.90	114.70	114.73
Analytical Solution		100						

The beam is made of a material with a modulus of elasticity and Poisson's ratio equal to 1500 and 0.25, respectively and its thickness is taken equal to 1 unit. A distortion parameter, e , controls the shape of the elements. The thickness of the beam is taken equal to 1. This beam is reanalyzed by

increasing distortion parameter, and the attained results for tip deflection are listed in Table 12. As it can be seen, the strain-based elements are completely insensitive to the mesh distortion, and increasing the distortion parameter has no remarkable effect on their performance, while the accuracy of Q4 and Q8 diminishes rapidly by the increase in the distortion parameter. Another interesting finding of this numerical example is the poor performance of S-T9D, which is one of the first suggested strain-based elements.

2.11 Cantilever shear wall

An important purpose of formulating efficient elements is to analyze practical structures with coarser meshes and consequently fewer degrees of freedom. Therefore, in order to investigate the efficiency of the strain-based elements in practical problems, two shear walls are examined with the strain-based elements. In the first problem, the shear wall shown in Figure 18 is analyzed [24].

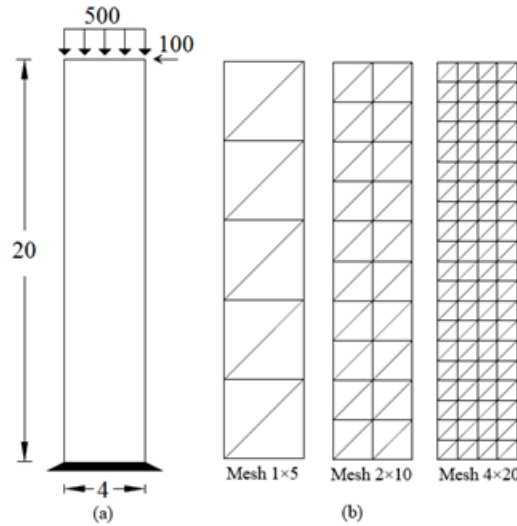


Figure 18: The shear wall and the utilized meshes

The modulus of elasticity and Poisson's ratio of the wall are 2×10^7 and 0.2, respectively. Here, to reevaluate the accuracy and efficiency of strain formulation, the conventional element Q8 is brought for comparison. Furthermore, to investigate the convergence, two finer meshes have been used. The normalized responses are provided in Table 13.

Based on the results presented in Table 13, the RGR-T14 element demonstrates the best performance among the compared elements. Two interesting

outcomes are the lower accuracy of Q8 and the inability of RY-Q14D, which becomes too flexible when using finer meshes. As it can be seen, all the reported strain-based elements except RGR-T10D have less than 5 percent error in their estimations when a coarse 1×5 meshes are used. Once again, this finding demonstrates the high efficiency of the assumed strain approach.

Table 13: Tip deflection of the cantilever beam with distortion parameter

Element		Mesh		
		1×5	2×10	4×20
Quadrilateral elements	Q8	62.17	80.10	89.17
	RY-R10-I	95.91	97.13	98.24
	RY-R10-II	95.87	96.99	98.19
	RY-Q14D	95.86	127.16	138.61
	RY-Q18	96.23	97.04	97.76
	RY-T10	96.86	97.53	98.35
Triangular elements	RGR-T10	96.62	97.78	98.12
	RGR-T10D	89.60	95.63	95.89
	RGR-T11D-I	96.21	98.56	99.01
	RGR-T11D-II	98.01	98.86	99.45
	RGR-T14	98.85	99.14	99.76
Near-exact solution		0.002570		

2.12 Coupled shear walls

In the last numerical example, two coupled shear walls are analyzed to study the performance of the elements in the presence of opening. This structure, which is demonstrated in Figure 19, is made of the elastic material with modulus of the elasticity and Poisson's ratio equal to 2×10^7 and 0.2, respectively [11].

The thickness of this structure is assumed 0.4. Lateral loads with an intensity of $P = 500$ are applied to each story level of the left shear wall. The structure is analyzed using two meshes consisting of 48 and 192 quadrilateral elements (96 and 384 triangular elements). To achieve a near-exact solution, the coupled wall is analyzed using 26880 eight-node isoparametric elements (Q8). The obtained results for lateral displacements at different story levels are reported in Table 14. It is evident that the RGR-T11D-II element provides the most accurate estimations. Based on the reported results for Q8 element, most of the strain-based membrane elements are more accurate and efficient. However, there is an exception about RY-Q14D, which becomes too flexible by using finer meshes and fails to converge to the exact response.

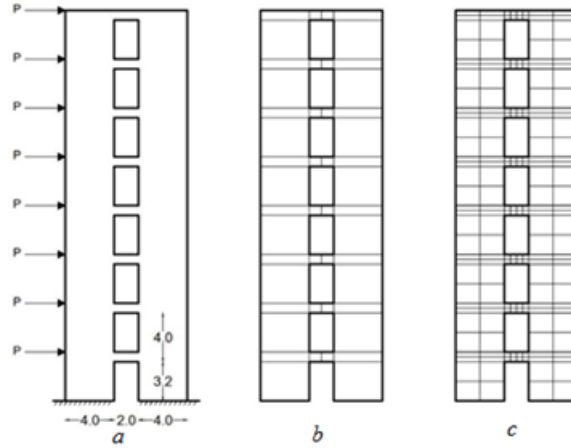


Figure 19: The Coupled shear wall and the utilized meshes a) applied lateral load b) coarse mesh with 48 elements c) fine mesh with 192 elements

3 Discussion

The performance of the existing strain-based plane elements reviewed in the first part of this study was evaluated using a series of benchmark problems in the previous section. First, a cantilever beam with distorted mesh was analyzed. The attained results showed low sensitivity of strain-based elements to mesh distortion compared to the classical displacement-based element, such as, Q4, Q8, and LST. Based on the reported results, the triangular elements are less sensitive than quadrilateral ones. In the next problem, the performance of the strain-based elements in the analysis of structures under distributed surface tractions with coarse mesh was evaluated. Once again, the superior performance of strain-based formulation in comparison with the displacement-based approach is demonstrated. It is also found that the higher-order elements provide better responses than others. However, the part of this better performance can be attributed to the larger number of degrees of freedom. To test the convergence trend of the elements, Cook's skew beam was analyzed using different plane elements. The derived results proved faster convergence of strain-based elements. However, their convergence trend is not uniform, that is, some elements converge to the exact solution from below and some other approaches the exact response from above.

The next two problems were devoted to assessing the performance of strain-based membrane elements in the analysis of structures with curved geometry. As it was expected, the triangular elements demonstrate better accuracy and faster convergence. It should be noted that some of the quadrilateral elements provided more accurate estimations than triangular ones in

Table 14: Lateral story displacements of the coupled shear wall

Element		Lateral displacement					
		Number of elements	Number of degrees-of-freedom	Story 2	Story 4	Story 6	Story 8
Quadrilateral elements	Q8	48	440	0.56	1.53	2.59	3.64
		192	1348	0.68	1.82	3.02	4.16
	RY-R10-I	48	264	0.77	2.07	3.40	4.71
		192	844	0.78	2.07	3.44	4.71
	RY-R10-II	48	216	0.69	1.88	3.13	4.28
		192	668	0.74	2.00	3.32	4.65
	RY-Q14D	48	348	0.90	2.62	4.61	6.63
		192	962	1.14	3.22	5.49	7.70
	RY-Q18	48	540	0.76	2.03	3.36	4.61
		192	1700	0.80	2.13	3.51	4.81
Triangular elements	RY-T10	96	402	0.71	1.92	3.18	4.38
		384	1272	0.80	2.12	3.50	4.79
	RGR-T10	96	396	0.76	2.03	3.29	4.54
		384	1252	0.85	2.26	3.63	4.96
	RGR-T10D	96	348	0.73	1.94	3.19	4.45
		384	1018	0.82	2.14	3.55	4.86
	RGR-T11D-I	96	530	0.75	2.07	3.26	4.63
		384	1800	0.83	2.25	3.56	5.02
	RGR-T11D-II	96	444	0.78	2.15	3.35	4.66
		384	1442	0.88	2.31	3.67	5.19
	RGR-T14	96	732	0.69	1.96	3.05	4.18
		384	2404	0.85	2.21	3.48	4.99
Near-exact solution				0.90	2.38	3.91	5.35

the coarse mesh. However, in the case of the finest utilized mesh, they tend to become a bit more flexible and therefore, predict responses higher than the exact values. To show the insensitivity of the strain-based formulation to trapezoidal locking, the McNeal's beam was analyzed. In fact, the trapezoidal locking is generally a problem for quadrilateral displacement-based elements, such as, Q4 and Q8. Once more, utilization of strain-based quadrilateral elements removes this problem and results in highly accurate responses irrespective of the mesh type. Another problem, which tested the performance of the strain-based elements with respect to mesh distortion was the higher-order patch test. The results of this numerical test proved considerable the insensitivity of the strain-based element to mesh distortion. The effect of distortion extent on the accuracy of the element responses was also investigated in the tenth studied problem. In this part, a cantilever beam loaded with a bending moment at its free end was reanalyzed considering different distorted meshes. Based on the attained results, elements, such as, RY-Q10, RY-T10D,

RGR-T10, and RGR-T10D are completely insensitive to the mesh distortion irrespective of its extent. The other element, however, showed some deviation from the exact responses by introducing severe mesh distortion.

Another problem that occurs for the classical plane elements is the Poisson's locking phenomenon, in which the finite elements face difficulty in predicting accurate responses for the structures made of nearly incompressible material. Solving a thick-walled cylinder under internal pressure for different values of Poisson's ratio, it is shown that higher-order strain-based elements are free from this locking phenomenon. To assess the influence of shear loading of the responses of strain-based elements for slender structures, a theoretically very slender cantilever beam with two different meshes was analyzed. Again, the elements such as RGR-T10, which the equilibrium conditions were applied on their assumed strain field, provided the most accurate estimations.

Finally, two problems tested the ability of the reviewed strain-based elements in the analysis of practical problems. For this purpose, two multistory single and coupled shear wall structures were analyzed to study the convergence and numerical efficiency of the strain-based formulation. The results of the single shear wall showed the fast convergence, as well, high accuracy of the strain-based element in coarse meshes compared to the classical elements. The coupled shear wall test provided a rough measure for evaluating the numerical efficiency of the studied elements by comparing the accuracy of the responses, as well as, the total number of degrees-of-freedom for two different types of meshes. It should be noted that by efficiency, the authors mean the number of elements and degrees of freedom required for a specific level of accuracy. From the numerical results in section 2, it is evident the strain-based elements provide enough accurate estimations with coarser meshes, in comparison with the classical displacement-based elements. However, to achieve a better judgment about the efficiency of the elements, the issue of computational time should also be investigated, which is not pursued in the present study and require further investigation in future research works.

4 Conclusion

Based on the performed review, many of the existing strain-based membrane elements were formulated by using linear assumed strain fields. On the other hand, most of the limited elements with higher-order strain fields were developed using incomplete higher-order polynomials, which do not provide any clear justification for the selected polynomial terms. Another interesting finding from the first part of this study was that in many of the available plane elements, the equilibrium criterion is not imposed on the assumed strain field. Moreover, it was shown that the inclusion of drilling degrees of freedom would improve the performance of resulting elements under in-plane bending. In

this part, several well-known benchmark problems were solved using the existing strain-based membrane elements and common displacement-based elements such as Q4, Q8, and LST. The obtained results clearly demonstrated the superiority of the strain-based formulation in accuracy and efficiency against displacement-based membrane elements. Various problems such as mesh sensitivity, shear, trapezoidal, and Poisson's locking were investigated, and the attained results showed that almost all the plane elements formulated by the assumed strain approach are free from these shortcomings, and even can compute response practical problems using a coarse mesh of elements. Therefore, the strain-based elements completely fit in the definition of robust finite elements. It must be added that the recently proposed higher-order triangular plane elements such as RGR-T11D-I, RGR-T11D-II, and RGR-T14 are among the best-performing elements in all the analyzed benchmark problems. This shows the merit of using higher-order assumed strain fields and imposing equilibrium equations to the opted strain components. The mentioned advantages make assumed strain formulation an interesting alternative for developing robust finite elements of different types, such as plates, shells, and solids.

Declarations

It is confirmed that the Availability of data and material, Funding, Authors' contributions, Acknowledgments, and all the subheadings of these and also the relevant information under each have been declared in this paper. Moreover, there is no conflict of interest.

References

1. Al Akhrass, D., Bruchon, J., Drapier, S. and Fayolle, S. *Integrating a logarithmic-strain based hyperelastic formulation into a three-field mixed finite element formulation to deal with incompressibility in finite-strain elastoplasticity*, Finite Elem. Anal. Des. 86 (2014) 61–70.
2. Belarbi, M.T. and Bourezane, M., *On improved Sabir triangular element with drilling rotation*, Rev. eur. génie civ., 9(9-10) (2005), 1151–1175.
3. Belarbi, M.T. and Bourezane, M. *An assumed strain based on triangular element with drilling rotation*, Courier de Savoir, 6 (2005), 117–123.
4. Belarbi, M.T. and Maalem, T. *On improved rectangular finite element for plane linear elasticity analysis*, Revue Européenne des Éléments Finis, 14(8) (2005), 985–997.

5. Cen, S., Chen, X.M. and Fu, X.R. *Quadrilateral membrane element family formulated by the quadrilateral area coordinate method*, Comput. Methods Appl. Mech. Eng. 196(41-44) (2007) 4337–4353.
6. Cen, S., Zhou, P.L., Li, C.F. and Wu, C.J. *An unsymmetric 4-node, 8-DOF plane membrane element perfectly breaking through MacNeal's theorem*, Int. J. Numer. Methods Eng. 103(7) (2015) 469–500.
7. Felippa, C.A. *A study of optimal membrane triangles with drilling freedoms*, Comput. Methods Appl. Mech. Eng. **192**(16-18) (2003), 2125–2168.
8. Hamadi, D., Ayoub, A. and Maalem, T. *A new strain-based finite element for plane elasticity problems*, Eng. Comput. 33(2) (2016), 562–579.
9. MacNeal, R.H., Harder, R.L. *A refined four-node membrane element with rotational degrees of freedom*, Comput. Struct. 28(1) (1988) 75–84.
10. Madeo, A., Casciaro, R., Zagari, G., Zinno, R. and Zucco, G. *A mixed isostatic 16 dof quadrilateral membrane element with drilling rotations, based on Airy stresses*, Finite Elem. Anal. Des. 89 (2014) 52–66.
11. Paknahad, M., Noorzaei, J., Jaafar, M.S. and Thanoon, W.A. *Analysis of shear wall structure using optimal membrane triangle element*, Finite Elem. Anal. Des. 43(11-12) (2007) 861–869.
12. Pian, T.H. and Sumihara, K. *Rational approach for assumed stress finite elements*, Int. J. Numer. Methods Eng. 20(9) (1984) 1685–1695.
13. Rebiai, C. *Finite element analysis of 2-D structures by new strain based triangular element*, J. Mech. 35(3) (2018) 1–9.
14. Rebiai, C. and Belounar, L. *A new strain based rectangular finite element with drilling rotation for linear and nonlinear analysis*, Arch. Civ. Mech. Eng. 13(1) (2013) 72–81.
15. Rebiai, C. and Belounar, L. *An effective quadrilateral membrane finite element based on the strain approach*, Measurement, 50 (2014) 263–269.
16. Rebiai, C., Saidani, N. and Bahloul, E. *A new finite element based on the strain approach for linear and dynamic analysis*, Res. J. Appl. Sci. 11(6) (2015) 639–644.
17. Rezaiee-Pajand, M., Gharaei-Moghaddam, N. and Ramezani, M. *Two triangular membrane element based on strain*, Int. J. Appl. Mech. 11(1) (2019), 1950010.
18. Rezaiee-Pajand, M., Gharaei-Moghaddam, N. and Ramezani, M.R. *A new higher-order strain-based plane element*, Scientia Iranica. Transaction A, Civil Engineering, 26(4) (2019), 2258–2275.

19. Rezaiee-Pajand, M., Gharaei-Moghaddam, N. and Ramezani, M., *Higher-order assumed strain plane element immune to mesh distortion*, Eng. Comput. 37(9) (2020), 2957–2981.
20. Rezaiee-Pajand, M., Gharaei-Moghaddam, N. and Ramezani, M., *Strain-based plane element for fracture mechanics' problems*, Theor. Appl. Fract. Mech. 108 (2020), 102569.
21. Rezaiee-Pajand, and Ramezani, M. *An evaluation of MITC and ANS elements in the nonlinear analysis of shell structures*, Mech. Adv. Mater. Struct. (2021) 1–21.
22. Rezaiee-Pajand, M., Ramezani, M. and Gharaei-Moghaddam, N. *Using higher-order strain interpolation function to improve the accuracy of structural responses*, Int. J. Appl. Mech. 12(3) (2020), 2050026.
23. Rezaiee-Pajand, M. and Yaghoobi, M. *Formulating an effective generalized four-sided element*, Eur. J. Mech. A Solids, 36 (2012), 141–155.
24. Rezaiee-Pajand, M. and Yaghoobi, M. *A free of parasitic shear strain formulation for plane element*, Research in Civil and Environmental Engineering, 1 (2013) 1–27.
25. Rezaiee-Pajand, M. and Yaghoobi, M. *A robust triangular membrane element*, Lat. Am. J. Solids Struct. 11(14) (2014), 2648–2671.
26. Rezaiee-Pajand, M. and Yaghoobi, M. *An efficient formulation for linear and geometric non-linear membrane elements*, Lat. Am. J. Solids Struct. 11(6) (2014), 1012–1035.
27. Rezaiee-Pajand, M. and Yaghoobi, M. *Two new quadrilateral elements based on strain states*, Civ. Eng. Infrastruct. J. 48(1) (2015), 133–156.
28. Sabir, A.B. *A rectangular and triangular plane elasticity element with drilling degrees of freedom*, Proceedings of the Second International Conference on Variational Methods in Engineering, Brebbia CA ed., Southampton University (1985), 17–25.
29. Sabir, A.B. and Sfindji, A. *Triangular and rectangular plane elasticity finite elements*, Thin-Walled Struct. 21(3) (1995), 225–232.
30. Tayeh, S.M. *New strain-based triangular and rectangular finite elements for plane elasticity problems*, Thesis, The Islamic University, Gaza, 2003.
31. Taylor, R.L., Beresford, P.J. and Wilson, E.L. *A non-conforming element for stress analysis*, Int. J. Numer. Methods Eng. 10(6) (1976) 1211–1219.
32. Zhang, G. and Wang, M. *Development of eight-node curved-side quadrilateral membrane element using chain direct integration scheme (SCDI) in area coordinates (MHCQ8-DI)*, Arabian Journal for Science and Engineering, 44(5) (2019) 4703–4724.

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] Stoer, J. and Bulirsch, R. *Introduction to Numerical Analysis*, Springer-verlag, New York, 2002.
- [2] 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.

Iranian Journal of Numerical Analysis and Optimization

Former : MRJMS

CONTENTS

Vol. 11, No.2, pp 235-511, 2021

Trainable fourth-order partial differential equations for image noise removal	235
N. Khoeiniha, S.M. Hosseini and R. Davoudi	
Exponentially fitted tension spline method for singularly perturbed differential difference equations	261
M.M. Woldaregay and G.F. Duressa	
New class of hybrid explicit methods for numerical solution of optimal control problems	283
M. Ebadi, I. Malih Maleki and A. Ebadian	
The strict complementarity in linear fractional optimization	305
M. Mehdiloo, K. Tone and M.B. Ahmadi	
Solving quantum optimal control problems by wavelets method	333
M. Rahimi, S. M.Karbassi and M.R. Hooshmandasl	
Singularly perturbed robin type boundary value problems with discontinuous source term in geophysical fluid dynamics	351
B.M. Abagero, G.F. Duressa and H.G. Debela	
Two new approximations to Caputo–Fabrizio fractional equation on non-uniform meshes and its applications	365
Z. Soori and A. Aminataei	

The rest of contents is on back of the page

web site: <http://ijnao.um.ac.ir>

Email: ijnao@um.ac.ir

ISSN-Print: [2423-6977](#)

ISSN-Online: [2423-6969](#)

Application of Newton–Cotes quadrature rule for nonlinear Hammerstein integral equations	385
A. Shahsavaran	
Investigating a claim about resource complexity measure	401
H.R. Yousefzadeh	
A new algorithm for solving linear programming problems with bipolar fuzzy relation equation constraints	407
S. Aliannezhadi and A. Abbasi Molai	
Review of the strain-based formulation for analysis of plane structures, Part I: Formulation of basics and the existing elements	437
M. Rezaiee-Pajand, N. Gharaei-Moghaddam and M. Ramezani	
Review of the strain-based formulation for analysis of plane structures, Part II: Evaluation of the numerical performance	485
M. Rezaiee-Pajand, N. Gharaei-Moghaddam and M. Ramezani	