Volume 25, Number 3, 2025 ISSN 1562-8353

NONLINEAR DYNAMICS AND SYSTEMS THEORY

An International Journal of Research and Surveys

Volume 25 Number 3 2025

CONTENTS

 ISSN 1562-8353 Printed in Ukraine
To receive contents and abstracts by e-mail, visit our Website at: **http://www.e-ndst.kiev.ua**

NONLINEAR DYNAMICS & SYSTEMS THEORY

Volume 25, No. 3, 2025

Nonlinear Dynamics and Systems Theory

An International Journal of Research and Surveys

InforMath Publishing Group
http://www.e-ndst.kiev.ua

Nonlinear Dynamics and Systems Theory
An International Journal of Research and Surveys

INSTRUCTIONS FOR CONTRIBUTORS

(1) General. Nonlinear Dynamics and Systems Theory (ND&ST) is an international journal devoted to publishing peer-refereed, high quality, original papers, brief notes and review articles focusing on nonlinear dynamics and systems theory and their practical applications in engineering, physical and life sciences. Submission of a manuscript is a representation that the submission has been approved by all of the authors and by the institution where the work was carried out. It also represents that the manuscript has not been previously published, has not been copyrighted, is not being submitted for publication elsewhere, and that the authors have agreed that the copyright in the article shall be assigned exclusively to InforMath Publishing Group by signing a transfer of copyright form. Before submission, the authors should visit the website:

http://www.e-ndst.kiev.ua

for information on the preparation of accepted manuscripts. Please download the archive Sample_NDST.zip containing example of article file (you can edit only the file Samplefilename.tex).

(2) Manuscript and Correspondence. Manuscripts should be in English and must meet common standards of usage and grammar. To submit a paper, send by e-mail a file in PDF format directly to

Professor A.A. Martynyuk, Institute of Mechanics,
Nesterov str.3, 03057, Kiev-57, Ukraine
e-mail: journalndst@gmail.com

or to one of the Regional Editors or to a member of the Editorial Board. Final version of the manuscript must typeset using LaTex program which is prepared in accordance with the style file of the Journal. Manuscript texts should contain the title of the article, name(s) of the author(s) and complete affiliations. Each article requires an abstract not exceeding 150 words. Formulas and citations should not be included in the abstract. AMS subject classifications and key words must be included in all accepted papers. Each article requires a running head (abbreviated form of the title) of no more than 30 characters. The sizes for regular papers, survey articles, brief notes, letters to editors and book reviews are: (i) 10-14 pages for regular papers, (ii) up to 24 pages for survey articles, and (iii) 2-3 pages for brief notes, letters to the editor and book reviews.

(3) Tables, Graphs and Illustrations. Each figure must be of a quality suitable for direct reproduction and must include a caption. Drawings should include all relevant details and should be drawn professionally in black ink on plain white drawing paper. In addition to a hard copy of the artwork, it is necessary to attach the electronic file of the artwork (preferably in PCX format).

(4) References. Each entry must be cited in the text by author(s) and number or by number alone. All references should be listed in their alphabetic order. Use please the following style:

Journal: [1] H. Poincare, Title of the article. *Title of the Journal* **volume** (issue) (year) pages. [Language]

Book: [2] A.M. Lyapunov, *Title of the Book.* Name of the Publishers, Town, year.

Proceeding: [3] R. Bellman, Title of the article. In: *Title of the Book.* (Eds.). Name of the Publishers, Town, year, pages. [Language]

(5) Proofs and Sample Copy. Proofs sent to authors should be returned to the Editorial Office with corrections within three days after receipt. The corresponding author will receive a sample copy of the issue of the Journal for which his/her paper is published.

(6) Editorial Policy. Every submission will undergo a stringent peer review process. An editor will be assigned to handle the review process of the paper. He/she will secure at least two reviewers' reports. The decision on acceptance, rejection or acceptance subject to revision will be made based on these reviewers' reports and the editor's own reading of the paper.

NONLINEAR DYNAMICS AND SYSTEMS THEORY
An International Journal of Research and Surveys
Published by InforMath Publishing Group since 2001

Volume 25 Number 3 2025

CONTENTS

Founded by A.A. Martynyuk in 2001.
Registered in Ukraine Number: KB No 5267 / 04.07.2001.

NONLINEAR DYNAMICS AND SYSTEMS THEORY
An International Journal of Research and Surveys

Impact Factor from SCOPUS for 2023: SJR – 0.298, SNIP – 1.032 and CiteScore – 1.6

Nonlinear Dynamics and Systems Theory (ISSN 1562–8353 (Print), ISSN 1813–7385 (Online)) is an international journal published under the auspices of the S.P. Timoshenko Institute of Mechanics of National Academy of Sciences of Ukraine and Curtin University of Technology (Perth, Australia). It aims to publish high quality original scientific papers and surveys in areas of nonlinear dynamics and systems theory and their real world applications.

AIMS AND SCOPE

Nonlinear Dynamics and Systems Theory is a multidisciplinary journal. It publishes papers focusing on proofs of important theorems as well as papers presenting new ideas and new theory, conjectures, numerical algorithms and physical experiments in areas related to nonlinear dynamics and systems theory. Papers that deal with theoretical aspects of nonlinear dynamics and/or systems theory should contain significant mathematical results with an indication of their possible applications. Papers that emphasize applications should contain new mathematical models of real world phenomena and/or description of engineering problems. They should include rigorous analysis of data used and results obtained. Papers that integrate and interrelate ideas and methods of nonlinear dynamics and systems theory will be particularly welcomed. This journal and the individual contributions published therein are protected under the copyright by International InforMath Publishing Group.

PUBLICATION AND SUBSCRIPTION INFORMATION

Nonlinear Dynamics and Systems Theory will have 6 issues in 2025, printed in hard copy (ISSN 1562–8353) and available online (ISSN 1813–7385), by InforMath Publishing Group, Nesterov str., 3, Institute of Mechanics, Kiev, MSP 680, Ukraine, 03057. Subscription prices are available upon request from the Publisher, EBSCO Information Services (`mailto:journals@ebsco.com`), or website of the Journal: `http://e-ndst.kiev.ua`. Subscriptions are accepted on a calendar year basis. Issues are sent by airmail to all countries of the world. Claims for missing issues should be made within six months of the date of dispatch.

ABSTRACTING AND INDEXING SERVICES

Papers published in this journal are indexed or abstracted in: Mathematical Reviews / MathSciNet, Zentralblatt MATH / Mathematics Abstracts, PASCAL database (INIST–CNRS) and SCOPUS.

Nonlinear Dynamics and Systems Theory, **25** (3) (2025) 231–242

A Primal-Dual IPM Algorithm for LO Problem Based on a New Kernel Function with a Logarithmic Barrier Term

Abderrahim Guemmaz [1*], Bachir Bounibane [2] and El Amir Djeffal [2]

[1] *Department of Mathematics, Institute of Science, University Center of Barika, Algeria.*
[2] *Department of Mathematics, University of Batna 2, Batna, Algeria.*

Received: April 7, 2024;   Revised: May 20, 2025

Abstract: In this paper, we consider a primal-dual Interior Point Method (IPM) for the linear optimization(LO) problem, based on a new kernel function with a logarithmic barrier term, which plays an important role for developing a new design of primal-dual IPM algorithms. New search directions and proximity functions are proposed based on this kernel function. We proved that our algorithm has $\mathbf{O}\left(qSn^{\frac{Sq+1}{2Sq}}\log\left(\frac{n}{\epsilon}\right)\right)$ iteration bound for large-update methods.

Keywords: *primal-dual interior point algorithm; kernel function; linear optimization problem; iteration bound; complexity.*

Mathematics Subject Classification (2020): 90C51, 49N15, 90C05, 68Q25.

1 Introduction

In this paper we deal with primal-dual IPMs for solving the standard linear optimization (LO) problem

$$(P) \qquad \min\left\{c^T x : Ax = b, x \geq 0\right\},$$

and the dual problem of (P) is given by

$$(D) \qquad \max\left\{b^T y + s = c, s \geq 0\right\},$$

where $A \in \mathbb{R}^{m\times n}, x, s, c \in \mathbb{R}^n$, and $y, b \in \mathbb{R}^m$.

* Corresponding author: mailto:abderrahim.guemmaz@cu-barika.dz

In 1984, Karmarkar [8] proposed a new polynomial-time method for solving linear programs. This method and its variants that were developed subsequently are now called IPMs, and they have become the most effective methods for solving LO problems. The new efficient algorithms of the interior-point methods (IPMs) have generated increased interest both in the application and the research of LO. In this paper, we deal with the so-called primal-dual IPMs. It is generally agreed that these IPMs are most efficient from a computational point of view [7]. Many researchers have designed different types of primal-dual interior-point methods. Among them, IPMs based on kernel functions have been designed. Several kernel functions have been introduced, including the so-called self-regular kernel functions [2,4] and the non-self-regular kernel functions [2,11]. In principle, a kernel function offers a search direction and hence the development of a primal-dual interior point method. Until now, all primal-dual IPMs have used the Newton direction as the search direction [6]; this direction is closely related to the well-known primal-dual logarithmic barrier function. In this paper, we consider the new kernel function with a logarithmic Barrier Term (1.1) from [11] as follows:

$$\psi_S(t) = \frac{\left(t^2-1\right)}{2} - \frac{\log(t)}{2} - \frac{1}{2S}\sum_{j=1}^{S}\frac{t^{1-jq}-1}{1-jq}, q>1, S\in\mathbb{N}\backslash\{0\}. \tag{1}$$

We will formulate an interior-point algorithm for LO by using a new proximity function and give its complexity analysis, and then we will show that the iteration bounds are $\mathbf{O}\left(qSn^{\frac{Sq+1}{2Sq}}\log\left(\frac{n}{\epsilon}\right)\right)$ and $\mathbf{O}\left(q^2S^2\sqrt{n}\log\left(\frac{n}{\epsilon}\right)\right)$ for large and small-update methods, respectively.

The remainder of this paper is organized as follows. First, in Section (2), we define the central path and the new search direction determined by Kernel Functions for LO, then we present the generic primal-dual IPM algorithm. The new kernel function and its properties are presented in Section (3). In Section (4), we analyse the algorithm and derive the complexity bound for LO. Finally, some concluding remarks follow in Section (5).

Some notations used throughout the paper are as follows. Let $\mathbb{R}^n$ be the n-dimensional Euclidean space with the inner product $\langle .,. \rangle$, and $\|.\|$ denote the 2-norm. $\mathbb{R}^n_+$ and $\mathbb{R}^n_{++}$ denote the set of n-dimensional nonnegative vectors and positive vectors, respectively. For $x, s \in \mathbb{R}^n$, x_{min} and xs denote the smallest component of the vector x, and the componentwise product of the vector x and s, respectively. We denote by $X = diag(x)$ the $n \times n$ diagonal matrix with the components of vector $x \in \mathbb{R}^n$ being the diagonal entries, e denotes the n-dimensional vector, where each coordinate takes the value 1. For two real-valued functions $f(x), g(x) : \mathbb{R}_{++} \longrightarrow \mathbb{R}_{++}$, $f(x) = O(g(x))$ if $f(x) \le cg(x)$ for some positive c, and $f(x) = \Theta(g(x))$ if $c_1g(x) \le f(x) \le c_2g(x)$ for some positive constants c_1 and c_2.

2 Preliminaries

It is well known that the optimality condition for (P) and (D) is equivalent to solving the following nonlinear system:

$$\begin{cases} Ax = b, & x \ge 0, \\ A^Ty + s = c, & s \ge 0, \\ xs = 0. \end{cases} \tag{2}$$

The basic idea of primal-dual IPMs for LO problems is to replace the third equation in (2), which is known as a complementarity condition for (P) and (D), by the parameterized equation $xs = \mu e$, with $\mu > 0$. Thus, the system (2) becomes

$$\begin{cases} Ax = b, & x \geq 0, \\ A^T y + s = c, & s \geq 0, \\ xs = \mu e. \end{cases} \tag{3}$$

Due to the last equation, any solution (x, y, s) of (3) will satisfy $x > 0$ and $s > 0$. So, a solution exists only if (P) and (D) satisfy the interior-point condition (IPC) [5], i.e., there exists (x^0, y^0, s^0) such that

$$\begin{cases} Ax^0 = b, & x^0 > 0, \\ A^T y^0 + s^0 = c, & s^0 > 0. \end{cases}$$

So, if the IPC is satisfied, the system (3) has only one solution $(x(\mu), y(\mu), s(\mu))$ for every $\mu > 0$ (see Lemma 4.3 in [13]), $x(\mu)$ is called the μ-center of (P), and $(y(\mu), s(\mu))$ is the μ-center of (D). The set of μ-centers is called the central path of (P) and (D). If $\mu \longrightarrow 0$, then the limit of the central path exists, and since the limit points satisfy the complementarity condition, the limit yields optimal solutions for (P) and (D) [5].

Let $\mu > 0$ be fixed. A direct application of the Newton method to (3) provides the following system for $\Delta x, \Delta y$ and Δs:

$$\begin{cases} A\Delta x = 0, \\ A^T \Delta y + \Delta s = 0, \\ x\Delta s + s\Delta x = \mu e - sx. \end{cases} \tag{4}$$

Since A has full row rank, the system (4) has a unique solution $(\Delta x, \Delta y, \Delta s)$ which is called the search direction (see [5, 9]). By taking a step along the search direction $(\Delta x, \Delta y, \Delta s)$, one constructs a new positive iterate (x_+, y_+, s_+) with

$$x_+ := x + \alpha\Delta x, y_+ := y + \alpha\Delta y,\ s_+ := s + \alpha\Delta s,$$

where α satisfies $0 < \alpha \leq 1$.

Now, we introduce the scaled vector v and the scaled search directions dx and ds as follows:

$$v := \sqrt{\frac{xs}{\mu}},\ d_x := \frac{v\Delta x}{x},\ d_s := \frac{v\Delta s}{s}. \tag{5}$$

The system (4) can be rewritten as follows:

$$\begin{cases} \overline{A} d_x = 0, \\ \overline{A}^T \Delta y + d_s = 0, \\ d_x + d_s = v^{-1} - v, \end{cases} \tag{6}$$

where $\overline{A} := \frac{1}{\mu} AV^{-1}X, V := diag\,(v)$ and $X := diag\,(x)$.

Note that

$$d_x = d_s = 0 \Leftrightarrow v^{-1} - v = 0 \Leftrightarrow v = e \Longleftrightarrow x = x(\mu), s = s(\mu).$$

A useful observation is that the right-hand side of the third equation in (6) equals to the minus gradient of the following proximity function:

$$\Phi(v) = \Phi(x, s; \mu) = \sum_{i=1}^{n} \psi(v_i) = \sum_{i=1}^{n} \left(\frac{v_i^2 - 1}{2} - \log v_i \right), v_i > 0.$$

Here, ψ is the so-called kernel function of Φ. And therefore, $d_x + d_s = -\nabla\Phi(v)$. We can rewrite the system (6) as

$$\begin{cases} \overline{A} d_x = 0, \\ \overline{A}^T \Delta y + d_s = 0, \\ d_x + d_s = -\nabla\Phi(v). \end{cases} \tag{7}$$

It is easy to notice that $\nabla\Phi(v) = 0$, therefore $\Phi(v)$ reaches its minimum value at $v = e$, with $\Phi(v) = 0$.

In order to measure the distance between the μ-center and the current iterate, we resort to using $\Phi(v)$, and this is for a given $\tau > 0$.

Now, we introduce a norm-based proximity measure $\delta(v) : \mathbb{R}^n_{++} \to \mathbb{R}_+$ in accordance with

$$\delta(v) = \frac{1}{2} ||\nabla\Phi(v)|| = \frac{1}{2} ||d_x + d_s||, \tag{8}$$

in terms of $\psi(v_i)$. Then we have $\psi(v_i) = 0 \Leftrightarrow \delta(v) = 0 \Leftrightarrow v = e$.

Using (5) and (8), we can write the system (4) in the form of a modified Newton system. We get the following:

$$\begin{cases} A\Delta x = 0, \\ A^T \Delta y + \Delta s = 0, \\ x\Delta s + s\Delta x = -\mu v \nabla\Phi(v). \end{cases} \tag{9}$$

In this paper, we replace $\psi(t)$ by a new kernel function $\psi_S(t)$, and $\Phi(v)$ by a new barrier function $\Phi_S(v)$, which will be defined in Section (3).

The new interior-point algorithm works as follows. Assume that we are given a strictly feasible point $(x; y; s)$ which is in a τ-neighbourhood of the given μ-centre. Then we decrease μ to $\mu_+ = (1 - \theta)\mu$ for some fixed $\theta \in (0, 1)$, and then solve the Newton system (4) to obtain the unique search direction. The positivity of a new iterate is ensured by an appropriate choice of the step size α which is defined by some line search rule. This procedure is repeated until we find a new iterate (x_+, y_+, s_+) that is in a τ-neighbourhood of the μ_+-centre, and then we let $\mu := \mu_+$ and $(x; y; s) := (x_+, y_+, s_+)$. Then μ is again reduced by the factor $(1 - \theta)$ and we solve the Newton system targeting at the new μ_+-centre, and so on. This process is repeated until μ is small enough, say until $n\mu < \epsilon$. The generic form of the algorithm is shown in Fig.1.

Generic Primal-Dual Algorithm for LO

Input:
a proximity function $\Phi_s(v)$; a threshold parameter $\tau > 0$;
an accuracy parameter $\epsilon > 0$; a fixed barrier update parameter $\theta, 0 < \theta < 1$;
(x^0, s^0) and $\mu^0 := 1$ such that $\Phi_s(x^0, s^0, \mu^0) \leq \tau$.
begin
$x := x^0; s := s^0; \mu := \mu^0;$
while $n\mu \geq \epsilon$ **do**
begin (outer iteration)
$\mu := (1-\theta)\mu; v := \sqrt{\frac{xs}{\mu}};$
while $\Phi_s(v) > \tau$ **do**
begin (inner iteration)
Solve the system (9) for $(\Delta x, \Delta y, \Delta s)$;
Determine a step size α;
$x := x + \alpha\Delta x; y := y + \alpha\Delta y; s := s + \alpha\Delta s; v := \sqrt{\frac{xs}{\mu}};$
end (inner iteration)
end (outer iteration)
end

Fig. 1 Generic algorithm.

We want to "optimize" the algorithm by minimizing the number of iterates in the algorithm. To do this, we must carefully choose the parameters τ, θ, and the step size α. Choosing the barrier update parameter θ is very important in application and theory. If θ is a constant number which is independent of the dimension n of the problem, i.e., $\theta = \mathbf{\Theta}(1)$, then the algorithm is called a large update method. If θ depends on the dimension n of the problem, then we call the algorithm a small update method. In this case, θ is usually chosen as follows: $\theta = \mathbf{\Theta}\left(\frac{1}{\sqrt{n}}\right)$.

Choosing the step size, $\alpha > 0$, is another key step in obtaining good convergence properties of the algorithm. It must be set in such a way that the closeness of the iterates to the current μ-center improves by a sufficient amount.

In this paper, we define a new kernel function and propose primal-dual interior point methods which improve all the results of the complexity bound for large-update methods based on a logarithmic kernel function for LO. More precisely, based on the proposed kernel function, we prove that the corresponding algorithm has $\mathbf{O}\left(qSn^{\frac{Sq+1}{2Sq}} \log\left(\frac{n}{\epsilon}\right)\right)$complexity bound for the large-update method, and $\mathbf{O}\left(q^2S^2\sqrt{n}\log\left(\frac{n}{\epsilon}\right)\right)$ for the small-update method. Another interesting choice is q dependences with n and S, which minimizes the iteration complexity bound. In fact, if we take $q = \frac{\log n}{2S}$, we obtain the best known complexity bound for large-update methods, namely, $\mathbf{O}\left(\sqrt{n}\log n \log \frac{n}{\varepsilon}\right)$. This bound improves the so far obtained complexity results for large-update methods based on a logarithmic kernel function given by El Ghami et al. [10].

3 The Properties of the New Kernel Function

We will now address a new kernel function with its properties being provided. Let the new univariate function be defined in [11].

$$\psi_S(t)=\frac{\left(t^2-1\right)}{2}-\frac{\log(t)}{2}-\frac{1}{2S}\sum_{j=1}^{S}\frac{t^{1-jq}-1}{1-jq},\ q>1,\ S\in\mathbb{N}\backslash\{0\}.$$

It is easy to observe that as $t\to 0$ or $t\to\infty$, then $\psi(t)\to\infty$. So, $\psi_S(t)$ is without a doubt a kernel function.

We will need the first three derivatives of $\psi_S(t)$, we provide them as follows:

$$\psi_S^{'}(t)=t-\frac{1}{2t}-\frac{1}{2S}\sum_{j=1}^{S}t^{-jq}, \tag{10}$$

$$\psi_S^{''}(t)=1+\frac{1}{2t^2}+\frac{1}{2S}\sum_{j=1}^{S}qjt^{-jq-1}, \tag{11}$$

$$\psi_S^{'''}(t)=-\frac{1}{t^3}-\frac{1}{2S}\sum_{j=1}^{S}jq(jq+1)t^{-jq-2}. \tag{12}$$

If $S=1$, we obtain the kernel function (12) given by Bouaafia et al. [10].

The following lemma establishes the efficiency of the new kernel function (1).

Lemma 3.1 *Let $\psi_S(t)$ be as defined in* (1) *and $t>0$. Then*

$$\psi_S^{''}(t)>1, \tag{13}$$
$$\psi_S^{'''}(t)<0, \tag{14}$$
$$t\psi_S^{''}(t)-\psi_S^{'}(t)>0, \tag{15}$$
$$t\psi_S^{''}(t)+\psi_S^{'}(t)>0. \tag{16}$$

The last property (16) in Lemma 3.1 is equivalent to the convexity of composed functions $t\to\psi_S(e^t)$ and this holds if and only if $\psi_S(\sqrt{t_1t_2})\leq\frac{1}{2}\left(\psi_S(t_1)+\psi_S(t_2)\right)$ for any $t_1,t_2\geq 0$. This property is well-known in the literature, and numerous researchers have demonstrated it (see [3, 12]).

We provide some technical findings of the new kernel function in preparation for further.

Lemma 3.2 *For $\psi_S(t)$, we get*

$$\frac{1}{2}(t-1)^2\leq\psi_S(t)\leq\frac{1}{2}\left[\psi_S^{'}(t)\right]^2,\ t>0. \tag{17}$$

$$\psi_S(t)\leq\left[\frac{6+q(S+1)}{8}\right](t-1)^2,\ t>1. \tag{18}$$

Proof. For (17), use (13). For (18), use Taylor's Theorem.

Let $\sigma:[0,\infty[\,\to[1,+\infty[$ be the inverse function of $\psi_S(t)$ for $t\geq 1$ and $\rho:[0,\infty[\,\to\,]0,1]$ be the inverse function of $-\frac{1}{2}\psi_S^{'}(t)$ for all $t\in\,]0,1]$. Then we have the following lemma.

Lemma 3.3 *[Lemma 3.3 from [11]] For $\psi_S(t)$, we have*

$$1+\sqrt{\frac{8s}{6+q(S+1)}} \leq \sigma(s) \leq 1+\sqrt{2s},\ s \geq 0. \tag{19}$$

$$\rho(z) > \left[\frac{1}{4z+2}\right]^{\frac{1}{Sq}},\ z > 0. \tag{20}$$

Lemma 3.4 *Let $\sigma : [0,\infty[\to [1,+\infty[$ be the inverse of $\psi_S(t)$. We have*

$$\Phi_S(\beta v) \leq n\psi_S\left(\beta\sigma\left(\frac{\Phi_S(v)}{n}\right)\right),\ v \in \mathbb{R}^*,\ \beta \geq 1.$$

Proof. Using (14) and (15), and Lemma 2.4 from [1], we can obtain the result. This completes the proof.

Lemma 3.5 *[Lemma 3.5 from [11]] Let $0 \leq \theta < 1$, $v_+ = \frac{v}{\sqrt{1-\theta}}$. If $\Phi_S(v) \leq \tau$, then we have*

$$\Phi_S(v_+) \leq \frac{\theta n + 2\tau + 2\sqrt{2\tau n}}{2(1-\theta)}.$$

Denote

$$(\Phi_S)_0 = \frac{\theta n + 2\tau + 2\sqrt{2\tau n}}{2(1-\theta)} = L(n,\theta,\tau).$$

So, during the algorithm's execution, $(\Phi_S)_0$ is the upper bound of $\Phi_S(v_+)$.

4 Complexity Analysis

In the next subsection, we compute a default step size α and the resulting decrease in the barrier function.

4.1 An estimation of the step size

We devoted this section to calculating a default step size α and the consequent decrease in the barrier function. And after the damping step, we obtain

$$x_+ = x + \alpha\Delta x,\ y_+ = y + \alpha\Delta y,\ s_+ = s + \alpha\Delta s.$$

By using (5), we get that

$$x_+ = x\left(e + \alpha\frac{\Delta x}{x}\right) = x\left(e + \alpha\frac{d_x}{v}\right) = \frac{x}{v}\left(v + \alpha d_x\right),$$
$$s_+ = s\left(e + \alpha\frac{\Delta s}{s}\right) = s\left(e + \alpha\frac{d_s}{v}\right) = \frac{s}{v}\left(v + \alpha d_s\right).$$

Hence, $v_+ = \sqrt{\frac{x_+ s_+}{\mu}} = \sqrt{(v+\alpha d_x)(v+\alpha d_s)}$. Define for $\alpha > 0$,

$$f(\alpha) = \Phi_S(v_+) - \Phi_S(v). \tag{21}$$

Therefore, $f(\alpha)$ represents the difference in proximities between a new iterate and a current iterate for a given value of μ . By (5), we can get

$$\Phi_S(v_+) = \Phi_S\left(\sqrt{(v+\alpha d_x)(v+\alpha d_s)}\right) \leq \frac{1}{2}\left(\Phi_S((v+\alpha d_x)) + \Phi_S((v+\alpha d_s))\right).$$

Thus, we have $f(\alpha) \leq f_1(\alpha)$ such that

$$f_1(\alpha) = \frac{1}{2}\left(\Phi_S((v+\alpha d_x)) + \Phi_S((v+\alpha d_s))\right) - \Phi_S(v). \tag{22}$$

Clearly, $f(0) = f_1(0) = 0$. We calculate $f_1^{'}(\alpha)$ and $f_1^{''}(\alpha)$, we find

$$f_1^{'}(\alpha) = \frac{1}{2}\sum_{i=1}^{n}\left(\psi_S^{'}(v_i+\alpha d_{x_i})d_{x_i} + \psi_S^{'}(v_i+\alpha d_{s_i})d_{s_i}\right),$$
$$f_1^{''}(\alpha) = \frac{1}{2}\sum_{i=1}^{n}\left(\psi_S^{''}(v_i+\alpha d_{x_i})d_{x_i}^2 + \psi_S^{''}(v_i+\alpha d_{s_i})d_{s_i}^2\right).$$

By using (5) and (8), we conclude that

$$f_1^{'}(0) = \frac{1}{2} < \nabla\Phi_S(v), (d_x + d_s) >= -\frac{1}{2} < \nabla\Phi_S(v), \nabla\Phi_S(v) >= -2\delta(v)^2.$$

We denote $v_1 = min(v)$, $\delta = \delta(v)$, $\Phi_S = \Phi_S(v)$.

Lemma 4.1 *Let $\delta(v)$ be defined in* (8)*. Then*

$$\delta(v) \geq \sqrt{\frac{\Phi_S(v)}{2}}. \tag{23}$$

Proof. Using (17), we have

$$\Phi_S(v) = \sum_{i=1}^{n}\psi_S(v_i) \leq \sum_{i=1}^{n}\frac{1}{2}\left[\psi_S^{'}(v_i)\right]^2 = \frac{1}{2}||\nabla\Phi_S(v)||^2 = 2\delta(v)^2.$$

Hence, $\delta(v) \geq \sqrt{\frac{1}{2}\Phi_S(v)}$. This completes the proof.

Remark 4.1 Throughout the paper, we assume that $\Phi_S(v) \geq \tau \geq 1$, and we have $\delta(v) \geq \frac{1}{2}$.

According to Lemmas 4.1-4.4 in [1], we get the following Lemmas 4.2 and 4.5 since $\psi_S(t)$ is a kernel function, and $\psi_S^{''}(t)$ decreases monotonically.

Lemma 4.2 *[Bai et al. [1]] Let $f_1(\alpha)$ be as defined in* (21) *and $\delta(v)$ be as defined in* (8)*. Then we have $f_1^{''}(\alpha) \leq 2\delta^2\psi_S^{''}(v_{min} - 2\alpha\delta)$. Because of the convexity of $f_1(\alpha)$, we will have $f_1^{'}(\alpha) \leq 0$ for any α less than or equal to the minimum value of $f_1(\alpha)$, and vice versa.*

The following three Lemmas result from the preceding Lemma.

Lemma 4.3 *[Bai et al. [1]] $f_1^{'}(\alpha) \leq 0$ certainly holds if α satisfies the inequality*

$$\psi_S^{'}(v_{min}) - \psi_S^{'}(v_{min} - 2\alpha\delta) \leq 2\delta. \tag{24}$$

Lemma 4.4 *[Bai et al. [1]] The largest step size $\overline{\alpha}$ satisfying* (24) *is given by*

$$\overline{\alpha} = \frac{\rho(\delta) - \rho(2\delta)}{2\delta}.$$

Lemma 4.5 *[Bai et al. [1]] Let $\overline{\alpha}$ be as defined in Lemma* 4.4*. Then*

$$\overline{\alpha} \geq \frac{1}{\psi_S^{''}(\rho(2\delta))}.$$

We are able to demonstrate the following Lemma.

Lemma 4.6 *[Lemma 4.6 from [11]] Let ρ and $\overline{\alpha}$ be as determined in Lemma* 4.5*. If $\Phi_S(v) \geq \tau \geq 1$, then we have*

$$\overline{\alpha} \geq \frac{2S}{2S + S(4\delta+2)^{\frac{2}{Sq}} + q\sum_{j=1}^{S} j(4\delta+2)^{\frac{jq+1}{Sq}}}.$$

Denoting

$$\widetilde{\alpha} = \frac{2S}{2S + S(4\delta+2)^{\frac{2}{Sq}} + q\sum_{j=1}^{S} j(4\delta+2)^{\frac{jq+1}{Sq}}}, \tag{25}$$

we have $\widetilde{\alpha}$ is the default step size, and $\widetilde{\alpha} \leq \overline{\alpha}$.

Lemma 4.7 *[Lemma 3.12 from [3]] Let h be a convex and twice differentiable function with $h(0) = 0$, $h^{'}(0) < 0$, which reaches its minimum at $t^* > 0$. If $h^{''}$ is increasing for $t \in [0, t^*]$, then*

$$h(t) \leq \frac{th^{'}(0)}{2},\ 0 \leq t \leq t^*.$$

The following result is of great importance.

Lemma 4.8 *[Lemma 4.5 from [1]] If the step size α satisfies $\alpha \leq \overline{\alpha}$, then*

$$f(\alpha) \leq -\alpha\delta^2.$$

Lemma 4.9 *Let $\Phi_S(v) \geq 1$ and let $\widetilde{\alpha}$ be the default step size as defined in* (25)*. Then we have*

$$f(\widetilde{\alpha}) \leq -\frac{2S}{8\sqrt{2}(S+8)(1+4qS)}\left[\Phi_S(v)\right]^{\frac{Sq-1}{2Sq}}. \tag{26}$$

Proof. Since $\Phi_S(v) \geq 1$, from (23), we have

$$\delta \geq \sqrt{\frac{1}{2}\Phi_S(v)} \geq \sqrt{\frac{1}{2}}.$$

Due to Lemma 4.8, with $\alpha = \widetilde{\alpha}$ and (25), this completes the proof.

4.2 Iteration bound

Following the updating of μ to $(1-\theta)\mu$, we obtain

$$\Phi_S(v_+) \leq (\Phi_S)_0 = \frac{\theta n + 2\tau + 2\sqrt{2\tau n}}{2(1-\theta)} = L(n, \theta, \tau).$$

After μ-update to $(1-\theta)\mu$, it is necessary to count how many inner iterations are required to come back to the situation where $\Phi_S(v_+) \leq \tau$. We declare the value of $\Phi_S(v)$ after the updating of μ as $(\Phi_S)_0$ and we denote by $(\Phi_S)_k$, $k = 1, 2, ..., K$, the subsequent values in the same outer iteration such that K represents the total number of inner iterations per the outer iteration.

Lemma 4.10 *[Lemma 14 from [3]] Let $t_0, t_1, ..., t_k$ be a sequence of positive numbers such that*

$$t_{k+1} \leq t_k - \beta t_k^{1-\gamma}, k = 0, 1, ..., K-1,$$

where $\beta > 0$ and $0 < \gamma \leq 1$, then $K \leq \left[\frac{t_0^\gamma}{\beta\gamma}\right]$.

Thus, it follows that

$$(\Phi_S)_{k+1} \leq (\Phi_S)_k - k(\Phi_S)^{1-\gamma}, k = 0, 1, ..., K-1,$$

with

$$\kappa = \frac{2S}{8\sqrt{2}(S+8)(1+4qS)}, \gamma = 1 - \frac{Sq-1}{2Sq} = \frac{Sq+1}{2Sq}.$$

Lemma 4.11 *Let K be the total number of inner iterations in the outer iteration. Then we have*

$$K \leq \frac{8\sqrt{2}q(S+8)(1+4qS)}{1+Sq}\left[(\Phi_S)_0\right]^{\frac{Sq+1}{2Sq}}.$$

Proof. By Lemma 1.3.2 from [3], we have
$K \leq \frac{\left[(\Phi_S)_0\right]^\gamma}{\kappa\gamma} = \frac{8\sqrt{2}q(S+8)(1+4qS)}{Sq+1}\left[(\Phi_S)_0\right]^{\frac{Sq+1}{2Sq}}$. This completes the proof.
Now, we estimate the total number of iterations of our algorithm.

We recall that the number of outer iterations is limited from above by $\frac{\log\left(\frac{n}{\epsilon}\right)}{\theta}$ (see Lemma II.17, page 116 in [5]). We can establish an upper bound on the total number of iterations by multiplying the number of outer iterations by the number of inner iterations such as

$$\frac{8\sqrt{2}q(S+8)(1+4qS)}{Sq+1}\left[(\Phi_S)_0\right]^{\frac{Sq+1}{2Sq}}\frac{\log\left(\frac{n}{\epsilon}\right)}{\theta}. \tag{27}$$

In the methods of large-update with $\tau = \mathbf{O}(n)$ and $\theta = \Theta(1)$, we have

$$\mathbf{O}\left(qSn^{\frac{Sq+1}{2Sq}}\log\left(\frac{n}{\epsilon}\right)\right) \text{ iterations complexity.}$$

This is the best well-known complexity result for large-update methods.

In the methods of small-update, the replacement of $\tau = \mathbf{O}(1)$ and $\theta = \Theta\left(\frac{1}{\sqrt{n}}\right)$ in (27) does not provide the best possible bound. The best bound is obtained as follows. By (18), with $\psi_S(t) \le \left[\frac{6+q(S+1)}{8}\right](t-1)^2$, $t > 1$, we have

$$\begin{aligned}\Phi_S(V_+) &\le n\psi_S\left(\frac{1}{\sqrt{1-\theta}}\sigma\left(\frac{\Phi_S(V)}{n}\right)\right)\\ &\le n\left[\frac{6+q(S+1)}{8}\right]\left(\frac{1}{\sqrt{1-\theta}}\sigma\left(\frac{\Phi_S(V)}{n}\right)-1\right)^2\\ &= \frac{n\,(6+q(S+1))}{8(1-\theta)}\left(\sigma\left(\frac{\Phi_S(V)}{n}\right)-\sqrt{1-\theta}\right)^2.\end{aligned}$$

Using (19), we have

$$\begin{aligned}\frac{n\,(6+q(S+1))}{8(1-\theta)}\left(\sigma\left(\frac{\Phi_S(V)}{n}\right)-\sqrt{1-\theta}\right)^2 &\le \frac{n\,(6+q(S+1))}{8(1-\theta)}\left(\left(1+\sqrt{2\frac{\Phi_S(V)}{n}}\right)-\sqrt{1-\theta}\right)^2\\ &= \frac{n\,(6+q(S+1))}{8(1-\theta)}\left(\left(1-\sqrt{1-\theta}\right)+\sqrt{2\frac{\Phi_S(V)}{n}}\right)^2\\ &\le \frac{n\,(6+q(S+1))}{8(1-\theta)}\left(\theta+\sqrt{2\frac{\tau}{n}}\right)^2\\ &= \frac{(6+q(S+1))}{8(1-\theta)}\left(\theta\sqrt{n}+\sqrt{2\tau}\right)^2 = (\Phi_S)_0\,,\end{aligned}$$

where we also use $1-\sqrt{1-\theta} = \frac{\theta}{1+\theta} \le \theta$ and $\Phi_S(v) \le \tau$, utilizing this upper bound for $(\Phi_S)_0$, we obtain the following iteration bound:

$$\frac{8\sqrt{2}q(S+8)\,(1+4qS)}{Sq+1}\left[(\Phi_S)_0\right]^{\frac{Sq+1}{2Sq}}\frac{\log\left(\frac{n}{\epsilon}\right)}{\theta}.$$

Note now that $(\Phi_S)_0 = \mathbf{O}(qS)$, and the iteration bound is given as follows:

$$\mathbf{O}\left(q^2S^2\sqrt{n}\log\left(\frac{n}{\epsilon}\right)\right) \text{ iterations complexity.}$$

5 Conclusion

In this work, we have improved the algorithmic complexity of IPM methods for LO problems by a new kernel function. More specifically, we have proved the large-update and small-update versions of the primal-dual algorithm based on a new kernel function with a logarithmic barrier term defined by (1). This new kernel function has never been mentioned before, and the resulting analysis is also different from others. Moreover, we intend to extend this work in the future to semi-definite linear complementarity problems (SDLCPs) based on this kernel function.

Acknowledgment

The authors would like to express their sincere gratitude to the reviewers for their valuable comments and suggestions, which have significantly improved the quality of this paper.

References

[1] Y.Q. Bai, M. El Ghami and C. Roos. A comparative study of kernel functions for primal-dual interior point algorithms in linear optimization. *SIAM J. Optim.* **15** (1) (2004) 101–128.

[2] Y.Q. Bai, M. El Ghami and C. Roos. A new efficient large-update primal-dual interior-point method based on a finite barrier. *SIAM J. Optim.* **13** (3) (2002) 766–782.

[3] J. Peng, C. Roos and T. Terlaky. Self-regular functions and new search directions for linear and semidefinite optimization. *Math. Programming* **93** (1) (2002) 129–171.

[4] M. Karimi and L. Tunçel. Primal–Dual Interior-Point Methods for Domain-Driven Formulations. *Mathematics of Operations Research* **45** (2) (2020) 591–621.

[5] C. Roos, T. Terlaky and J.P. Vial. *Theory and Algorithms for Linear Optimization, An Interior Point Approach.* Wiley, Chichester, UK, 1997.

[6] F. Leulmi and A. Leulmi. A New Efficient Step-Size in Karmarkar's Projective Interior Point Method for Optimization Problems. *Nonlinear Dyn. Syst. Theory* **24** (1) (2024) 65–79.

[7] S. F. Hafshejani, D. R. Gaur and R. Benkoczi. On Implementing a Two-Step Interior Point Method for Solving Linear Programs. *Algorithms* **17** (7) (2024) 303.

[8] T. Kitahara and T. Tsuchiya. An extension of Chubanov's polynomial-time linear programming algorithm to second-order cone programming. *Optimization Methods and Software* **33** (1) (2018) 1–25.

[9] L. Derbal. Implementation of Infeasible Interior-Point Methods Based on a New Search Direction. *Nonlinear Dynamics and Systems Theory* **23** (2) (2023) 157–166.

[10] M. Bouaafia, D. Benterki and Y. Adnan. An efficient parameterized logarithmic kernel function for linear optimization. *Optim. Lett.* **12** (19) (2018) 1079–1097.

[11] A. Guemmaz, EL. Djeffal and B. Bounibane. A primal-dual ipms for sdo problem based on a new kernel function with a logarithmic barrier term. *Adv. Math. Sci. J.* **11** (19) (2022) 449–467.

[12] S. Guerdouh, Wi. Chikouche and I. Touil. A Primal-Dual Interior-Point Algorithm Based on a Kernel Function with a New Barrier Term. Stat. *Optim. Inf. Comput.* **11** (12) (2023) 773–784.

[13] Z. Darvay and A-S. Orbán. Implementation of an interior-point algorithm for linear complementarity problem working in a wide neighborhood. *Műszaki Tudományos Közlemények* **11** (4) (2019) 47–50.

Nonlinear Dynamics and Systems Theory, **25** (3) (2025) 243–254

A Note on Linear Matrix Functions and Applications

Sihem Guerarra [1*], Souad Allihoum [2] and Shubham Kumar [3]

[1] *Dynamic Systems and Control Laboratory, Department of Mathematics and Informatics, University of Oum El Bouaghi, 04000, Algeria.*
[2] *Mathematics and Computer Science Department, Ghardaia University, Algeria.*
[3] *Mathematics Discipline,* PDPM-Indian Institute of Information Technology, Design and Manufacturing, Jabalpur - 482005, Madhya Pradesh, India.

Received: July 23, 2024; Revised: May 24, 2025

Abstract: This paper focuses on some algebraic characterizations between linear matrix functions (LMFs) and their domains defined over the field of complex numbers $\mathbb{C}$. We discuss the intersection, as well as the inclusion of two domains of some LMFs. By applying specific algebraic methods on ranks and ranges, we consider certain forms of LMFs, where the general solutions can be expressed via specific explicit LMFs to establish some relationships between their domains. As a consequence, we have obtained a well-known result of Lin and Wang.

Keywords: *linear matrix function; algebraic method; generalized inverse; general solutions; rank.*

Mathematics Subject Classification (2020): 15A03, 15A09, 15A24. 93B30, 93B25.

1 Introduction

In this work, we use the notation $\mathbb{C}^{n\times m}$ to represent the set of all $n \times m$ complex matrices. The symbols A^*, $\mathfrak{R}(A)$, $r(A)$ and I_n denote the conjugate transpose, the range, the rank of the matrix A and the identity matrix of order n, respectively. The Moore-Penrose inverse of a matrix $A \in \mathbb{C}^{n\times m}$ is defined as the unique $m \times n$ complex matrix denoted by A^+ satisfying the following four equations:

$$AA^+A = A,\ \ A^+AA^+ = A^+,\ \ \left(AA^+\right)^* = AA^+,\ \ \left(A^+A\right)^* = A^+A. \tag{1}$$

* Corresponding author: mailto:guerarra.siham@univ-oeb.dz

Extensive studies and results regarding the Moore-Penrose inverse can be found in [3,4, 10]. Additionally, we introduce two orthogonal projectors induced by $A \in \mathbb{C}^{m\times n}$, namely $F_A = I_n - A^+A$ and $E_A = I_m - AA^+$.

A linear matrix function

$$Y = f(X_1, X_2, ..., X_p),$$

where X_1, X_2, ..., X_p are the variables over the field of complex numbers $\mathbb{C}$ and Y is the matrix value associated with the matrix function corresponding to $X_1, X_2, \ldots, X_p$. In addition, we define the domain of the function f mentioned above as

$$S = \{Y \mid Y = f(X_1, X_2, ..., X_p)\}.$$

The majority of problems with linear or nonlinear matrix functions should be understood in terms of their analytic or algebraic aspects and behaviors, and used to solve matrix function-related problems in both computational and pure mathematics. Further, matrix equations play an important role in nonlinear dynamics, control engineering, mathematical models, for a variety of reasons, including the analysis, modeling, and simulation of complex systems to linearize nonlinear systems for local analysis, determine stability through eigenvalue analysis, analyze normal modes in oscillatory systems, their use ranges from fundamental stability analysis to advanced control and bifurcation studies. For instance, Baddi et al. [1] studied the stabilization problem of inhomogeneous semilinear control systems; they established the existence and uniqueness of solutions of the system using the semigroup theory. By algebraic method, Tian and Yuan [17] studied and suggested connections between specific LMFs, then explored some specific subjects about the algebraic relationships between the reduced equations and solutions of a certain linear matrix problems. Guerarra [5] investigated the inclusion relationships between the set of persymmetric solutions and the set of minimal rank persymmetric solutions of the quaternion matrix equation $AXA^{(*)} = B$. Özgüler and Akar [9] provided equivalent conditions for the existence of a common solution to a pair of linear matrix equations over a principal ideal domain. Jiang et al. [6] studied the relationships between the set of solutions to $AXB = C$ and the set of solutions of its reduced equations. Therefore, all matrix functions possess a class of fundamental types called LMFs, and they can be defined consistently using matrix additions and multiplications. On the other hand, nonlinear matrix functions have been studied in many works, one may refer to [15,16] and references therein.

Here, we just provide a common illustration of an LMF

$$f(X_1, X_2, \ldots, X_p) = A + B_1X_1C_1 + B_2X_2C_2 + \cdots + B_pX_pC_p,$$

where $A \in \mathbb{C}^{m\times n}$, $B_i \in \mathbb{C}^{m\times l_i}$, $C_i \in \mathbb{C}^{n_i\times n}$ are given, and $X_i \in \mathbb{C}^{l_i\times n_i}$ are matrices with variable entries, where $i = 1, 2, \ldots, p$. Hence, its domain is given as

$$S = \left\{Y = A + B_1X_1C_1 + B_2X_2C_2 + ... + B_pX_pC_p \mid X_i \in \mathbb{C}^{l_i\times n_i},\, i = 1, 2, \ldots, p\right\}.$$

The rank of a matrix is one of the most basic quantities and useful methods and tools that are widely used in linear algebra, specifically in matrix theory. This finite nonnegative integer can be used to represent many properties of matrices such as singularity or nonsingularity of a matrix, identification of matrices, consistency of a matrix equation, etc. For further details, see [2,10,12,13]. The rank of matrices or partitioned matrices was first studied by Matsaglia and Styan [8], where they provided various formulas that simplify complicated matrix expressions or equalities.

Based on the results of Tian and Yuan [17], this work aims to explore and suggest some basic aspects concerning the domains of some specific examples of LMF using the matrix rank method. Because of this fact, we will consider the following new domains of LMFs:

$$S_1 = \left\{A_1 + B_1X_1C_1 \mid X_1 \in \mathbb{C}^{p_1\times n_1}\right\}, \tag{2}$$

$$S_2 = \left\{A_2 + B_2X_2C_2 + B_3X_3C_3 \mid X_2 \in \mathbb{C}^{p_2\times n_2},\ X_3 \in \mathbb{C}^{p_3\times n_3}\right\}, \tag{3}$$

where $A_1,\ A_2 \in \mathbb{C}^{l\times n}$, $B_i \in \mathbb{C}^{l\times p_i}$, $C_i \in \mathbb{C}^{n_i\times n}$, for $i = \overline{1,3}$, are given. This paper is organized as follows. In Section 2, we recall some results. In Section 3, we establish the necessary and sufficient conditions for the two relations $S_1 \cap S_2 \neq \varnothing$, $S_1 \subseteq S_2$ to hold. As a consequence, we give conditions for some matrix equations to have common solutions. We conclude our discussion in Section 4.

2 Preliminaries

To advance this objective, we require the following basic lemmas.

Lemma 1 *[8] Let $A \in \mathbb{C}^{l\times n}$, $D \in \mathbb{C}^{l\times k}$, and $C \in \mathbb{C}^{p\times n}$. Then*

$$r\begin{bmatrix} A & D\end{bmatrix} - r(E_AD) = r(A),\ r\begin{bmatrix} A & D\end{bmatrix} - r(E_DA) = r(D), \tag{4}$$

$$r\begin{bmatrix} A \\ C\end{bmatrix} - r(CF_A) = r(A),\ r\begin{bmatrix} A \\ C\end{bmatrix} - r(AF_C) = r(C), \tag{5}$$

$$r\begin{bmatrix} A & D \\ C & 0\end{bmatrix} - r(E_DAF_C) = r(D) + r(C), \tag{6}$$

from (4)-(6), it follows

$$r\begin{bmatrix} A & BF_P \\ E_QC & 0\end{bmatrix} = r\begin{bmatrix} A & B & 0 \\ C & 0 & Q \\ 0 & P & 0\end{bmatrix} - r(P) - r(Q),$$

$$r\begin{bmatrix} E_{B_1}AF_{C_1} & E_{B_1}B \\ CF_{C_1} & 0\end{bmatrix} = r\begin{bmatrix} A & B & B_1 \\ C & 0 & 0 \\ C_1 & 0 & 0\end{bmatrix} - r(B_1) - r(C_1).$$

Lemma 2 *[10] Consider the matrix equation*

$$AXB = D, \tag{7}$$

where $A \in \mathbb{C}^{l\times n}$, $B \in \mathbb{C}^{p\times q}$, and $D \in \mathbb{C}^{l\times q}$ are given, and $X \in \mathbb{C}^{n\times p}$ is an unknown matrix. Then the following are equivalent:
(i) Eq (7) is consistent.
(ii) $AA^+DB^+B = D$.
(iii) $r\begin{bmatrix} A & D\end{bmatrix} = r(A)$ and $r\begin{bmatrix} B \\ D\end{bmatrix} = r(B)$.
(iv) $\Re(D) \subseteq \Re(A)$ and $\Re(D^*) \subseteq \Re(B^*)$.

In this case, the general solution can be expressed as

$$X = A^{+}DB^{+} + F_A V + U E_B,$$

where V, U *are arbitrary with appropriate sizes. In particular, Eq (7) holds for matrix* $X \in \mathbb{C}^{n\times p}$ *if and only if*

$$\begin{bmatrix} A & D \end{bmatrix} = 0 \text{ or } \begin{bmatrix} B \\ D \end{bmatrix} = 0.$$

Lemma 3 *[9] The matrix equation*

$$A_1X_1B_1 + A_2X_2B_2 = D \tag{8}$$

is solvable for X_1 *and* X_2 *of suitable sizes if and only if all the following equalities*

$$r\begin{bmatrix} D & A_1 & A_2 \end{bmatrix} = r\begin{bmatrix} A_1 & A_2 \end{bmatrix},\ r\begin{bmatrix} D & A_1 \\ B_2 & 0 \end{bmatrix} = r(A_1) + r(B_2),$$
$$r\begin{bmatrix} D & A_2 \\ B_1 & 0 \end{bmatrix} = r(A_2) + r(B_1),\ r\begin{bmatrix} D \\ B_1 \\ B_2 \end{bmatrix} = r\begin{bmatrix} B_1 \\ B_2 \end{bmatrix}$$

hold, or, equivalently,

$$E_A D = 0,\ E_{A_1}DF_{B_2} = 0,\ E_{A_2}DF_{B_1} = 0,\ DF_B = 0 \ \text{ hold},$$

where $A = \begin{bmatrix} A_1 & A_2 \end{bmatrix}$ *and* $B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}$.

Lemma 4 *[12] Eq (8) holds for all* X_1 *and* X_2 *of suitable sizes if and only if any one of the following equalities*

$$\begin{bmatrix} D & A_1 & A_2 \end{bmatrix} = 0,\ \begin{bmatrix} D & A_1 \\ B_2 & 0 \end{bmatrix} = 0,\ \begin{bmatrix} D & A_2 \\ B_1 & 0 \end{bmatrix} = 0,\ \begin{bmatrix} D \\ B_1 \\ B_2 \end{bmatrix} = 0$$

holds.

Lemma 5 *[11] The matrix equation*

$$A_1X_1B_1 + A_2X_2B_2 + A_3X_3B_3 = C \tag{9}$$

is solvable for X_1, X_2 *and* X_3 *of suitable sizes if and only if all the following equalities* *hold*

$$r\begin{bmatrix}A_1 & A_2 & A_3 & C\end{bmatrix} = r\begin{bmatrix}A_1 & A_2 & A_3\end{bmatrix},\ r\begin{bmatrix}B_1\\ B_2\\ B_3\\ C\end{bmatrix} = r\begin{bmatrix}B_1\\ B_2\\ B_3\end{bmatrix},$$

$$r\begin{bmatrix}C & A_1 & A_2\\ B_3 & 0 & 0\end{bmatrix} = r\begin{bmatrix}A_1 & A_2\end{bmatrix} + r(B_3),\ r\begin{bmatrix}C & A_1 & A_3\\ B_2 & 0 & 0\end{bmatrix} = r\begin{bmatrix}A_1 & A_3\end{bmatrix} + r(B_2),$$

$$r\begin{bmatrix}C & A_2 & A_3\\ B_1 & 0 & 0\end{bmatrix} = r\begin{bmatrix}A_2 & A_3\end{bmatrix} + r(B_1),\ r\begin{bmatrix}C & A_3\\ B_1 & 0\\ B_2 & 0\end{bmatrix} = r\begin{bmatrix}B_1\\ B_2\end{bmatrix} + r(A_3),$$

$$r\begin{bmatrix}C & A_2\\ B_1 & 0\\ B_3 & 0\end{bmatrix} = r\begin{bmatrix}B_1\\ B_3\end{bmatrix} + r(A_2),\ r\begin{bmatrix}C & A_1\\ B_2 & 0\\ B_3 & 0\end{bmatrix} = r\begin{bmatrix}B_2\\ B_3\end{bmatrix} + r(A_1),$$

$$r\begin{bmatrix}C & 0 & A_1 & 0 & A_3\\ 0 & -C & 0 & A_2 & A_3\\ B_2 & 0 & 0 & 0 & 0\\ 0 & B_1 & 0 & 0 & 0\\ B_3 & B_3 & 0 & 0 & 0\end{bmatrix} = r\begin{bmatrix}B_2 & 0\\ 0 & B_1\\ B_3 & B_3\end{bmatrix} + r\begin{bmatrix}A_1 & 0 & A_3\\ 0 & A_2 & A_3\end{bmatrix}.$$

Lemma 6 *[18] Let $T \in \mathbb{C}^{l\times n}$, $N \in \mathbb{C}^{l\times p}$, $B \in \mathbb{C}^{p\times k}$ and $D \in \mathbb{C}^{n\times k}$ be given. Then the system of matrix equations $TX = N$ and $XB = D$ has a solution if and only if*

$$TT^{+}N = N,\ \ DB^{+}B = D \quad \text{and} \quad TD = NB.$$

In this case, the general solution can be written as

$$X = T^{+}N + F_T DB^{+} + F_T V E_B,$$

where $V \in C^{n\times p}$ is arbitrary.

Lemma 7 *[14] Define two domains as*

$$\Gamma_1 = \left\{D_1 + B_1X_1C_1 \mid X_1 \in \mathbb{C}^{s_1\times t_1}\right\} \text{ and } \Gamma_2 = \left\{D_2 + B_2X_2C_2 \mid X_2 \in \mathbb{C}^{s_2\times t_2}\right\},$$

where $D_i \in \mathbb{C}^{l\times n}$, $B_i \in \mathbb{C}^{l\times s_i}$, and $C_i \in \mathbb{C}^{t_i\times n}$ are given, and $X_i \in \mathbb{C}^{s_i\times t_i}$ are variable for $i = 1, 2$. Then
(a) $\Gamma_1 \cap \Gamma_2 \neq \varnothing$ if and only if all the following conditions hold:

$$\mathfrak{R}(D_1 - D_2) \subseteq \mathfrak{R}\begin{bmatrix}B_1 & B_2\end{bmatrix},\ \mathfrak{R}(D_1^* - D_2^*) \subseteq \mathfrak{R}\begin{bmatrix}C_1^* & C_2^*\end{bmatrix},$$

$$r\begin{bmatrix}D_1 - D_2 & B_1\\ C_2 & 0\end{bmatrix} = r(B_1) + r(C_2),\ r\begin{bmatrix}D_1 - D_2 & B_2\\ C_1 & 0\end{bmatrix} = r(B_2) + r(C_1).$$

(b) $\Gamma_1 \subseteq \Gamma_2$ if and only if

$$\mathfrak{R}\begin{bmatrix}D_1 - D_2 & B_1\end{bmatrix} \subseteq \mathfrak{R}(B_2) \text{ and } \mathfrak{R}\begin{bmatrix}D_1^* - D_2^* & C_1^*\end{bmatrix} \subseteq \mathfrak{R}(C_2^*).$$

(c) $\Gamma_1 = \Gamma_2$ if and only if

$$\mathfrak{R}(D_1 - D_2) \subseteq \mathfrak{R}(B_1) = \mathfrak{R}(B_2) \text{ and } \mathfrak{R}(D_1^* - D_2^*) \subseteq \mathfrak{R}(C_1^*) = \mathfrak{R}(C_2^*).$$

3 Relationship between Linear Matrix Functions

In this section, we consider two domains given in (2) and (3), we discuss the necessary and sufficient conditions for two relations $S_1 \cap S_2 \neq \varnothing$, $S_1 \subseteq S_2$ to hold. We also present connections between two domains of some well known linear matrix functions.

Theorem 8 *Let S_1 and S_2 be as given in (2) and (3), respectively. Then*
(a) $S_1 \cap S_2 \neq \varnothing$ if and only if all the following equalities hold:

$$r\begin{bmatrix} A_2 - A_1 & B_1 & B_2 & B_3 \end{bmatrix} = r\begin{bmatrix} B_1 & B_2 & B_3 \end{bmatrix},\ r\begin{bmatrix} A_2 - A_1 \\ C_1 \\ C_2 \\ C_3 \end{bmatrix} = r\begin{bmatrix} C_1 \\ C_2 \\ C_3 \end{bmatrix},$$

$$r\begin{bmatrix} A_2 - A_1 & B_1 & B_2 \\ C_3 & 0 & 0 \end{bmatrix} = r\begin{bmatrix} B_1 & B_2 \end{bmatrix} + r(C_3),$$

$$r\begin{bmatrix} A_2 - A_1 & B_1 & B_3 \\ C_2 & 0 & 0 \end{bmatrix} = r\begin{bmatrix} B_1 & B_3 \end{bmatrix} + r(C_2),$$

$$r\begin{bmatrix} A_2 - A_1 & B_2 & B_3 \\ C_1 & 0 & 0 \end{bmatrix} = r\begin{bmatrix} B_2 & B_3 \end{bmatrix} + r(C_1),$$

$$r\begin{bmatrix} A_2 - A_1 & B_3 \\ C_1 & 0 \\ C_2 & 0 \end{bmatrix} = r\begin{bmatrix} C_1 \\ C_2 \end{bmatrix} + r(B_3),\ r\begin{bmatrix} A_2 - A_1 & B_2 \\ C_1 & 0 \\ C_3 & 0 \end{bmatrix} = r\begin{bmatrix} C_1 \\ C_3 \end{bmatrix} + r(B_2),$$

$$r\begin{bmatrix} A_2 - A_1 & B_1 \\ C_2 & 0 \\ C_3 & 0 \end{bmatrix} = r\begin{bmatrix} C_2 \\ C_3 \end{bmatrix} + r(B_1),$$

$$r\begin{bmatrix} A_2 - A_1 & 0 & B_1 & 0 & B_3 \\ 0 & A_1 - A_2 & 0 & B_2 & B_3 \\ C_2 & 0 & 0 & 0 & 0 \\ 0 & C_1 & 0 & 0 & 0 \\ C_3 & C_3 & 0 & 0 & 0 \end{bmatrix} = r\begin{bmatrix} C_2 & 0 \\ 0 & C_1 \\ C_3 & C_3 \end{bmatrix} + r\begin{bmatrix} B_1 & 0 & B_3 \\ 0 & B_2 & B_3 \end{bmatrix}.$$

(b) $S_1 \subseteq S_2$ if and only if any one of the following equalities holds:

$$\begin{gathered} r\begin{bmatrix} B_2 & B_3 \end{bmatrix} = l \quad or\ r\begin{bmatrix} A_2 - A_1 & B_1 & B_2 & B_3 \end{bmatrix} = r\begin{bmatrix} B_2 & B_3 \end{bmatrix}, \\ or\ r\begin{bmatrix} A_2 - A_1 & B_2 & B_3 \\ C_1 & 0 & 0 \end{bmatrix} = r\begin{bmatrix} B_2 & B_3 \end{bmatrix}, \end{gathered} \tag{10}$$

$$\begin{gathered} r(B_2) = l \quad or\ r\begin{bmatrix} A_2 - A_1 & B_1 & B_2 \\ C_3 & 0 & 0 \end{bmatrix} = r(B_2) + r(C_3), \\ or\ r\begin{bmatrix} A_2 - A_1 & B_2 \\ C_1 & 0 \\ C_3 & 0 \end{bmatrix} = r(C_3) + r(B_2) \quad or\ r(C_3) = n, \end{gathered} \tag{11}$$

$$\begin{gathered} r(B_3) = l \quad or\ r\begin{bmatrix} A_2 - A_1 & B_1 & B_3 \\ C_2 & 0 & 0 \end{bmatrix} = r(B_3) + r(C_2), \\ or\ r\begin{bmatrix} A_2 - A_1 & B_3 \\ C_1 & 0 \\ C_2 & 0 \end{bmatrix} = r(C_2) + r(B_3) \quad or\ r(C_2) = n, \end{gathered} \tag{12}$$

$$r\begin{bmatrix} A_2-A_1 & B_1 \\ C_2 & 0 \\ C_3 & 0 \end{bmatrix} = r\begin{bmatrix} C_2 \\ C_3 \end{bmatrix} \quad or\; r\begin{bmatrix} A_2-A_1 \\ C_1 \\ C_2 \\ C_3 \end{bmatrix} = r\begin{bmatrix} C_2 \\ C_3 \end{bmatrix} \quad or\; r\begin{bmatrix} C_2 \\ C_3 \end{bmatrix} = n. \tag{13}$$

Proof. (a) The intersection $S_1 \cap S_2 \neq \varnothing$ is obviously equivalent to

$$A_1 + B_1X_1C_1 = A_2 + B_2X_2C_2 + B_3X_3C_3. \tag{14}$$

Eq (14) can be written as

$$B_1X_1C_1 - B_2X_2C_2 - B_3X_3C_3 = A_2 - A_1. \tag{15}$$

By applying Lemma 5 to the Eq (15), we get (a).

(b) Eq (14) can be written as

$$B_2X_2C_2 + B_3X_3C_3 = A_1 - A_2 + B_1X_1C_1. \tag{16}$$

From Lemma 3, Eq (16) holds for two matrices X_2 and X_3 if and only if all the following four conditions hold:

$$E_{[B_2,B_3]}(A_1 - A_2 + B_1X_1C_1) = 0, \tag{17}$$
$$E_{B_2}(A_1 - A_2 + B_1X_1C_1)F_{C_3} = 0, \tag{18}$$
$$E_{B_3}(A_1 - A_2 + B_1X_1C_1)F_{C_2} = 0, \tag{19}$$
$$((A_1 - A_2) + B_1X_1C_1)F_Z = 0, \tag{20}$$

where $Z = \begin{bmatrix} C_2 \\ C_3 \end{bmatrix}$. By Lemma 2, Eq (17) holds for all X_1 if and only if

$$E_{[B_2,B_3]} = 0 \;\text{ or }\; \begin{bmatrix} E_{[B_2,B_3]}B_1 & E_{[B_2,B_3]}(A_2-A_1) \end{bmatrix} = 0 \;\text{ or }\; \begin{bmatrix} C_1 \\ E_{[B_2,B_3]}(A_2-A_1) \end{bmatrix} = 0,$$

which are equivalent, respectively, to

$$r\begin{bmatrix} B_2 & B_3 \end{bmatrix} = l \;\text{ or } r\begin{bmatrix} A_2-A_1 & B_1 & B_2 & B_3 \end{bmatrix} = r\begin{bmatrix} B_2 & B_3 \end{bmatrix},$$
$$\text{or } r\begin{bmatrix} A_2-A_1 & B_2 & B_3 \\ C_1 & 0 & 0 \end{bmatrix} = r\begin{bmatrix} B_2 & B_3 \end{bmatrix}.$$

This proves (10). Eq (18) holds for all X_1 if and only if

$$E_{B_2} = 0 \;\text{ or }\; \begin{bmatrix} E_{B_2}B_1 & E_{B_2}(A_2-A_1)F_{C_3} \end{bmatrix} = 0 \;\text{ or }\; \begin{bmatrix} C_1F_{C_3} \\ E_{B_2}(A_2-A_1)F_{C_3} \end{bmatrix} \;\text{ or } F_{C_3} = 0,$$

which, in consequence, is equivalent to

$$r(B_2) = l \;\text{ or } r\begin{bmatrix} A_2-A_1 & B_1 & B_2 \\ C_3 & 0 & 0 \end{bmatrix} = r(B_2) + r(C_3),$$
$$\text{or } r\begin{bmatrix} A_2-A_1 & B_2 \\ C_1 & 0 \\ C_3 & 0 \end{bmatrix} = r(C_3) + r(B_2) \;\text{ or } r(C_3) = n.$$

Then (11) holds. Similarly, Eq (19) holds for all X_1 if and only if

$$r\left(B_3\right)=l \text{ or } r\begin{bmatrix} A_2-A_1 & B_1 & B_3 \\ C_2 & 0 & 0 \end{bmatrix}=r\left(B_3\right)+r\left(C_2\right),$$

$$\text{or } r\begin{bmatrix} A_2-A_1 & B_3 \\ C_1 & 0 \\ C_2 & 0 \end{bmatrix}=r\left(C_2\right)+r\left(B_3\right) \text{ or } r\left(C_2\right)=n.$$

Then we get (12). Eq (20) holds for all X_1 if and only if

$$\begin{bmatrix} B_1 & (A_2-A_1)F_Z \end{bmatrix}=0 \text{ or } \begin{bmatrix} C_1F_Z \\ (A_2-A_1)F_Z \end{bmatrix}=0 \text{ or } F_Z=0,$$

which then is equivalent to

$$r\begin{bmatrix} A_2-A_1 & B_1 \\ C_2 & 0 \\ C_3 & 0 \end{bmatrix}=r\begin{bmatrix} C_2 \\ C_3 \end{bmatrix} \text{ or } r\begin{bmatrix} A_2-A_1 \\ C_1 \\ C_2 \\ C_3 \end{bmatrix}=r\begin{bmatrix} C_2 \\ C_3 \end{bmatrix} \text{ or } r\begin{bmatrix} C_2 \\ C_3 \end{bmatrix}=n,$$

which proves (13). Hence we establish (b).

Setting $B_3=I_p$, $C_2=I_n$ in Theorem 8, we get the following result.

Corollary 9 *Consider two domains of two linear matrix functions*

$$S_1=\left\{A_1+B_1X_1C_1 \mid X_1\in\mathbb{C}^{p_1\times n_1}\right\},$$
$$S_2=\left\{A_2+B_2X_2+X_3C_3 \mid X_2\in\mathbb{C}^{p_2\times n},\ X_3\in\mathbb{C}^{l\times p_3}\right\},$$

where $A_1,\ A_2\in\mathbb{C}^{l\times n},\ B_1\in\mathbb{C}^{l\times p_1},\ B_2\in\mathbb{C}^{l\times p_2}$ *and* $C_1\in\mathbb{C}^{n_1\times n},\ C_3\in\mathbb{C}^{p_3\times n}$ *are known matrices. Then*

(a) $S_1\cap S_2\neq\varnothing$ *if and only if the following rank equalities hold:*

$$r\begin{bmatrix} A_2-A_1 & B_1 & B_2 \\ C_3 & 0 & 0 \end{bmatrix}=r\begin{bmatrix} B_1 & B_2 \end{bmatrix}+r(C_3),$$

$$r\begin{bmatrix} A_2-A_1 & B_2 \\ C_1 & 0 \\ C_3 & 0 \end{bmatrix}=r\begin{bmatrix} C_1 \\ C_3 \end{bmatrix}+r(B_2),$$

$$r\begin{bmatrix} A_2-A_1 & B_1 & B_2 \\ C_1 & 0 & 0 \\ C_3 & 0 & 0 \end{bmatrix}=r\begin{bmatrix} C_1 \\ C_3 \end{bmatrix}+r\begin{bmatrix} B_1 & B_2 \end{bmatrix}.$$

(b) $S_1\subseteq S_2$ *if and only if*

$$r\left(B_2\right)=l \text{ or } r\begin{bmatrix} A_2-A_1 & B_1 & B_2 \\ C_3 & 0 & 0 \end{bmatrix}=r\left(B_2\right)+\left(C_3\right)$$

$$\text{or } r\begin{bmatrix} A_2-A_1 & B_2 \\ C_1 & 0 \\ C_3 & 0 \end{bmatrix}=r\left(C_3\right)+r\left(B_2\right) \text{ or } r\left(C_3\right)=n.$$

From Colrollary 9, we can deduce the following result.

Corollary 10 *Let $A_4 \in \mathbb{C}^{l\times n}$, $B_4 \in \mathbb{C}^{p\times s}$, $C_4 \in \mathbb{C}^{l\times p}$, $D_4 \in \mathbb{C}^{n\times s}$, $A_5 \in \mathbb{C}^{k\times t}$, $B_5 \in \mathbb{C}^{k\times n}$, $C_5 \in \mathbb{C}^{p\times t}$be given, X_4, $X_5 \in \mathbb{C}^{n\times p}$ be unknown matrices, and assume that the system $A_4X = C_4$, $XB_4 = D_4$, and the matrix equation $B_5XC_5 = A_5$ are solvable for X_4 and X_5, respectively. Denote*

$$S_1 = \left\{X_4 \in \mathbb{C}^{n\times p} \mid \ A_4X_4 = C_4,\ X_4B_4 = D_4\right\}, \tag{21}$$

$$S_2 = \left\{X_5 \in \mathbb{C}^{n\times p} \mid B_5X_5C_5 = A_5\right\}. \tag{22}$$

Then

(a) $S_1 \cap S_2 \neq \varnothing$, that is, the system $A_4X_4 = C_4$, $X_4B_4 = D_4$ and $B_5X_5C_5 = A_5$ have a common solution if and only if

$$r\begin{bmatrix} A_4 & C_4C_5 \\ B_5 & A_5 \end{bmatrix} = r\begin{bmatrix} A_4 \\ B_5 \end{bmatrix},$$

$$r\begin{bmatrix} B_4 & C_5 \\ B_5D_4 & A_5 \end{bmatrix} = r\begin{bmatrix} B_4 & C_5 \end{bmatrix},$$

$$r\begin{bmatrix} 0 & B_4 & C_5 \\ A_4 & -C_4B_4 & 0 \\ B_5 & 0 & A_5 \end{bmatrix} = r\begin{bmatrix} B_4 & C_5 \end{bmatrix} + r\begin{bmatrix} A_4 \\ B_5 \end{bmatrix}.$$

(b) $S_1 \subseteq S_2$, that is, all solutions of $A_4X_4 = C_4$, $X_4B_4 = D_4$ are solutions of $B_5X_5C_5 = A_5$ if and only if

$$r\begin{bmatrix} A_4 & C_4C_5 \\ B_5 & A_5 \end{bmatrix} = r(A_4) \quad or \quad r\begin{bmatrix} B_4 & C_5 \\ B_5D_4 & A_5 \end{bmatrix} = r(B_4).$$

Proof. It follows from Lemmas 6 and 2 that, the solutions of system $A_4X_4 = C_4$, $X_4B_4 = D_4$ and equation $B_5X_5C_5 = A_5$ can be expressed, respectively, as

$$X_4 = A_4^+C_4 + F_{A_4}D_4B_4^+ + F_{A_4}VE_{B_4},$$
$$X_5 = B_5^+A_5C_5^+ + F_{B_5}U + WE_{C_5},$$

where V, U and W are arbitrary. So, two sets in (21) and (22) can be represented, respectively, as

$$S_1 = \left\{A_4^+C_4 + F_{A_4}D_4B_4^+ + F_{A_4}VE_{B_4}\right\},$$
$$S_2 = \left\{B_5^+A_5C_5^+ + F_{B_5}U + WE_{C_5}\right\}.$$

From Corollary 9, the relation $S_1 \cap S_2 \neq \varnothing$ holds if and only if the following equalities hold:

$$r\begin{bmatrix} B_5^+A_5C_5^+ - A_4^+C_4 - F_{A_4}D_4B_4^+ & F_{A_4} & F_{B_5} \\ E_{C_5} & 0 & 0 \end{bmatrix} = r\begin{bmatrix} F_{A_4} & F_{B_5} \end{bmatrix} + r(E_{C_5}), \tag{23}$$

$$r\begin{bmatrix} B_5^+A_5C_5^+ - A_4^+C_4 - F_{A_4}D_4B_4^+ & F_{B_5} \\ E_{B_4} & 0 \\ E_{C_5} & 0 \end{bmatrix} = r\begin{bmatrix} E_{B_4} \\ E_{C_5} \end{bmatrix} + r(F_{B_5}), \tag{24}$$

$$r\begin{bmatrix} B_5^+A_5C_5^+ - A_4^+C_4 - F_{A_4}D_4B_4^+ & F_{A_4} & F_{B_5} \\ E_{B_4} & 0 & 0 \\ E_{C_5} & 0 & 0 \end{bmatrix} = r\begin{bmatrix} E_{B_4} \\ E_{C_5} \end{bmatrix} + r\begin{bmatrix} F_{A_4} & F_{B_5} \end{bmatrix}. \tag{25}$$

By Lemma 1, and simplifying by $C_4B_4 = A_4D_4$, $A_4A_4^+C_4 = C_4$, $D_4B_4^+B_4 = D_4$, $B_5B_5^+A_5 = A_5$, $A_5C_5^+C_5 = A_5$, we find that the rank equalities in (23)-(25) are equivalent, respectively, to

$$\begin{aligned} & r\begin{bmatrix} A_4 & C_4C_5 \\ B_5 & A_5 \end{bmatrix} = r\begin{bmatrix} A_4 \\ B_5 \end{bmatrix}, \\ & r\begin{bmatrix} B_4 & C_5 \\ B_5D_4 & A_5 \end{bmatrix} = r\begin{bmatrix} B_4 & C_5 \end{bmatrix}, \\ & r\begin{bmatrix} 0 & B_4 & C_5 \\ A_4 & -C_4B_4 & 0 \\ B_5 & 0 & A_5 \end{bmatrix} = r\begin{bmatrix} B_4 & C_5 \end{bmatrix} + r\begin{bmatrix} A_4 \\ B_5 \end{bmatrix}. \end{aligned}$$

Thus (a) is proved.

(b) $S_1 \subseteq S_2$ holds if and only if

$$\begin{aligned} & r\begin{bmatrix} B_5^+A_5C_5^+ - A_4^+C_4 - F_{A_4}D_4B_4^+ & F_{A_4} & F_{B_5} \\ E_{C_5} & 0 & 0 \end{bmatrix} = r(F_{B_5}) + r(E_{C_5}), \\ \text{or}\quad & r\begin{bmatrix} B_5^+A_5C_5^+ - A_4^+C_4 - F_{A_4}D_4B_4^+ & F_{B_5} \\ E_{B_4} & 0 \\ E_{C_5} & 0 \end{bmatrix} = r(E_{C_5}) + r(F_{B_5}), \end{aligned}$$

which then is equivalent to

$$r\begin{bmatrix} A_4 & C_4C_5 \\ B_5 & A_5 \end{bmatrix} = r(A_4) \ \text{ or } \ r\begin{bmatrix} B_4 & C_5 \\ B_5D_4 & A_5 \end{bmatrix} = r(B_4).$$

Then we establish (b).

Remark 3.1 *Result (a) of Corollary 10 is the same as in [7, Theorem 2.4].*

4 Conclusion

In this study, we discussed and examined some fundamental questions associated with the connections between two domains of linear matrix functions and specific types of linear matrix equations. The general solutions can be expressed via particular explicit linear matrix functions to establish some connections between their domains through the methodical application of various established or well-known relations to ranks and ranges of matrices. Thus, they show that a variety of matrix equality and matrix set inclusion problems may be solved with the help of the matrix rank and range method.

Acknowledgment

The authors are grateful for the detailed comments from the referees, which significantly improved the quality of the paper.

References

[1] M. Baddi, M. Chqondi and Y. Akdim. Exponential and Strong Stabilization for Inhomogeneous Semilinear Control Systems by Decomposition Method. *Nonlinear Dyn. Syst. Theory* **23** (1) (2023) 1–13.

[2] R. Belkhiri and S. Guerarra. Some structures of submatrices in solution to the paire of matrix equations $AX = C, XB = D$. *Math. Found. Comput.* **6** (2) (2023) 231–252.

[3] A. Ben-Israel and T. N. E. Greville. *Generalized Inverses: Theory and Applications.* 2^{nd} *ed*. Springer, 2003.

[4] S. L. Campbell and C. D. Meyer. Generalized inverses of linear transformations. *SIAM* (2009).

[5] S. Guerarra. Relationship between Persymmetric Solutions and Minimal Persymmetric Solutions of $AXA^{(*)} = B$. *Nonlinear Dyn. Syst. Theory* **24** (5) (2024) 485–494.

[6] B. Jiang, Y. Tian and R. Yuan. On Relationships between a Linear Matrix Equation and Its Four Reduced Equations. *Axioms.* **11** (9) (2022) 440.

[7] C. Lin and Q. W. Wang. New Solvable Conditions and a New Expression of the General Solution to a System of Linear Matrix Equations over an Arbitrary Division Ring. *Southeast Asian Bull. Math.* **29** (4) (2005) 131–142.

[8] G. Matsaglia and G.P.H. Styan. Equalities and inequalities for ranks of matrices. *Linear Multilinear Algebra* **2** (1974) 269–292.

[9] A. Özgüler Bülent and Akar, Nail. A common solution to a pair of linear matrix equations over a principal ideal domain. *Linear Algebra & Appl.* **144** (1991) 85–99.

[10] R. Penrose. A generalized inverse for matrices. *Mathematical proceedings of the Cambridge philosophical society* **51** (3) (1955) 406–413.

[11] Y. Tian. The solvability of two linear matrix equations. *Linear Multilinear Algebra* **48** (2) (2000) 123–147.

[12] Y. Tian. Upper and lower bounds for ranks of matrix expressions using generalized inverses. *Linear Algebra & Appl.* **355** (1-3) (2002) 187–214.

[13] Y. Tian. Equalities and inequalities for inertias of Hermitian matrices with applications. *Linear Algebra & Appl.* **433** (1) (2010) 263–296.

[14] Y. Tian. Relations between matrix sets generated from linear matrix expressions and their applications. *Comput. Math. Appl.* **61** (6) (2011) 1493–1501.

[15] Y. Tian. Formulas for calculating the extremum ranks and inertias of a four-term quadratic matrix-valued function and their applications. *Linear Algebra & Appl.* **437** (3) (2012) 835–859.

[16] Y. Tian. Some optimization problems on ranks and inertias of matrix-valued functions subject to linear matrix equation restrictions. *Banach J. Math. Anal.* **8** (1) (2014) 148–178.

[17] Y. Tian, Yongge and R. Yuan. Algebraic Characterizations of Relationships between Different Linear Matrix Functions. *Mathematics* **11** (3) (2023) 756.

[18] Q. W. Wang. The general solution to a system of real quaternion matrix equations. *Comput. Math. Appl.* **49** (5-6) (2005) 665–675.

Nonlinear Dynamics and Systems Theory, **25** (3) (2025) 255–265

Analysis and Existence of Optimal Control in Industrial Economic Growth with Investment Using the Ramsey-Cass-Koopmans Model

Alvian Alif Hidayatullah, Subchan Subchan * and Devi Try Lestari

Department of Mathematics, Faculty of Science and Data Analytics, Institut Teknologi Sepuluh Nopember

Received: July 14, 2024; Revised: May 20, 2025

Abstract: Economic growth is associated with an increase in the production of goods and services. Consumption and investment influence the increased production of goods and services. Consumption parameters can be assessed based on utility, while investments can be affected by the level of capital stock. This paper applies a modification and analysis of the Ramsey-Cass-Koopmans model to the economic growth of two industries, focusing on investment strategies to maximize consumption utility. The analysis of the Ramsey-Cass-Koopmans model showed that the model is valid as it has a positive and unique solution. This study performs optimal control by maximizing the consumption utility of each industry, where the control is given in the form of per capita consumption. In this paper, consumption control can be interpreted as a form of savings. In addition, the existence of optimal control is proven, indicating that the problem can be solved.

Keywords: *Ramsey-Cass-Koopmans model; utility; optimal control; investment.*

Mathematics Subject Classification (2020): 49J15, 49K15, 49L20, 93-10.

* Corresponding author: mailto:subchan@its.ac.id

1 Introduction

A country's development is determined by economic growth. This is demonstrated by increasing a country's ability to provide goods and services to its population [1]. Consumption and investment drive the growth in the production of goods and services. Consumption and investment factors cannot be separated. Therefore, consumption and investment are interconnected and influence each other. Consumption allows firms to generate the income needed to increase investment in the capital stock, while investment can increase production. Utility is the value or benefit obtained from consumption activities or use of goods and services. Furthermore, utility is an essential factor that influences consumption. The importance of consumption utility lies in its ability to increase productivity and efficiency in the production of goods and services [2]. One approach to maximizing utility is modelling economic growth problems using the Ramsey model, which can then be analyzed using optimal control theory.

Optimal control theory focuses on determining controls that influence processes while adhering to specific constraints [3–5]. Optimal control theory also serves as an alternative for solving economic growth problems, including those related to the Ramsey model. The Ramsey model was first introduced in 1928 by Frank P. Ramsey [6]. The Ramsey model is a neoclassical economic growth model that maximizes the utility of capital-bound consumption under dynamic constraints. David Cass and Tjalling Koopmans further developed this model in separate works, and it is now known as the Ramsey-Cass-Koopmans model [7, 8]. In their developments, several previous studies have used this model [9–13]. Olivia Bundau and Adina Juratoni [14] discussed the Ramsey growth model in infinite and continuous time with the aim of maximizing global utility using the Pontryagin Maximum Principle. Then Kajanovičová et al. [15] discussed optimal control of the Ramsey-Cass-Koopmans economic growth model with non-constant population growth using the Maximum Principle, which aims to maximize consumption utility with control in the form of per capita consumption. Further research by Frerick et al. [16] discussed the multi-object Ramsey-Cass-Koopmans model for Ramsey-type equilibrium problems with heterogeneous agents.

This paper discusses a modification and analysis of the nonlinear dynamics of the Ramsey-Cass-Koopmans model of the economic growth of two industries that are interrelated by investment. The optimal control is aimed to maximize the utility of the amount of consumed production. The analyses conducted in this study are the positivity analysis, the uniqueness analysis, and the existence of optimal control [4]. Positivity and uniqueness aim to validate the model, while the existence of optimal control verifies whether a control that maximizes utility exists. The control variable for maximizing utility is defined as per capita consumption. When consumption is controlled, consumption expenditure is reduced, and the remaining production output can be reinvested or saved as savings.

2 Mathematical Model of Economic Growth with Investment

The mathematical model of economic growth used in this paper is a modification of the multi-object Ramsey-Cass-Koopmans model [16]. In this paper, it is assumed that the second industry has a high demand for goods. Thus, the first industry provides some of its capital by investing in the second industry. The investment return is assumed to be a profit of 5% from the investment, and this problem can be illustrated as follows.

Individual A owns and manages two industries in different regions, namely, the first industry in region X and the second industry in region Y. In the first industry, consumer demand for goods in region X is not too high; thus, the first industry can invest in the second industry. While consumer demand for goods in region Y in the second industry exceeds the demand for goods in the first industry in region X, the first industry helps by providing capital or investing in the second industry. With this additional capital, it is expected to maximize *output* or production results. For example, if individual A owns and manages two industries and has problems in one industry, then individual A can solve the problems in the first industry. This process aims to maximize consumption utility and increase capital stock growth in both industries.

The relationship between the two industries can be formulated by the following mathematical model of economic growth:

$$\begin{aligned} \frac{dK_1(t)}{dt} &= F_1(\mathcal{A}_1, K_1(t), L_1(t)) - \delta_1 K_1(t) + \delta_2 K_2(t) - C_1(t), \\ \frac{dK_2(t)}{dt} &= F_2(\mathcal{A}_2, K_2(t), L_2(t)) + \delta_1 K_1(t) - \delta_2 K_2(t) - C_2(t) \end{aligned} \tag{1}$$

with

$K_1(t)$	:	First industry capital stock at time t
$K_2(t)$	:	Second industry capital stock at time t
$L_1(t)$	:	Total labor of the first industry at time t
$L_2(t)$	:	Total labor of the second industry at time t
$C_1(t)$	:	Amount of production output consumed by the first industry at time t
$C_2(t)$	:	Amount of production output consumed by the second industry at time t
δ_1	:	Investment rate
δ_2	:	Investment return rate
$F_1(\mathcal{A}_1, K_1(t), L_1(t))$	:	First industry output
$F_2(\mathcal{A}_2, K_2(t), L_2(t))$	:	Second industry output
$\mathcal{A}$	:	Technological advancement factor.

$K(t)$ and $C(t)$ are continuous functions, and the production function F used in the model is the Cobb-Douglas production function

$$F(\mathcal{A}, K(t), L(t)) = \mathcal{A}K(t)^{\alpha}L(t)^{1-\alpha} \tag{2}$$

with $\mathcal{A} > 0$ being a constant. The production output (F) describes the relationship between the technological advancement factor ($\mathcal{A}$), the capital stock (K), and the amount of labor (L) with $\dot{L}(t) = nL(t)$, where $L(t)$ experiences exponential growth with a constant growth rate of the amount of labor (n).

In economic analysis, to enable more accurate and fair comparisons between groups with different populations, it is necessary to convert the Equation (1) into per capita model:

- Capital stock per capita (k):

$$k(t) = \frac{K(t)}{L(t)},$$
$$K(t) = k(t)L(t).$$

Then

$$\dot{K}(t) = \dot{k}(t)L(t) + k(t)\dot{L}(t),$$
$$\dot{K}(t) = \dot{k}(t)L(t) + k(t)nL(t).$$

- Amount of production *output* consumed per capita (c):
$$c(t) = \frac{C(t)}{L(t)}.$$
- Production output per capita (f):
$$f(t) = \frac{F(t)}{L(t)},$$
where F is
$$\begin{aligned} F(t) &= \mathcal{A}K(t)^{\alpha}L(t)^{1-\alpha} \\ &= \mathcal{A}k(t)^{\alpha}L(t)^{\alpha}L(t)^{1-\alpha} \\ &= \mathcal{A}k(t)^{\alpha}L(t). \end{aligned}$$
Therefore, the output of per capita production is
$$f(t) = \frac{F(t)}{L(t)} = \frac{\mathcal{A}k(t)^{\alpha}L(t)}{L(t)} = \mathcal{A}k(t)^{\alpha}.$$

Furthermore, assume labor $L(t) = L_1(t) = L_2(t)$, then Equation (1) becomes:

1. The capital stock of the first industry (k_1) can be given as follows:
$$(\dot{k}_1(t) + k_1(t)n_1)L(t) = (f_1(t) - \delta_1 k_1(t) + \delta_2 k_2(t) - c_1(t))L(t).$$
Then, simplify both segments by multiplying by $\frac{1}{L(t)}$:
$$\dot{k}_1(t) = f_1(t) - \delta_1 k_1(t) + \delta_2 k_2(t) - k_1(t)n_1 - c_1(t).$$
Substitute $f_1(t) = \mathcal{A}_1 k_1(t)^{\alpha_1}$ so that
$$\dot{k}_1(t) = \mathcal{A}_1 k_1(t)^{\alpha_1} - \delta_1 k_1(t) + \delta_2 k_2(t) - k_1(t)n_1 - c_1(t).$$

2. The capital stock of the second industry (k_2) can be given as follows:
$$(\dot{k}_2(t) + k_2(t)n_2)L(t) = (f_2(t) + \delta_1 k_1(t) - \delta_2 k_2(t) - c_2(t))L(t).$$
Then, simplify both segments by multiplying by $\frac{1}{L(t)}$:
$$\dot{k}_2(t) = f_2(t) + \delta_1 k_1 - \delta_2 k_2(t) - k_2(t)n_2 - c_2(t).$$
Substitute $f_2(t) = \mathcal{A}_2 k_2(t)^{\alpha_2}$ so that
$$\dot{k}_2(t) = \mathcal{A}_2 k_2(t)^{\alpha_2} + \delta_1 k_1(t) - \delta_2 k_2 - k_2(t)n_2 - c_2(t).$$

Thus, Equation (1) can be expressed as

$$\begin{array}{rcl} \dot{k}_1(t) &=& \mathcal{A}_1 k_1(t)^{\alpha_1} - \delta_1 k_1(t) + \delta_2 k_2(t) - n_1 k_1(t) - c_1(t), \\ \dot{k}_2(t) &=& \mathcal{A}_2 k_2(t)^{\alpha_2} + \delta_1 k_1(t) - \delta_2 k_2(t) - n_2 k_2(t) - c_2(t) \end{array} \qquad (3)$$

for $k_i(0) = k_{i0},\ i = 1, 2$.

3 Positive and Unique Solution

It can be seen from Equation (3) that the model is considered valid if it has a positive solution at any time t. It means that if the model has initial conditions $k_1(t_0) > 0$ and $k_2(t_0) > 0$, then $k_1(t) > 0$ and $k_2(t) > 0$ for every $t > t_0$. First, it will be shown that Equation (3) is valid. Suppose X is the set of all $x(t) = (k_1, k_2)$ for each time t as the solution of a controlled model with the initial conditions $x(t_0) = (k_1(t_0), k_2(t_0))$ and the set

$$\Omega_{(k_1(t_0),k_2(t_0))} := \{k_1(t), k_2(t)|t_0 \leq t \leq t_f, 0 < k_1(t), k_2(t)\}\,. \tag{4}$$

If the initial conditions in the Equation (3) satisfy $k_1(t_0 = 0) > 0$ and $k_2(t_0 = 0) > 0$, then it can be said that the Equation (3) is valid if the set $\Omega_{(k_1(t_0),k_2(t_0))}$ is a positive invariant set. The definition of a positive invariant set can be given as follows.

Definition 3.1 *(Positively invariant set). Let $\dot{\boldsymbol{x}} = \boldsymbol{f}(t, \boldsymbol{x})$ be a dynamic system with the initial conditions $\boldsymbol{x}_0 = \boldsymbol{x}(t_0)$. Suppose Ω is a subset of $\mathbb{R}^n$. Then Ω is said to be a positive invariant set if $\boldsymbol{x}_0 \in \Omega$ implies $\boldsymbol{x}(t, x_0) \in \Omega$ for every $t \geq t_0$.*

The positive invariant set of (4) can be proven by the following theorem.

Theorem 3.1 *Let*

$$\Omega_{(k_1(t_0),k_2(t_0))} := \{k_1(t), k_2(t)|t_0 \leq t \leq t_f, 0 < k_1(t), k_2(t)\}$$

be a subset of all solutions of the Equation (3) with the initial conditions $k_1(t_0 = 0), k_2(t_0 = 0)$. If $k_1(t_0), k_2(t_0) > 0$, then $\Omega_{(k_1(t_0),k_1(t_0))}$ is a positive invariant set.

Proof. Define the functions $\dot{k}_1(t)$ and $\dot{k}_2(t)$ as follows:

$$\begin{aligned} \dot{k}_1(t) &= f_1(t) - \delta_1 k_1(t) + \delta_2 k_2(t) - n_1 k_1(t) - c_1(t) \\ &= f_1(t) - c_1(t) - \delta_1 k_1(t) + \delta_2 k_2(t) - n_1 k_1(t), \end{aligned} \tag{5}$$

$$\begin{aligned} \dot{k}_2(t) &= f_2(t) + \delta_1 k_1(t) - \delta_2 k_2(t) - n_2 k_2(t) - c_2(t) \\ &= f_2(t) - c_2(t) + \delta_1 k_1(t) - \delta_2 k_2(t) - n_2 k_2(t). \end{aligned} \tag{6}$$

Based on Krasovskii et al. [13], the consumption function is defined as

$$\begin{aligned} C(t) &= (1 - s(t))F(t), \\ c(t)L(t) &= (1 - s(t))f(t)L(t). \end{aligned}$$

Both segments are multiplied by $\frac{1}{L(t)}$:

$$\begin{aligned} c(t) &= (1 - s(t))f(t), \\ c(t) &= f(t) - s(t)f(t), \\ s(t)f(t) &= f(t) - c(t), \end{aligned}$$

where $s(t)$ represents the investment of current savings invested in capital growth at time t with $s(t) < 1$. Thus, the Equations (5) – (6) can be written as

$$\begin{aligned} \dot{k}_1(t) &= f_1(t) - c_1(t) - \delta_1 k_1(t) + \delta_2 k_2(t) - n_1 k_1(t) \\ &= s_1(t) f_1(t) - \delta_1 k_1(t) + \delta_2 k_2(t) - n_1 k_1(t) \\ &= s_1(t)\mathcal{A}_1 k_1(t)^{\alpha_1} - \delta_1 k_1(t) + \delta_2 k_2(t) - n_1 k_1(t), \end{aligned}$$

$$\begin{aligned}
\dot{k}_2(t) &= f_2(t) - c_2(t) + \delta_1 k_1(t) - \delta_2 k_2(t) - n_2 k_2(t) \\
&= s_2(t) f_2(t) + \delta_1 k_1(t) - \delta_2 k_2(t) - n_2 k_2(t) \\
&= s_2(t) \mathcal{A}_2 k_2(t)^{\alpha_2} + \delta_1 k_1(t) - \delta_2 k_2(t) - n_2 k_2(t).
\end{aligned}$$

Assume that there exists $t \in (0, t_f]$ such that $k_1(t) \leq 0$ or $k_2(t) \leq 0$. First, suppose $k_1(t) \leq 0$ and $k_{1*} = \{t \in (0, t_f] \mid k_1(t) \leq 0\}$, then let $t^* = \inf k_{1*}$. It can be seen that $t^* \neq 0$, so there exists $k_1(t) > 0, \forall t \in [0, t^*)$ and

$$\begin{aligned}
\dot{k}_2(t) &= s_2(t) \mathcal{A}_2 k_2(t)^{\alpha_2} + \delta_1 k_1(t) - \delta_2 k_2(t) - n_2 k_2(t), \\
\dot{k}_2(t) &> -\delta_2 k_2(t) - n_2 k_2(t), \quad \forall t \in [0, t_f), \\
\dot{k}_2(t) + \delta_2 k_2(t) + n_2 k_2(t) &> 0.
\end{aligned}$$

Assume that there exists $t \in (0, t^*)$ such that $k_2(t) \leq 0$. Then suppose $k_{2*} = \{t \in (0, t^*) \mid k_2(t) \leq 0\}$ and $t^*_{k_2} = \inf k_{2*}$. It can be seen that $t^*_{k_2} \neq 0$. Then there exists $k_2(t) > 0, \forall t \in [0, t^*_{k_2})$ and

$$\begin{aligned}
\dot{k}_2(t) + \delta_2 k_2(t) + n_2 k_2(t) &> 0, \quad \forall t \in [0, t^*_{k_2}), \\
\dot{k}_2(t) + (\delta_2 + n_2) k_2(t) &> 0, \\
e^{(\delta_2 + n_2)t} \dot{k}_2(t) + e^{(\delta_2 + n_2)t} (\delta_2 + n_2) k_2(t) &> 0, \\
\frac{d}{dt} \left(e^{(\delta_2 + n_2)t} k_2(t) \right) &> 0, \\
\int_0^{t^*_{k_2}} \frac{d}{dt} \left(e^{(\delta_2 + n_2)t} k_2(t) \right) dt &> 0, \\
e^{(\delta_2 + n_2) t^*_{k_2}} k_2(t) - k_2(0) &> 0, \\
k_2(t^*_{k_2}) &> k_2(0) e^{-(\delta_2 + n_2) t^*_{k_2}}.
\end{aligned}$$

We obtain that $k_2(t^*_{k_2}) > k_2(0) e^{-(\delta_2 + n_2) t^*_{k_2}} > 0$. However, this contradicts the statement $k_2(t^*_{k_2}) \leq 0$. Therefore, it can be concluded that $k_2(t) > 0$ for any $t \in [0, t^*)$. It means that

$$\begin{aligned}
\dot{k}_1(t) &= s_1(t) \mathcal{A}_1 k_1(t)^{\alpha_1} - \delta_1 k_1(t) + \delta_2 k_2(t) - n_1 k_1(t), \\
\dot{k}_1(t) &> -\delta_1 k_1(t) - n_1 k_1(t), \qquad \forall t \in [0, t^*), \\
\dot{k}_1(t) + \delta_1 k_1(t) + n_1 k_1(t) &> 0, \\
\dot{k}_1(t) + (\delta_1 + n_1) k_1(t) &> 0, \\
e^{(\delta_1 + n_1)t} \dot{k}_1(t) + e^{(\delta_1 + n_1)t} (\delta_1 + n_1) k_1(t) &> 0, \\
\frac{d}{dt} \left(e^{(\delta_1 + n_1) t^*_{k_1}} k_1(t) \right) &> 0, \\
\int_0^{t^*} \frac{d}{dt} \left(e^{(\delta_1 + n_1)t} k_1(t) \right) dt &> 0, \\
e^{(\delta_1 + n_1) t^*} k_1(t) - k_1(0) &> 0, \\
k_1(t^*) &> k_1(0) e^{-(\delta_1 + n_1) t^*}
\end{aligned}$$

so that $k_1(t^*) > k_1(0) e^{-(\delta_1 + n_1) t^*} > 0$ holds. However, this contradicts the statement $k_1(t^*) \leq 0$. Therefore, it can be concluded that $k_1(t) > 0$ for any $t \in [0, t_f]$. Furthermore, in the same way, for $k_2(t)$, it is obtained that $k_2(t) > 0$ for every $t \geq t_0$.

It has been proved that the set $\Omega_{(k_1(t_0),k_2(t_0))}$ defined in Equation (4) is a positive invariant set. It means that if the initial condition in Equation (3) with control is positive, then the solution of the model is positive for any time t. However, it is not guaranteed that this model has a unique solution for a given initial condition.

Now, to guarantee that the solution of Equation (3) exists and is unique, we can use the concept of the Lipschitz condition in Equation (3) [17]. For that, we prove that Equation (3) satisfies the Lipschitz condition for $\alpha_1 = \alpha_2 = 1$ as given in the following theorem.

Theorem 3.2 *The mathematical model* (3) *that satisfies a given initial condition* $k_1(t_0), k_2(t_0) > 0$*, has a unique solution.*

Proof. Let $X = (k_1, k_2)$ and

$$\varphi(X) = \begin{bmatrix} \frac{dk_1}{dt} \\ \\ \frac{dk_2}{dt} \end{bmatrix}.$$

The Ramsey-Cass-Koopmans model with investment can be written as

$$\varphi(X) = \begin{bmatrix} \mathcal{A}_1 k_1 - \delta_1 k_1 + \delta_2 k_2 - n_1 k_1 - c_1 \\ \\ \mathcal{A}_2 k_2 + \delta_1 k_1 - \delta_2 k_2 - n_2 k_2 - c_2 \end{bmatrix}.$$

Note that for $X_a = (k_{1a}, k_{2a})$ and $X_b = (k_{1b}, k_{2b})$,

$$\varphi(X_a) - \varphi(X_b) = \begin{bmatrix} (\mathcal{A}_1 - \delta_1 - n_1)(k_{1a} - k_{1b}) + \delta_2(k_{2a} - k_{2b}) \\ \\ (\mathcal{A}_2 - \delta_2 - n_2)(k_{2a} - k_{2b}) + \delta_1(k_{1a} - k_{1b}) \end{bmatrix}.$$

Furthermore, by using the Euclidean norm of $\mathbb{R}^2$ and based on the triangle inequality, we obtain

$$\| \varphi(X_a) - \varphi(X_b) \| \leq \left\| \begin{bmatrix} (\mathcal{A}_1 - \delta_1 - n_1)(k_{1a} - k_{1b}) \\ \\ (\mathcal{A}_2 - \delta_2 - n_2)(k_{2a} - k_{2b}) \end{bmatrix} \right\| + \left\| \begin{bmatrix} \delta_2(k_{2a} - k_{2b}) \\ \\ \delta_1(k_{1a} - k_{1b}) \end{bmatrix} \right\|.$$

Since $\mathcal{A}_1, \mathcal{A}_2, \delta_1, \delta_2, n_1, n_2$ are constant, then there exists $M > 0$ such that

$$|\mathcal{A}_1 - \delta_1 - n_1|, |\mathcal{A}_2 - \delta_2 - n_2|, |\delta_1|, |\delta_2| \leq M.$$

Then

$$\begin{aligned} \| \varphi(X_a) - \varphi(X_b) \| &\leq M \left\| \begin{bmatrix} k_{1a} - k_{1b} \\ \\ k_{2a} - k_{2b} \end{bmatrix} \right\| + M \left\| \begin{bmatrix} k_{2a} - k_{2b} \\ \\ k_{1a} - k_{1b} \end{bmatrix} \right\| \\ &= M \left\| \begin{bmatrix} k_{1a} - k_{1b} \\ \\ k_{2a} - k_{2b} \end{bmatrix} \right\| + M \left\| \begin{bmatrix} k_{2a} - k_{2b} \\ \\ k_{1a} - k_{1b} \end{bmatrix} \right\| \\ &= M \| X_a - X_b \| + M \| X_a - X_b \| \\ &= 2M \| X_a - X_b \| \end{aligned}$$

so that $\varphi(X)$ is a Lipschitz function. It means that we obtain

$$X(t) = X(t_0) + \int_{t_0}^{t} \varphi(X)\ dt.$$

It is proved that X has a unique solution for the initial condition $k_i(t_0) > 0,\ i = 1, 2$.

4 The Existence of Optimal Control

In this paper, the objective function aims to maximize consumption utility through per capita consumption control. This approach is generally used when utility is directly based on consumption level. Consumption level is considered as the main factor affecting the consumption utility. The utility used is the logarithmic utility function, that is,

$$u(c(t)) = \ln c(t).$$

Then the objective function based on the Ramsey-Cass-Koopman model can be defined as follows:

$$J = \max_{c_1(t)\in U_1} \int_0^\infty \ln c_1(t) e^{-\rho t} dt + \max_{c_2(t)\in U_2} \int_0^\infty \ln c_2(t) e^{-\rho t} dt \tag{7}$$

for the discount factor ρ adjusts future consumption utility values according to individual time preferences, with constraints based on Equation (3). From Equation (7), we obtain that this statement is equivalent to

$$J = \max_{(c_1(t),c_2(t))\in U} \int_0^\infty (\ln c_1(t) + \ln c_2(t)) e^{-\rho t} dt \tag{8}$$

for $U = (U_1, U_2)$. Furthermore, by considering the equation

$$c(t) = (1 - s(t)) f(t) = (1 - s(t)) \mathcal{A} k(t),$$

we can represent the objective function (7) as

$$J = \max \int_0^\infty (\ln(1 - s_1(t)) \mathcal{A} k_1(t) + \ln(1 - s_2(t)) \mathcal{A} k_2(t)) e^{-\rho t} dt, \tag{9}$$

where $s_1(t)$ and $s_2(t)$ denote investment in the form of a portion of current output saved and invested in capital growth at time t with $s(t) < 1$ in the first and second industries, respectively.

We have shown in Section 3 that the Equation (3) has a unique and positive solution. Using the result from Fleming and Rishel [18]) we prove the existence of the optimal control by checking the following points.

1. The set S defined as

$$S = \{(s_1(t), s_2(t)) \mid 0 \le s_1(t), s_2(t) \le g, \forall t \in [0, t_f]\}$$

is a nonempty set. This can be seen from Theorems 3.1 and 3.2, where every control $s \in S$ has a unique and positive solution.

2. The set S is a closed convex set.

   Let $s_1(t),\ s_2(t) \in S$. It can be easily seen that $0 \leq s_1(t), s_2(t) \leq g$ for every $t \in [0, t_f]$; so, with every $\lambda \in [0,1]$, we obtain

   $$0 \leq \lambda s_1(t) + (1-\lambda)\, s_2(t) \leq g, t \in [0, t_f].$$

   Therefore

   $$\lambda s_1(t) + (1-\lambda)\, s_2(t) \in S.$$

   This shows that S is a convex set. Now, we show that S is a closed set. It is enough to show that for every convergent sequence $(s_n(t))_{n\in\mathbb{N}} = (s_{1n}(t), s_{2n}(t)) \subseteq S$, $\lim_{n\to\infty} s_n(t) \in S$. It means that for $s_n(t) \to (s_1(t), s_2(t))$ with $s_1(t) = \lim_{n\to\infty} s_{1n}(t)$ and $s_2 = \lim_{n\to\infty} s_{2n}(t)$, we obtain $(s_1(t), s_2(t)) \in S$. Now we define

   $$\| u - v \| := \sup\{|u\,(t) - v\,(t)\,|\ |\ t \in [0, t_f]\}\ .$$

   Then we know that $s_n(t)$ is a convergent sequence such that for every $\varepsilon > 0$, there exists $K(\varepsilon) \in \mathbb{N}$ that satisfies

   $$\| s_{1n}(t) - s_1(t) \| < \varepsilon$$

   and

   $$\| s_{2n}(t) - s_2(t) \| < \varepsilon$$

   for every $n \geq K(\varepsilon)$. Therefore, we obtain

   $$\begin{array}{cc} & \| s_{in}(t) - s_i(t) \| < \varepsilon \\ (\Rightarrow) & |s_{in}(t) - s_i(t)| < \sup\{|s_{in}(t) - s_i(t)|\ |\ t \in [0, t_f]\} < \varepsilon \\ (\Rightarrow) & -\varepsilon < s_i(t) - s_{in}(t) < \varepsilon \\ (\Rightarrow) & -\varepsilon \leq s_{in}(t) -\ \varepsilon < s_i(t) < \varepsilon + s_{in}(t) \leq \varepsilon + g \\ (\Rightarrow) & -\varepsilon < s_i(t) < g + \varepsilon \end{array}$$

   for $i = 1, 2$. Since it holds for every $\varepsilon > 0$, we obtain $0 \leq s_1(t), s_2(t) \leq g$ and $(s_1(t), s_2(t)) \in S$. Consequently, we show that S is a closed set. Furthermore, it can be proven that S is a closed convex set.

3. Note that the dynamic Equation (3) can be expressed as

   $$\begin{array}{rcl} \dot{k}_1(t) & = & s_1(t)\mathcal{A}_1 k_1(t) - \delta_1 k_1(t) + \delta_2 k_2(t) - n_1 k_1(t), \\ \dot{k}_2(t) & = & s_2(t)\mathcal{A}_2 k_2(t) + \delta_1 k_1(t) - \delta_2 k_2(t) - n_2 k_2(t). \end{array} \tag{10}$$

   It can be seen that the right-hand side of equation (10) is a linear function in the state and control variables. Then we know that $k_i(t)$ is continuous in the interval $[0, t_f]$ and $s_i(t)$ is a bounded function with $0 \leq s_i(t) \leq g$ for $i = 1, 2$. So we prove that the Equation (10) is bounded.

4. Let

   $$U(s_1(t), s_2(t)) = e^{-\rho t} \ln((1 - s_1(t))\mathcal{A}k_1(t)) + \ln((1 - s_2(t))\mathcal{A}k_2(t))$$

   and we know that $s_1(t), s_2(t), k_1(t)$ and $k_2(t)$ are bounded functions on $[0, t_f]$. It means that $U(s_1(t), s_2(t))$ is a bounded function. Suppose that $U_1(s_1(t) = \ln((1 - s_1(t))\mathcal{A}k_1(t))$ and $U_2(s_2(t) = \ln((1 - s_2(t))\mathcal{A}k_2(t))$. Then we obtain

   $$\frac{\partial^2 U_1}{\partial s_1^2} = -\frac{1}{(1 - s_1(t))^2}\mathcal{A}k_1(t) < 0$$

and
$$\frac{\partial^2 U_2}{d\partial s_2^2} = -\frac{1}{(1-s_2(t))^2}\mathcal{A}k_2(t) < 0.$$
It means that U_1 and U_2 are concave functions. Thus, we prove that U is a concave function.

5 Conclusion

The dynamic model of industrial economic growth with investment, as developed in this paper, extends the Ramsey-Cass-Koopmans model by incorporating a strategy focused on maximizing consumption utility. Specifically, the model is adapted to two industries, with investment flowing from the first industry to the second to achieve optimal consumption utility across both sectors. Control variables are introduced in the form of per capita consumption in the first industry (c_1) and the second industry (c_2) to maximize utility in both industries. In this context, control through per capita consumption can be interpreted as control through savings, represented by s_1 and s_2, which denote savings in the first and second industries, respectively. Analytically, the model is valid and has a unique solution, as demonstrated through the concept of positive invariant sets and the Lipschitz continuity of the model. The positivity of the resulting solution ensures that the capital stocks in both industries, K_1 and K_2, remain non-negative. This paper also analyzes the existence of optimal control, establishing that any introduced control leads to a positive solution and confirming the existence of optimal control for maximizing consumption utility.

Acknowledgements

The authors gratefully acknowledge the Institut Teknologi Sepuluh Nopember for financial support to this work, within the 2024 Publication Writing and IPR Incentive Program (PPHKI) project programme.

References

[1] D. Aemoglu. *Introduction to Modern Economic Growth*. Princeton University Press, 2008.

[2] N. Prawoto and A. T. Basuki. The influence of macroeconomic variables, processing industry, and education services on economic growth in Indonesia. *Entrepreneurship and Sustainability Issues* **8** (1) (2020) 1–20.

[3] D. Idayani and S. Subchan. Optimal control of multi-supplier inventory management with lead time. *International Journal of Computing Science and Applied Mathematics* **6** (1) (2020) 23–32.

[4] I. Fitria, S. Subchan, D. A. Maulina, and A. A. Hidayatullah. Analysis and existence of rumor spreading with campaign and punishment control. *Nonlinear Dynamics and Systems Theory* **4** (1) (2024) 380–391.

[5] M. Bohner, C. Duque and H. Leiva. Controllability of dynamic equations with memory. *Nonlinear Dynamics and Systems Theory* **22** (5) (2022) 489–502.

[6] F. P. Ramsey. A mathematical theory of saving. *The Economic Journal* **38** (152) (1928) 543–559.

[7] T. C. Koopmans. On the concept of optimal economic growth. *The Economic Approach to Development Planning* (1965) 225–287.

[8] D. Cass. Optimum growth in an aggregative model of capital accumulation. *The Review of Economic Studies* **32** (3) (1965) 233–240.

[9] E. Accinelli and J. G. Brida. The Ramsey model with logistic population growth. Available at SSRN 881518 (2006).

[10] G. Sorger. On the long-run distribution of capital in the Ramsey model. *Journal of Economic Theory* **105** (1) (2002) 226–243.

[11] R. Boucekkine, C. Camacho, and B. Zou. Bridging the gap between growth theory and the new economic geography: The spatial Ramsey model. *Macroeconomic Dynamics* **13** (1) (2009) 20–45.

[12] M. N. Dudin, et al. Methodology making management decisions based on a modified Ramsey model. *Asian Social Science* **10** (17) (2014) 292–301.

[13] A. A. Krasovskii, P. D. Lebedev, and A. M. Tarasyev. Some facts about the Ramsey model. *Proceedings of the Steklov Institute of Mathematics* **299** (1) (2017) 123–131.

[14] O. Bundău and A. Juratoni. Optimal conditions for a control problem associated to an economical model. *ITM Web of Conferences* **29** (2019) 01012. EDP Sciences.

[15] V. Kajanovičová, B. Novotnỳ, and M. Pospíšil. Ramsey model with non-constant population growth. *Mathematical Social Sciences* **104** (2020) 40–46.

[16] L. Frerick, G. Müller-Fürstenberger, M. Schmidt, and M. Spaeth. Complementary modelling of a Ramsey-type equilibrium problem with heterogeneous agents. *Computational Economics* **60** (2022) 1135–1154.

[17] X. Chen and B. Liu. Existence and uniqueness theorem for uncertain differential equations. *Fuzzy Optimization and Decision Making* **9** (2010) 69–81.

[18] W. H. Fleming and R. W. Rishel. *Deterministic and Stochastic Optimal Control*. Springer Science & Business Media, 2012.

Nonlinear Dynamics and Systems Theory, **25** (3) (2025) 266–287

Modified Parameter of the Dai–Liao Conjugacy Condition of the Conjugate Gradient Method with Some Applications

A. Jaradat [1*], S. Masmali [2], A. Alhawarat [1], R. Sabra [2], S. Ismail [3] and A. S. Al-Jawarneh [1]

[1] *Department of Mathematics, Faculty of Arts and Science, Amman Arab University, Amman, 11953, Jordan.*
[2] *Department of Mathematics, College of Science, Jazan University, Jazan, Saudi Arabia.*
[3] *Financial Mathematics Program, Faculty of Science and Technology, Universiti Sains Islam Malaysia, Bandar Baru Nilai, 71800, Negeri Sembilan, Malaysia.*
[3] *Risk and Analytics Research Group, Faculty of Science and Technology, Universiti Sains Islam Malaysia.*

Received: December 25, 2024; Revised: June 4, 2025

Abstract: This study introduces a novel modification of the conjugate gradient (CG) method by refining the Dai–Liao conjugacy parameter and incorporating a restart property. The proposed method, which is established in the Hestenes–Stiefel framework, is designed to ensure global convergence and satisfies the sufficient descent condition for both convex and non-convex functions. Utilizing the Lipschitz constant as a foundation, the method's efficiency and robustness were benchmarked against CG Descent across over 200 functions from the CUTEst library. Numerical experiments revealed superior performance in terms of CPU time, iterations, gradient evaluations, and function evaluations. Additionally, practical applications in heat conduction and image restoration demonstrate the method's versatility and effectiveness.

Keywords: *conjugate gradient; inexact line search; conjugacy condition; global convergence; CUTEst library.*

Mathematics Subject Classification (2020): 49M37, 65K05, 90C30.

* Corresponding author: `mailto:a.jaradat@aau.edu.jo`

1 Introduction

Conjugate gradient (CG) methods have been widely used for solving nonlinear unconstrained optimization problems due to their low memory requirements for implementation. Moreover, CG methods have been used in many applications such as regression analysis, image restoration, electrical circuits, and many others.

The CG method is used to determine optimal solutions for the following optimization problem:

$$\min f(x), \quad x \in \mathbb{R}^n,$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a continuously differentiable function, and its gradient $\nabla f(x_k) = g_k = g(x_k)$ should exist. From the starting point (arbitrary or standard) $x_1 \in \mathbb{R}^n$, the CG method generates a sequence of vectors x_k by the iterative rule

$$x_{k+1} = x_k + \alpha_k d_k, \quad k = 1, 2, \dots, \tag{1}$$

in which x_k represents the present iteration and $\alpha_k > 0$ represents a step size obtained from the exact line search or an inexact line search. The search direction d_k of the CG method is defined by

$$d_k = \begin{cases} -g_k & \text{if } k = 1, \\ -g_k + \beta_k d_{k-1} & \text{if } k \geq 2, \end{cases} \tag{2}$$

where β_k is the update parameter. The following exact line search can be utilized to obtain the step size α_k:

$$f(x_k + \alpha_k d_k) = \min_{\alpha} f\ (x_k + \alpha d_k). \tag{3}$$

However, Eq.(3) is computationally expensive because it requires unidimensional optimization to achieve the step size and many iterations to reach convergence. To avoid this problem, the inexact line search is a dominant approach in computing the step size. The most popular inexact line search is the strong Wolfe–Powell (SWP) line search [1,2], which is defined as

$$f(x_k + \alpha_k d_k) \leq f(x_k) + \delta \alpha_k g_k^T d_k, \tag{4}$$

$$|g(x_k + \alpha_k d_k)^T d_k| \leq \sigma |g_k^T d_k| \tag{5}$$

so that $0 < \delta < \sigma < 1$.

A version of the Wolfe–Powell line search is the weak Wolfe–Powell (WWP) line search, which is defined by (4) and

$$g(x_k + \alpha_k d_k)^T d_k \geq \sigma g_k^T d_k.$$

The most famous classical formulae of the CG methods are the Hestenes–Stiefel (HS) [3], Fletcher–Reeves (FR) [4], and Polak–Ribiere–Polyak (PRP) [5] methods, which are defined by the following update parameters, respectively:

$$\beta_k^{HS} = \frac{g_k^T y_{k-1}}{d_{k-1}^T y_{k-1}}, \quad \beta_k^{FR} = \frac{\|g_k\|^2}{\|g_{k-1}\|^2}, \quad \text{and}$$

$$\beta_k^{PRP} = \frac{g_k^T y_{k-1}}{\|g_{k-1}\|^2}, \quad \text{where} \quad y_{k-1} = g_k - g_{k-1}.$$

Powell in [6] provided a counterexample showing that there exists a non-convex function for which the PRP and HS methods fail to satisfy the convergence properties even if the exact line search is employed. Powell recommended the use of nonnegative paremeters β_k^{HS} and β_k^{PRP} to achieve the convergence properties of the CG method. Gilbert and Nocedal [7] proved that the nonnegative PRP or HS method defined by $\beta_k = \max\{\beta_k^{PRP}, 0\}$, is globally convergent with arbitrary line searches.

The descent condition (downhill condition) plays a crucial role in the convergence of the CG method and its robustness, and it is defined by

$$g_k^T d_k < 0. \tag{6}$$

Al-Baali [8] proposed another version of the downhill condition called the sufficient descent condition, which also plays a significant role in the convergence of the CG method. Al-Baali proposed the condition

$$g_k^T d_k \leq -c\|g_k\|^2 \quad \forall k \in \mathbb{N} \tag{7}$$

to establish the global convergence properties of β_k^{FR}. More precisely, if there exists a constant $c > 0$ satisfying (7), then the search direction d_k guarantees the sufficient descent condition.

Based on the quasi-Newton method, the Broyden–Fletcher–Goldfarb–Shanno (BFGS) method and the limited-memory BFGS (LBFGS) method, and using Eq.(2), Dai and Liao [9] proposed the conjugacy condition

$$d_k^T y_{k-1} = -t g_k^T s_{k-1} \tag{8}$$

such that $s_{k-1} = x_k - x_{k-1}$ and $t \geq 0$. In the case of $t = 0$, Eq.(8) is considered as the classical conjugacy condition. Using Eqs. (2) and (8), Dai and Liao [9] proposed the following CG formula:

$$\beta_k^{DL} = \frac{g_k^T y_{k-1}}{d_{k-1}^T y_{k-1}} - t\frac{g_k^T s_{k-1}}{d_{k-1}^T y_{k-1}}. \tag{9}$$

However, β_k^{DL} cannot satisfy the descent condition and convergence properties similar to β_k^{PRP} and β_k^{HS} because β_k^{DL} is not nonnegative in general. Thus, Dai and Liao [9] replaced the formula (9) by

$$\beta_k^{DL+} = \max\{\beta_k^{HS}, 0\} - t\frac{g_k^T s_{k-1}}{d_{k-1}^T y_{k-1}}. \tag{10}$$

However, β_k^{DL+} cannot satisfy the descent property in some cases. Therefore, Dai and Liao [9] restarted (10) using a negative gradient (steepest descent) when β_k^{DL+} fails to satisfy inequality (7). Another method for determining the optimal parameter t was proposed by Babaie-Kafaki and Ghanbari [10,11], where they rewrote the search direction (Eq.(2)) with β_k^{DL}, and based on Perry [12], as follows: $d_{k+1} = -Q_{k+1}g_{k+1}$, where $Q_{k+1} = I - \frac{s_k y_k^T}{s_k^T y_k} + t\frac{s_k s_k^T}{s_k^T y_k}$. Babaie-Kafaki and Ghanbari [10] proposed the following adaptive choices for t:

$$t = \frac{s_k y_k^T}{\|s_k\|^2} + \frac{\|y_k\|}{\|s_k\|} \quad \text{and} \quad t = \frac{\|y_k\|}{\|s_k\|}.$$

Andrei in [13] proposed a CG method with the parameter

$$\beta_k^{DL*} = \max\left\{\frac{y_k^T g_k}{y_k^T s_k}, 0\right\} - t_k^* \frac{s_k^T g_{k+1}}{y_k^T s_k},$$

where $t_k^* = \frac{y_k^T s_k}{\|s_k\|^2}$. Hager and Zhang [14, 15] presented a modified CG parameter that satisfies the descent property for any inexact line search with $g_k^T d_k \leq -(7/8)\|g_k\|^2$. This new version of the CG method is globally convergent whenever the line search satisfies the WWP line search. This formula is expressed by

$$\beta_k^{HZ} = \max\{\beta_k^N, \eta_k\},$$

where $\beta_k^N = \frac{1}{d_k^T y_k}\left(y_k - 2d_k \frac{\|y_k\|^2}{d_k^T y_k}\right)^T g_k$, $\eta_k = -\frac{1}{\|d_k\| \min\{\eta, \|g_k\|\}}$, and $\eta > 0$ is a constant.

Note that, if $t = 2\frac{\|y_k\|^2}{s_k^T y_k}$, then $\beta_k^N = \beta_k^{DL}$. Zhang *et al.* [16] proposed a new parameter for Eq.(9) as follows:

$$t = \frac{\|y_k\|^2}{s_k^T y_k} - \frac{1}{4}\frac{s_k^T y_k}{\|s_k\|^2}.$$

Yao *et al.* [2] proposed three terms of CG with a new choice of t as follows:

$$d_{k+1} = -g_{k+1} + \left(\frac{g_k^T y_k - t_k g_{k+1}^T s_k}{y_k^T d_k}\right) d_k + \frac{g_{k+1}^T d_k}{y_k^T d_k} y_k.$$

Based on the SWP line search, Yao *et al.* [2] selected t_k to satisfy the descent condition

$$t_k > \frac{\|y_k\|^2}{y_k^T s_k}.$$

Yao *et al.* [2] also proposed a theorem stating that if t_k is close to $\frac{\|y_k\|^2}{y_k^T s_k}$, then the search direction results in a zigzag search path. Therefore, they selected the following choice for t_k:

$$t_k = 1 + 2\frac{\|y_k\|^2}{y_k^T s_k}.$$

Al-Baali *et al.* [17] proposed a new CG version called the G3TCG that offers many selections of CG parameters. They found that the G3TCG method is more efficient than β_k^{HZ} in some cases and competitive in some other cases.

In this research, we propose a new CG iterative formula based on a modified parameter of the Dai–Liao conjugacy condition of the CG method with the restart property. The convergence of the proposed modified CG method is analyzed under standard assumptions. Numerical experiments are performed to illustrate the superiority of the proposed method.

The highlighted results are achieved in the subsequent sections organized in the following manner. A novel CG formula is proposed in Section 2, as well as underlying motivation. The convergence analysis of the modified CG method is presented in Section 3. Section 4 includes the results of numerical experiments and their discussion.

2 Proposed CG Formula and Its Motivation

The CG method with β_k^{DL} cannot satisfy the descent condition, but β_k^{DL} inherits the conjugacy condition. To improve the properties of β_k^{DL}, we used β_k^{AZPRP} as presented by Alhawarat et al. [23] to propose a new nonnegative CG method that can satisfy the sufficient descent condition and global convergence properties with the SWP line search as follows:

$$\beta_k^{AZPRP} = \begin{cases} \frac{\|g_k\|^2 - \mu_k \left|g_k^T g_{k-1}\right|}{\|g_{k-1}\|^2}, & \|g_k\|^2 > \mu_k \left|g_k^T g_{k-1}\right|, \\ 0, & \text{otherwise.} \end{cases}$$

The proposed CG update parameter is a modification of β_k^{DL} and β_k^{HS}, with the restart criterion depending on the Lipschitz constant used in the study conducted by Alhawarat *et al.* [23]. The modified formula is expressed as

$$\beta_k^{AZHS} = \begin{cases} \frac{\|g_k\|^2 - \mu_k \left|g_k^T g_{k-1}\right|}{d_{k-1}^T y_{k-1}} - \frac{1}{\alpha_k}\mu_k \frac{g_k^T s_{k-1}}{d_{k-1}^T y_{k-1}}, & \|g_k\|^2 > \mu_k \left|g_k^T g_{k-1}\right|, \\ -\frac{1}{\alpha_k}\mu_k \frac{g_k^T s_{k-1}}{d_{k-1}^T y_{k-1}}, & \text{otherwise,} \end{cases} \tag{11}$$

where $\|\cdot\|$ represents the Euclidean norm and μ_k is defined as follows:

$$\mu_k = \frac{\|s_{k-1}\|}{\|y_{k-1}\|}.$$

In the first case of the equality (11), we can note that

$$\beta_k^{AZHS} \leq \frac{\|g_k\|^2}{d_{k-1}^T y_{k-1}} - \frac{1}{\alpha_k}\mu_k \frac{g_k^T s_{k-1}}{d_{k-1}^T y_{k-1}}. \tag{12}$$

It is worth noting that the formula (11) inherits the advantages of β_k^{DL}, β_k^{HS}, and β_k^{AZPRP}. Moreover, as we will see in the next sections, the new formula satisfies the descent condition and the global convergence properties. The usage of the proposed parameter β_k^{AZHS} in (11) leads to the novel CG method described in Algorithm 2.1.

Algorithm 2.1 CG method based on β_k^{AZHS}.

Step 1 Set a starting point x_1. The initial point can be arbitrary or standard for scientific functions. The initial search direction is the negative gradient, i.e., $d_1 = -g_1$. Let $k \leftarrow 1$.

Step 2 If the stopping condition is satisfied, then stop.

Step 3 Compute the search direction d_k based on Eq.(2) using Eq.(11).

Step 4 Compute the step size α_k using Eqs.(4) and (5).

Step 5 Update x_{k+1} based on Eq.(1).

Step 6 Set $k \leftarrow k + 1$ and go to Step 2.

3 Convergence Analysis of β_k^{AZHS}

To perform the convergence analysis of the modified CG method, we consider the following assumptions.

Assumption 1

A. The level set $\varPhi = \{x | f(x) \leq f(x_1)\}$ is bounded. In other words, a positive constant B exists so that

$$\|x\| \leq B, \quad \forall x \in \varPhi.$$

B. In some neighborhood P of $\varPhi$, f is continuously differentiable, and its gradient is Lipschitz continuous. In other words, $\forall x, y \in P, \exists L > 0$ so that

$$\|g(x) - g(y)\| \leq L \|x - y\| .$$

This assumption implies that there exists a positive constant $\hat{\gamma}$ such that

$$\|g(x)\| \leq \widehat{\gamma}, \quad \forall x \in P.$$

Theorem 3.1 *Let the sequences* $\{g_k\}$ *and* $\{d_k\}$ *be obtained using Eqs.*(1) *and* (2), *and* β_k^{AZHS}, *where* α_k *is computed using the SWP line search in Eqs.*(4) *and* (5). *If* $\sigma \in (0, 0.5)$, *then the descent condition provided in* (7) *holds.*

Proof. The proof is carried out for two cases.

Case 1: $\|g_k\|^2 > \mu_k \left|g_k^T g_{k-1}\right|$.

This assumption implies

$$\beta_k^{AZHS} = \frac{\|g_k\|^2 - \mu_k \left|g_k^T g_{k-1}\right|}{d_{k-1}^T y_{k-1}} - \frac{1}{\alpha_k} \mu_k \frac{g_k^T s_{k-1}}{d_{k-1}^T y_{k-1}}.$$

Multiplying (2) by g_k^T, we can conclude that

$$\begin{aligned} g_k^T d_k &= g_k^T(-g_k + \beta_k d_{k-1}) = -\|g_k\|^2 + \beta_k g_k^T d_{k-1} \\ &\leq -\|g_k\|^2 + \frac{\|g_k\|^2}{\left|d_{k-1}^T y_{k-1}\right|} \left|g_k^T d_{k-1}\right| - \mu_k \frac{\left\|g_k^T d_{k-1}\right\|^2}{d_{k-1}^T y_{k-1}}. \end{aligned}$$

The usage of the SWP line search leads to the inequality

$$\frac{\left|g_k^T d_{k-1}\right|}{\left|d_{k-1}^T y_{k-1}\right|} \leq \frac{\sigma}{1-\sigma}.$$

Thus,

$$g_k^T d_k \leq -\|g_k\|^2 + \frac{\sigma \|g_k\|^2}{(1-\sigma)} = -\|g_k\|^2 \left(1 - \frac{\sigma}{1-\sigma}\right) = -c\|g_k\|^2.$$

Case 2: $\|g_k\|^2 \leq \mu_k \left|g_k^T g_{k-1}\right|$.

This assumption implies

$$\beta_k^{AZHS} = -\frac{1}{\alpha_k} \mu_k \frac{g_k^T s_{k-1}}{d_{k-1}^T y_{k-1}}$$

and further

$$\begin{aligned} g_k^T d_k &= g_k^T(-g_k + \beta_k d_{k-1}) = -\|g_k\|^2 + \beta_k g_k^T d_{k-1} \\ &\leq -\|g_k\|^2 + \left(-\frac{\mu_k}{\alpha_{k-1}} \frac{g_k^T s_{k-1}}{d_{k-1}^T y_{k-1}}\right) g_k^T d_{k-1} \\ &= -\|g_k\|^2 - \mu_k \frac{\left\|g_k^T d_{k-1}\right\|^2}{d_{k-1}^T y_{k-1}}. \end{aligned}$$

Since the SWP line search is used, it follows that $d_{k-1}^T y_{k-1} > 0$, and further

$$g_k^T d_k \leq -c\|g_k\|^2,$$

which completes the proof. □

Lemma 3.1 shows that if $L > 1$, then equation (13) holds. Note that if $L \ll 1$, then $\|g_k\|^2 > \mu_k \left|g_k^T g_{k-1}\right|$ can not be satisfied.

Lemma 3.1 *If* $\|g_k\|^2 > \mu_k \left|g_k^T g_{k-1}\right|$ *and* $L > 1$*, then*

$$\|g_k\|^2 - \frac{1}{L}\left|g_k^T g_{k-1}\right| \leq L \left|\|g_k\|^2 - \left|g_k^T g_{k-1}\right|\right|. \tag{13}$$

***Proof*.** The proof is performed using contradiction. Suppose

$$\|g_k\|^2 - \frac{1}{L}\left|g_k^T g_{k-1}\right| > L \left|\|g_k\|^2 - \left|g_k^T g_{k-1}\right|\right|,$$

and divide both sides by L:

$$\frac{\|g_k\|^2}{L} - \frac{1}{L^2}\left|g_k^T g_{k-1}\right| > \left|\|g_k\|^2 - \left|g_k^T g_{k-1}\right|\right|. \tag{14}$$

Using Assumption 1, the following relationship is derived:

$$\|g_k\|^2 > \mu_k \left|g_k^T g_{k-1}\right| > \frac{1}{L}\left|g_k^T g_{k-1}\right|.$$

If $L > 1$, we conclude that inequality (14) is not true, which results in a contradiction. Thus, inequality (13) holds. □

The following Lemma 3.2 indicates the step length always has a lower bound.

Lemma 3.2 [25]. *Suppose that the objective function satisfies Assumption 1. If the step length* α_k *fulfills the SWP line search conditions* (4) *and* (5)*, then*

$$\alpha_k \geq \frac{(1-\sigma)\left|g_k^T d_k\right|}{L\|d_k\|^2}.$$

Lemma 3.3 *Let Assumption 1 hold. Consider any form of Eqs.*(1) *and* (2) *with the step size* α_k *satisfying the SWP line search, where the search direction* d_k *is descent. The following inequality is obtained:*

$$\sum_{k=0}^{\infty} \frac{(g_k^T d_k)^2}{\|d_k\|^2} < \infty. \tag{15}$$

The condition presented in inequality (15) is called the Zoutendijk condition [25] and plays an important role in proving the convergence properties of the CG method. We use the contradiction technique with (15) to prove $\liminf_{k\to\infty} \|g_k\| = 0$.

Moreover, (15) holds for the exact and SWP line searches. By substituting (7) into (15), we obtain

$$\sum_{k=0}^{\infty} \frac{\|g_k\|^4}{\|d_k\|^2} < \infty. \tag{16}$$

Lemma 3.4 *If $\|g_k\|^2 > \mu_k \left|g_k^T g_{k-1}\right|$ is satisfied, then $\mu_k = \frac{\|s_{k-1}\|}{\|y_{k-1}\|}$ is bounded above and below.*

Proof. Since $\|g_k\|^2 > \mu_k \left|g_k^T g_{k-1}\right| > \frac{1}{L} \left|g_k^T g_{k-1}\right|$, based on Assumption 1, it follows that $0 < \mu_k \le E$, where E denotes a positive constant. Moreover, if $y_{k+1} = 0$, this means $x_{k+1} = x_k$ and it is known that $x_{k+1} = x_k + \alpha_k d_k$. Thus, $\alpha_k d_k = 0$. However, by Lemma 3.2, we conclude that $\alpha_k > 0$. This means that $d_k = 0$. The usage of Theorem 3.1 and Lemma 3.3 leads to a contradiction. □

Dai *et al.* [26] presented the following Theorem 3.2, which is also useful for proving the global convergence properties of CG methods.

Theorem 3.2 *Suppose that Assumption 1 holds. Consider any CG method in the form of Eqs.*(1) *and* (2)*, where d_k is a descent direction and α_k is obtained using the SWP line search. If*

$$\sum_{k\ge 1}^{\infty} \frac{1}{\|d_k\|^2} = \infty,$$

then

$$\liminf_{k\to\infty} \|g_k\| = 0.$$

Global convergence properties for the convex functions

In the following theorem, if $f(x)$ is a uniformly convex function, then the CG method satisfies β_k^{AZHS} strong global convergence properties.

Theorem 3.3 *Suppose that Assumption 1 holds. Consider the CG method in the form of Eqs.*(1) *and* (2) *with β_k^{AZHS}, $L > 1$, and d_k as a descent direction, where α_k is obtained using the SWP line search. If $f(x)$ is a uniformly convex function, then $\liminf_{k\to\infty} \|g_k\| = 0$.*

Proof. Since the function $f(x)$ is uniformly convex, there exists a positive constant ϖ satisfying

$$\varpi \|x - y\|^2 \le (\nabla f(x) - \nabla f(y))^T (x - y)$$

for all $x, y \in \mathcal{P}$. Thus,

$$d_{k-1} y_{k-1} \ge \varpi \alpha_{k-1} \|d_{k-1}\|^2 \tag{17}$$

and

$$\begin{aligned} \beta_k^{AZHS} &= \frac{\|g_k\|^2 - \mu_k \left|g_k^T g_{k-1}\right|}{d_{k-1}^T y_{k-1}} - \frac{\mu_k}{\alpha_{k-1}} \frac{g_k^T s_{k-1}}{d_{k-1}^T y_{k-1}} \\ &\le \frac{\|g_k\|^2 - \frac{1}{L} \left|g_k^T g_{k-1}\right|}{d_{k-1}^T y_{k-1}} - \frac{\mu_k}{\alpha_{k-1}} \frac{g_k^T s_{k-1}}{d_{k-1}^T y_{k-1}}. \end{aligned}$$

An application of inequalities (17) and (13) gives

$$\begin{aligned}\beta_k^{AZHS} &\leq \frac{L\left\|g_k\right\|\left(\left\|g_k-g_{k-1}\right\|\right)}{\varpi\alpha_{k-1}\left\|d_{k-1}\right\|^2}+E\frac{\left|g_k^T s_{k-1}\right|}{\varpi\alpha_{k-1}^2\left\|d_{k-1}\right\|^2} \\ &\leq \frac{L\left\|g_k\right\|\left\|g_k-g_{k-1}\right\|}{\varpi\alpha_{k-1}\left\|d_{k-1}\right\|^2}+E\frac{\left\|g_k\right\|\left\|s_{k-1}\right\|}{\varpi\alpha_{k-1}^2\left\|d_{k-1}\right\|^2}.\end{aligned}$$

Applying Assumption 1, we obtain

$$\begin{aligned}\beta_k^{AZHS} &\leq \frac{L^2\left\|g_k\right\|\alpha_{k-1}\left\|d_{k-1}\right\|}{\varpi\alpha_{k-1}\left\|d_{k-1}\right\|^2}+E\frac{\left\|g_k\right\|\left\|d_{k-1}\right\|}{\varpi\alpha_{k-1}\left\|d_{k-1}\right\|^2} \\ &\leq \frac{L^2\left\|g_k\right\|}{\varpi\left\|d_{k-1}\right\|}+E\frac{\left\|g_k\right\|}{\varpi\alpha_{k-1}\left\|d_{k-1}\right\|}.\end{aligned}$$

Based on Eq.(2), it can be obtained that

$$\begin{aligned}\left\|d_k\right\| &\leq \left\|g_k\right\|+\left|\beta_k\right|\left\|d_{k-1}\right\| \\ &\leq \left\|g_k\right\|+\frac{\left\|g_k\right\|}{\left\|d_{k-1}\right\|}\left(\frac{L^2}{\varpi}+\frac{E}{\varpi\alpha_{k-1}}\right)\left\|d_{k-1}\right\| \\ &\leq \hat{\gamma}\left(1+\left(\frac{L^2}{\varpi}+\frac{E}{\varpi\alpha_{k-1}}\right)\right).\end{aligned}$$

Thus, Theorem 3.2 leads to the conclusion

$$\liminf_{k\to\infty}\left\|g_k\right\|=0$$

and completes the proof. □

Global convergence for β_k^{AZHS} with the SWP line search for general functions

Using Property(*) and some lemmas, Gilbert and Nocedal [7] proved the global convergence of nonnegative PRP and HS methods. Because our modification is nonnegative and satisfies Property(*), by using the other lemmas presented below, we perform our proof in the same way as in [7]. This property is defined as follows.

Property(*)

Consider any CG method in the form of Eqs.(1) and (2). Assume

$$0<\gamma\leq\left\|g_k\right\|\leq\hat{\gamma} \tag{18}$$

for all $k\geq 1$. The CG method then inherits Property(*) if for $\forall k$, there exist constants $b>1$ and $\lambda>0$ such that $|\beta_k|\leq b$ and $\|s_k\|\leq\lambda$, which implies that $|\beta_k|\leq\frac{1}{2b}$.

Lemma 3.5 shows that β_k^{AZHS} satisfies Property(*).

Lemma 3.5 *Consider a CG method in the form of Eqs.*(1) *and* (2) *using* β_k^{AZHS} *with* $L>1$. *Lemma 3.1 holds true, then* β_k^{AZHS} *satisfies Property(*).*

Proof. Let $b=\frac{2L\alpha_{k-1}\hat{\gamma}^2+B\hat{\gamma}}{\alpha_{k-1}L(1-\sigma)c\gamma^2}\geq 1$, and let $\lambda\leq\frac{(1-\sigma)c\gamma^2}{2(L^2+\frac{E}{\alpha_{k-1}})\hat{\gamma}b}$. Then the following inequality holds:

$$\beta_k^{AZHS}\leq\frac{\left\|g_k\right\|^2-\mu_k\left|g_k^T g_{k-1}\right|}{d_{k-1}^T y_{k-1}}-\frac{\mu_k}{\alpha_{k-1}}\frac{g_k^T s_{k-1}}{d_{k-1}^T y_{k-1}}.$$

Inequalities (13) and (18) are a basis for the inequalities

$$\begin{aligned}\beta_k^{AZHS} &\leq \frac{\|g_k\|^2 + |g_k^T g_{k-1}|}{d_{k-1}^T y_{k-1}} + \frac{E}{\alpha_{k-1}} \frac{|g_k^T s_{k-1}|}{d_{k-1}^T y_{k-1}} \\ &\leq \frac{2\hat{\gamma}^2}{(1-\sigma)c\gamma^2} + \frac{EB\hat{\gamma}}{\alpha_{k-1}L(1-\sigma)c\gamma^2} = \frac{2L\alpha_{k-1}\hat{\gamma}^2 + EB\hat{\gamma}}{\alpha_{k-1}(1-\sigma)c\gamma^2} \\ &= b > 1.\end{aligned}$$

Further, $\|s_k\| \leq \lambda$ gives

$$\begin{aligned}\beta_k^{AZHS} &\leq \frac{\|g_k\|^2 - \mu_k |g_k^T g_{k-1}|}{d_{k-1}^T y_{k-1}} - \frac{\mu_k}{\alpha_{k-1}} \frac{g_k^T s_{k-1}}{d_{k-1}^T y_{k-1}} \\ &\leq \frac{L\|g_k\| \|g_k - g_{k-1}\|}{d_{k-1}^T y_{k-1}} + \frac{E}{\alpha_{k-1}} \frac{\|g_k\| \|s_{k-1}\|}{d_{k-1}^T y_{k-1}} \\ &\leq \frac{L^2 \|g_k\| \|s_{k-1}\|}{d_{k-1}^T y_{k-1}} + \frac{E}{\alpha_{k-1}} \frac{\|g_k\| \|s_{k-1}\|}{d_{k-1}^T y_{k-1}} \\ &\leq \frac{(L^2 + \frac{E}{\alpha_{k-1}}) \|g_k\| \|s_{k-1}\|}{d_{k-1}^T y_{k-1}} \\ &\leq \frac{(L^2 + \frac{E}{\alpha_{k-1}})\hat{\gamma}\lambda}{(1-\sigma)c\gamma^2} = \frac{1}{2b}.\end{aligned}$$

Thus, the proof is complete. $\square$

Lemma 3.6 and Lemma 3.7 are similar to Lemma 4.1 and Lemma 4.2 presented by Gilbert and Nocedal in [7].

Lemma 3.6 *Assume that Assumption 1 holds and the sequences $\{g_k\}$ and $\{d_k\}$ are generated using Algorithm 1, where the step size α_k is computed via the SWP line search so that the sufficient descent condition holds. If $\beta_k \geq 0$, there exists a constant $\gamma > 0$ such that $\|g_k\| > \gamma$ for all $k \geq 1$. Then $d_k \neq 0$ and*

$$\sum_{k=0}^{\infty} \|u_{k+1} - u_k\|^2 < \infty,$$

where $u_k = \frac{d_k}{\|d_k\|}$.

Proof. The assumption $d_k = 0$, based on the sufficient descent condition, leads to $g_k = 0$. So, $d_k \neq 0$ as well as

$$\|g_k\| \geq \gamma, \quad \text{where } \gamma > 0. \tag{19}$$

Eq.(11) can be divided into two parts as follows:

$$\beta_k^{(1)} = \frac{\|g_k\|^2 - \mu_k |g_k^T g_{k-1}|}{d_{k-1}^T y_{k-1}}, \quad \beta_k^{(2)} = -\frac{\mu_k}{\alpha_{k-1}} \frac{g_k^T s_{k-1}}{d_{k-1}^T y_{k-1}}.$$

Then the following values can be defined:

$$\xi = \frac{\left\| -g_k + \beta_k^{(2)} d_{k-1} \right\|}{\|d_k\|}, \quad \zeta = \frac{\beta_k^{(1)} \|d_{k-1}\|}{\|d_k\|}.$$

From the definition of u_k, it can be derived that

$$u_k = \frac{d_k}{\|d_k\|} = \frac{-g_k + (\beta_k^{(1)} + \beta_k^{(2)}) d_{k-1}}{\|d_k\|} = \xi + \zeta \frac{d_{k-1}}{\|d_k\|} = \xi + \zeta u_{k-1}.$$

Since u_k is a unit vector, it follows that $\|\xi\| = \|u_k - \zeta u_{k-1}\| = \|\zeta u_k - u_{k-1}\|$. By using the triangle inequality and $\zeta > 0$, one concludes

$$\|u_k - u_{k-1}\| = 2\|\xi\|. \tag{20}$$

Using the definition of ξ, we obtain

$$\|\xi\| \|d_k\| = \left\| -g_k + \beta_{k-1}^{(2)} d_{k-1} \right\| \le \|g_k\| + \left\| \beta_{k-1}^{(2)} \right\| \|d_{k-1}\|. \tag{21}$$

By using the equations of SWP (Eq.(5)) and line search (Eq.(6)), one gets

$$d_{k-1}^T y_{k-1} \ge (\sigma - 1) g_{k-1}^T d_{k-1}, \quad \left| \frac{g_k^T d_{k-1}}{d_{k-1}^T y_{k-1}} \right| \le \left(\frac{\sigma}{1-\sigma} \right).$$

Thus,

$$\beta_k^{(2)} = -\frac{\mu_k}{\alpha_{k-1}} \frac{g_k^T s_{k-1}}{d_{k-1}^T y_{k-1}} \le \frac{E}{\alpha_{k-1}} \frac{|g_k^T s_{k-1}|}{d_{k-1}^T y_{k-1}} \le \frac{E}{\alpha_{k-1}} \frac{\|g_k\| \|s_{k-1}\|}{d_{k-1}^T y_{k-1}}.$$

By using Eq.(21), we obtain the following:

$$\begin{aligned} \|\xi\| \|d_k\| = \left\| -g_k + \beta_{k-1}^{(2)} d_{k-1} \right\| &\le \|g_k\| + \frac{E}{\alpha_{k-1}} \left| \frac{g_k^T d_{k-1}}{d_{k-1}^T y_{k-1}} \right| \|s_{k-1}\| \\ &\le \gamma + \frac{E}{\alpha_{k-1}} \left(\frac{\sigma}{1-\sigma} \right) B. \end{aligned}$$

The application of Eq.(20) leads to

$$\|u_k - u_{k-1}\| = 2\|\xi\| = 2 \frac{\gamma + \frac{E}{\alpha_{k-1}} \left(\frac{\sigma}{1-\sigma} \right) B}{\|d_k\|},$$

$$\|u_k - u_{k-1}\|^2 = 4 \frac{\left(\gamma + \frac{E}{\alpha_{k-1}} \left(\frac{\sigma}{1-\sigma} \right) B \right)^2}{\|d_k\|^2}.$$

Utilizing Eq.(19), we obtain the following:

$$\sum_{k=1}^{\infty} \frac{1}{\|d_k\|^2} \le \infty,$$

which completes the proof. $\square$

Lemma 3.7 *Assume that Assumption 1 holds and the sequences* $\{g_k\}$ *and* $\{d_k\}$ *are generated using Algorithm 1, where* α_k *is computed via the WWP line search so that the sufficient descent condition given in Eq.*(7) *holds and consider that the method satisfies Property(*). Suppose also that Eq.*(19) *holds. Then there exists a constant* $\lambda > 0$ *so that for any* $\Delta \in \mathbb{N}$ *and any index* k_0*, there exists an index* $k > k_0$ *that satisfies the following inequality:*

$$\left|\kappa_{k,\Delta}^{\lambda}\right| > \frac{\lambda}{2},$$

where $\kappa_{k,\Delta}^{\lambda} = \{i \in \mathbb{N} : k \le i \le k+\Delta-1, \|s_i\| > \lambda\}$, $\mathbb{N}$ *denotes the set of positive integers, and* $\left|\kappa_{k,\Delta}^{\lambda}\right|$ *denotes the number of elements in* $\kappa_{k,\Delta}^{\lambda}$.

From Lemmas 3.5, 3.6 and 3.7, the convergence properties of Algorithm 1 with the SWP line search can be satisfied in a manner similar to that used in Theorem 3.6 presented by Gilbert and Nocedal [7]. Therefore, the proof of the following theorem is omitted.

Theorem 3.4 *Assume that the sequences* $\{g_k\}$ *and* $\{d_k\}$ *are generated using Eqs.*(1) *and* (2) *with the CG formula* β_k^{AZHS}*, and let the step length satisfy Eqs.*(4) *and* (5)*. If Lemmas 3.5, 3.6, and 3.7 are true, then* $\liminf_{k\to\infty}\|g_k\| = 0$.

Note that if Lemma 3.1 does not hold true, then it is enough to show that

$$\beta_k^{AZHS} = \frac{\|g_k\|^2 - \mu_k \left|g_k^T g_{k-1}\right|}{d_{k-1}^T y_{k-1}}$$

satisfies Property (*) similar to Lemma 3.3 in [23].

The following theorem shows that if the second case of equation (11) holds, i.e.,

$$\beta_k^{AZHS} = -\frac{1}{\alpha_k}\mu_k \frac{g_k^T s_{k-1}}{d_{k-1}^T y_{k-1}}, \tag{22}$$

then we will obtain the result stated in Theorem 3.6.

Theorem 3.5 *Assume that Assumption 1 holds. Consider the conjugate gradient method in* (1) *and* (2) *with equation* (22)*, where* d_k *is a descent direction and* α_k *is obtained by the strong Wolfe line search. Then* $\liminf_{k\to\infty}\|g_k\| = 0$.

Proof. We will prove this theorem by contradiction. Suppose Theorem 3.5 is not true. Then equation (19) holds and

$$\begin{aligned}
\|d_k\|^2 &= \|g_k\|^2 - 2\beta_k g_k^T d_{k-1} + \beta_k^2 \|d_{k-1}\|^2 \\
&\le \|g_k\|^2 + 2\left|\beta_k\right| \left|g_k^T d_{k-1}\right| + \beta_k^2 \|d_{k-1}\|^2 \\
&\le \|g_k\|^2 + \frac{2E}{\alpha_k}\frac{\|g_k\|\,\|s_{k-1}\|}{(1-\sigma)\left|g_{k-1}^T d_{k-1}\right|}(\sigma)\left|g_{k-1}^T d_{k-1}\right| + \frac{E^2}{\alpha_k^2}\frac{(\sigma g_{k-1}^T d_{k-1})^2 \|s_{k-1}\|^2}{(1-\sigma)^2\left|g_{k-1}^T d_{k-1}\right|^2} \\
&\le \|g_k\|^2 + \frac{2E}{\alpha_k}\frac{\|g_k\|\,\|s_{k-1}\|}{(1-\sigma)}\sigma + \frac{E^2}{\alpha_k^2}\frac{\sigma^2\|s_{k-1}\|^2}{(1-\sigma)^2}.
\end{aligned}$$

Further calculation gives

$$\begin{aligned}\frac{\|d_k\|^2}{\|g_k\|^4} &\le \frac{\|g_k\|^2}{\|g_k\|^4} + \frac{2E}{\alpha_k}\frac{\|g_k\|\,\|s_{k-1}\|}{(1-\sigma)\|g_k\|^4}\sigma + \frac{E}{\alpha_k^2}\frac{\sigma^2\,\|s_{k-1}\|^2}{(1-\sigma)^2\|g_k\|^4}\\ &\le \frac{1}{\|g_k\|^2} + \frac{2E}{\alpha_k}\frac{\|s_{k-1}\|}{(1-\sigma)\|g_k\|^3}\sigma + \frac{E^2}{\alpha_k^2}\frac{\sigma^2\,\|s_{k-1}\|^2}{(1-\sigma)^2\|g_k\|^4}\\ &\le \frac{1}{\|g_k\|^2} + \frac{2E}{\alpha_k}\frac{\|s_{k-1}\|}{(1-\sigma)\|g_k\|^3}\sigma + \frac{E^2}{\alpha_k^2}\frac{\sigma^2\,\|s_{k-1}\|^2}{(1-\sigma)^2\|g_k\|^4}.\end{aligned}$$

If

$$\|g_k\|^m = \min\left\{\|g_k\|^2, \|g_k\|^3, \|g_k\|^4\right\}, \quad m \in \mathbb{N},$$

then it follows that

$$\frac{\|d_k\|^2}{\|g_k\|^4} \le \frac{1}{\|g_k\|^m}\left(1 + \frac{2E}{\alpha_k}\frac{\lambda}{(1-\sigma)}\sigma + \frac{E^2}{\alpha_k^2}\frac{\sigma^2\lambda^2}{(1-\sigma)^2}\right).$$

Also,

$$R = \left(1 + \frac{2E}{\alpha_k}\lambda\sigma + \frac{E^2}{\alpha_k^2}\frac{\sigma^2\lambda^2}{(1-\sigma)^2}\right)$$

initiates

$$\frac{\|d_k\|^2}{\|g_k\|^4} \le \frac{R}{\|g_k\|^m} \le R\sum_{i=1}^{k}\frac{1}{\|g_i\|^m} \quad \text{and} \quad \frac{\|g_k\|^4}{\|d_k\|^2} \ge \frac{\epsilon^m}{kR}.$$

Therefore,

$$\sum_{k=0}^{\infty}\frac{\|g_k\|^4}{\|d_k\|^2} = \infty.$$

This result contradicts (15). Therefore, $\liminf_{k\to\infty}\|g_k\| = 0$, completing the proof. □

4 Numerical Results and Discussion

To analyze the efficiency of the proposed method, we use more than 200 standard test functions presented in Table 1. These test functions are available from the CUTEst library [28] with the CUTEr/st test functions and SIF extension available on the website

`http://www.cuter.rl.ac.uk/Problems/mastsif.shtml`

The numerical results of CG_Descent 5.3 were obtained by running the code provided by Hager and Zhang [29] with memory set to 0. The numerical results of AZHS are obtained using a modified CG_Descent code with the SWP line search, employing $\sigma = 0.1$ and $\delta = 0.01$. If $\mu_k > 1$, then we conclude that $L < 1$ and $\frac{\|g_k\|^2}{|g_k^T g_{k-1}|} > 1$. Thus, it is reasonable to modify Eq.(11) as follows:

$$\beta_k^{AZHS} = \begin{cases} \frac{\|g_k\|^2 - |g_k^T g_{k-1}|}{d_{k-1}^T y_{k-1}}, & \text{if } \|g_k\|^2 > |g_k^T g_{k-1}|,\\ \frac{\|g_k\|^2 - \mu_k|g_k^T g_{k-1}|}{d_{k-1}^T y_{k-1}} - \frac{1}{\alpha_k}\mu_k\frac{g_k^T s_{k-1}}{d_{k-1}^T y_{k-1}}, & \text{if } \|g_k\|^2 > \mu_k\,|g_k^T g_{k-1}|,\\ -\frac{1}{\alpha_k}\mu_k\frac{g_k^T s_{k-1}}{d_{k-1}^T y_{k-1}}, & \text{otherwise.}\end{cases}$$

Note that if $\beta_k^{AZHS} = \frac{\|g_k\|^2 - |g_k^T g_{k-1}|}{d_{k-1}^T y_{k-1}}$, then $\beta_k^{AZHS} \leq \beta_k^{HS}$, thus the proof will be similar to that presented in [7].

The host computer used was an AMD A4-7210 APU with AMD Radeon R3 Graphics, 4 GB RAM, and a 64-bit operating system. The graphs on the following results were obtained using SigmaPlot, a performance measure introduced by Dolan and Moré [30].

This performance measure compares the performance of a set of solvers S on a set of problems ρ. For n_s solvers and n_p problems in S and ρ, respectively, the measure $t_{p,s}$ is the computation time (e.g., the number of iterations or CPU time) required for solver s to solve problem p.

To establish a baseline for comparison, the performance of solver s on problem p is scaled relative to the best performance of any solver in S on that problem, yielding the ratio

$$r_{p,s} = \frac{t_{p,s}}{\min\{t_{p,s} : s \in S\}}.$$

A parameter $r_M \geq r_{p,s}$ for all p, s is selected such that $r_{p,s} = r_M$ if and only if solver s cannot solve problem p. To obtain an overall assessment of the performance of each solver, we define the measure

$$P_s(t) = \frac{1}{n_p} \text{size}\{p \in \rho : r_{p,s} \leq t\}.$$

$P_s(t)$ is the probability for solver $s \in S$ that the performance ratio $r_{p,s}$ will be within a factor $t \in \mathbb{R}$ of the best possible ratio. If we denote the cumulative distribution function of the performance ratio as p_s, then the performance measure $p_s : \mathbb{R} \to [0, 1]$ for a given solver is non-decreasing and piecewise continuous from the right. The value of $p_s(1)$ is the probability that the solver will achieve the best performance among all solvers. In general, a solver with higher values of $P_s(t)$, which will lie closer to the upper right corner of the figure, is preferable.

The numerical results are shown in Figures 1, 2, 3 and 4. Figure 1 depicts the number of iterations, showing that the new modification significantly outperforms CG_Descent 5.3. Figure 2 illustrates that the new modification, AZHS, outperforms CG_Descent 5.3 in the number of function evaluations. Figures 3 and 4 show the performance based on the number of gradient evaluations and CPU time, respectively. It is observed that AZHS outperforms CG_Descent 5.3 in CPU time and is significantly competitive with CG_Descent 5.3 in the number of function evaluations and gradient evaluations as the latter used an approximate Wolfe line search with $\sigma = 0.9$ and $\delta = 0.1$. Thus, we can conclude that β_k^{AZHS} outperforms CG_Descent 5.3 in all figures.

5 Application to Heat Conduction Problem [32]

Suppose a rectangular flat plate with dimensions of 5×4 units generates heat [33]. Suppose the thermal conductivity k is fixed, and the heat production per unit area f is a nonlinear function of the temperature M. Our objective is to define the temperature of the slab such that the temperature outside the perimeter of the slab is zero. Poisson's equation classifies the temperature distribution within this region as follows:

$$k\left[\frac{\partial^2 M}{\partial x^2} + \frac{\partial^2 M}{\partial y^2}\right] + f(M) = 0.$$

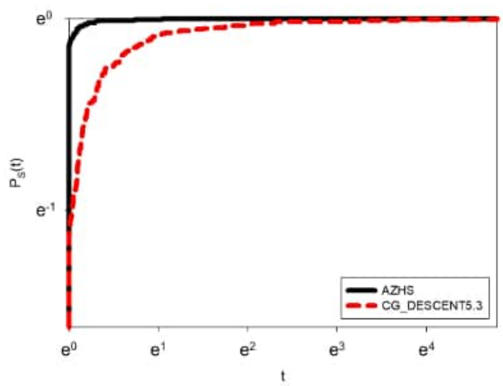


Figure 1: Performance measure based on the number of iterations.

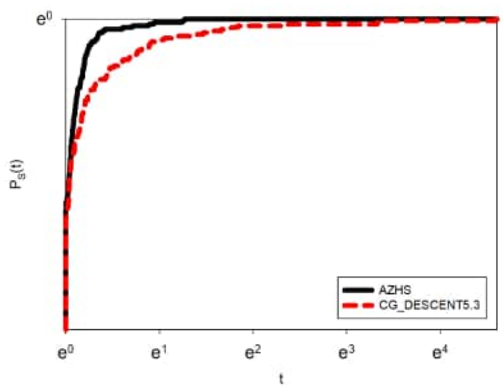


Figure 2: Performance measure based on the function evaluation.

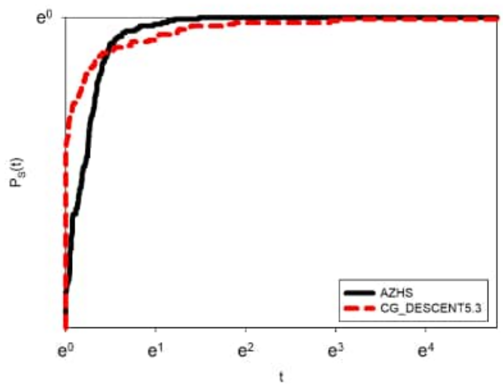


Figure 3: Performance measure based on the gradient evaluation.

Function	Dim	Function	Dim	Function	Dim
AKIVA	2	FBRAIN2LS	4	OSCIPATH	10
ALLINITU	4	FLETCBV2	5000	PALMER1C	8
ARGLINB	200	FLETCHCR	1000	PALMER1D	7
ARGLINC	200	FMINSRF2	5625	PALMER2C	8
ARWHEAD	5000	FMINSURF	5625	PALMER3C	8
BARD	3	GENHUMPS	5000	PALMER4C	8
BDEXP	5000	GROWTHLS	3	PALMER5C	6
BDQRTIC	5000	GULF	3	PALMER6C	8
BEALE	2	HAHN1LS	7	PALMER7C	8
BIGGS3	6	HAIRY	2	PALMER8C	8
BIGGS5	6	HATFLDD	3	PARKCH	15
BIGGS6	6	HATFLDE	3	PENALTY1	1000
BIGGSB1	5000	HATFLDFL	3	PENALTY2	200
BOX2	3	HATFLDFLS	3	PENALTY3	200
BOX3	3	HEART6LS	6	PENALTY3	200
BOX	10000	HEART8LS	8	POWELLBSLS	2
BRKMCC	2	HELIX	3	POWELLSG	5000
BROYDNBDLS	10	HIELOW	3	POWER	10000
BROWNAL	200	HILBERTA	2	POWERSUM	4
BROWNBS	2	HILBERTB	10	PRICE3	2
BROWNDEN	4	HIMMELBB	2	PRICE4	2
BROYDN7D	5000	HIMMELBF	4	QING	100
BRYBND	5000	HIMMELBG	2	QUARTC	5000
CAMEL6	2	HIMMELBH	2	RAT43LS	4
CHNROSNB	50	HUMPS	2	RECIPELS	3
CLIFF	2	HYDCAR6LS	29	ROSENBR	2
COSINE	10000	INDEF	5000	ROSENBRTU	2
CUBE	2	INDEFM	100000	S308	2
CURLY10	10000	INTEQNELS	12	SCHMVETT	5000
CURLY20	10000	JENSMP	2	SENSORS	100
CURLY30	10000	JIMACK	3549	SINEVAL	2
DENSCHNA	2	JUDGE	2	SINQUAD	5000
DENSCHNB	2	KOWOSB	4	SISSER	2
DENSCHNC	2	KSSLS	1000	SNAIL	2
DENSCHND	3	LANCZOS1LS	6	SPMSRTLS	4999
DENSCHNE	3	LANCZOS2LS	6	SROSENBR	5000
DENSCHNF	2	LANCZOS3LS	6	SSCOSINE	5000
DIXMAANA	3000	LIARWHD	5000	SSI	3
DIXMAANB	3000	LOGHAIRY	2	STREG	4
DIXMAANC	3000	LSC1LS	3	STRATEC	10
DIXMAAND	3000	LSC2LS	3	STRTCHDV	10
DIXMAANE	3000	LUKSAN11LS	100	TESTQUAD	5000
DIXMAANF	3000	LUKSAN12LS	98	THURBERLS	7
DIXMAANG	3000	LUKSAN13LS	98	TOINTGOR	50
DIXMAANH	3000	LUKSAN14LS	98	TOINTGSS	5000
DIXMAANI	3000	LUKSAN15LS	100	TOINTPSP	50
DIXMAANJ	3000	LUKSAN16LS	100	TOINTQOR	50
DIXMAANK	3000	MANCINO	100	TQUARTIC	5000
DIXMAANL	3000	MARATOSB	2	TRIDIA	5000
DIXMAANP	3000	MEXHAT	2	TRIGON1	10
DIXON3DQ	10000	MEYER3	3	TRIGON2	10
DJTL	2	MGH09LS	4	VANDANMSLS	22
DMN15332LS	66	MGH10LS	3	VARDIM	200
DQDRTIC	5000	MGH10SLS	3	VAREIGVL	50
ECKERLE4LS	3	MGH17LS	5	VESUVIALS	8
EDENSCH	2000	MISRA1BLS	2	VESUVIOULS	8
EGGCRATE	2	MISRA1CLS	2	VIBRBEAM	8
EG2	1000	MISRA1DLS	2	WAYSEA1	2
EIGENALS	2550	MODBEALE	20000	WAYSEA2	2
EIGENBLS	2550	MOREBV	5000	WOODS	4000
EIGENCLS	2652	MSQRTALS	1024	YATP1CLS	123200
ELATVIDU	2	MSQRTBLS	1024	YATP2CLS	123200
ENGVAL1	5000	NCB20	5010	YFITU	3
ENGVAL2	3	NELSONLS	3	ZANGWIL2	2
ENSOLS	9	NONCVXU2	5000		
EXPFIT	2	NONDIA	5000		

Table 1: Test functions.

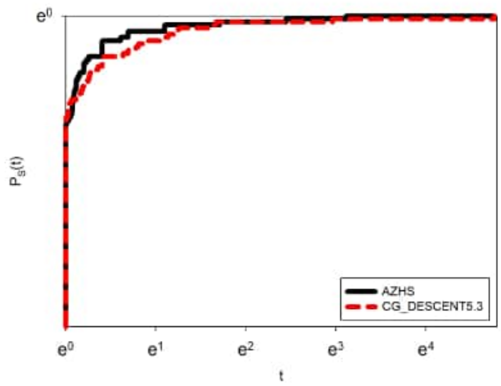


Figure 4: Performance measure based on the CPU time.

If $k = 2$ and $f(M) = 20 - \frac{3}{2}M + \frac{1}{20}M^2$, there are 12 mesh points in total. Symmetry reduces the problem to only four distinct temperatures.

$$2(M_2 + M_3 - 4M_1) = -20 + \frac{3}{2}M_1 - \frac{1}{20}M_1^2,$$

$$2(M_3 + M_1 + M_4 - 4M_3) = -20 + \frac{3}{2}M_3 - \frac{1}{20}M_3^2,$$

$$2(M_1 + M_4 + 4M_2) = -20 + \frac{3}{2}M_2 - \frac{1}{20}M_2^2,$$

$$2(2M_3 + M_2 - 3M_4) = -20 + \frac{3}{2}M_4 - \frac{1}{20}M_4^2.$$

These equations, expressed in powers of M_1, are as follows:

$$\left(M_1^2 - 190M_1\right) + 40\left(M_2 + M_3 + 10\right) = 0,$$

$$M_1 + \frac{M_3^2 - 150M_3 + 400}{40} + M_4 = 0,$$

$$2M_1 + \frac{M_2^2 - 190M_2 + 400}{40} + M_4 = 0,$$

$$\left(M_4^2 - 150M_4\right) + 40M_2 + 80M_3 + 400 = 0.$$

The objective function f is constructed by summing the squares of the functions connected with each nonlinear equation as follows:

$$f(M_1, M_2, M_3, M_4, H_1, H_2, H_3, H_4, H_5, H_6) = Q_1 + Q_2 + Q_3 + Q_4,$$

where

$$Q_1 = Q_5^2, \quad Q_2 = Q_6^2, \quad Q_3 = Q_7^2, \quad Q_4 = Q_8^2,$$

$$Q_5 = \frac{1}{20}\left[M_1^2 + H_1 M_1 + H_2\left(M_2 + M_3 + H_3\right)\right],$$

$$Q_6 = 2\left[M_1 + \frac{M_3^2 + H_4 M_3}{H_2} + H_5 + M_4\right],$$

$$Q_7 = 2\left[H_6 M_1 + \frac{M_2^2 + H_1 M_2}{H_2} + H_5 + M_4\right],$$

$$Q_8 = \frac{1}{20}\left[M_4^2 + H_4 M_4 + H_2 M_2 + H_2 H_6 M_3 + H_2 H_5\right].$$

If

$$H_1 = -190, \quad H_2 = 40, \quad H_3 = 10, \quad H_4 = -150, \quad H_5 = 10, \quad H_6 = 2,$$

let

$$M_1 = x_1, \quad M_2 = x_2, \quad M_3 = x_3, \quad M_4 = x_4.$$

Then, we obtain the following function:

$$f(x_1,x_2,x_3,x_4) = \left(2(x_2+x_3-4x_1)+20-1.5x_1+\frac{x_1^2}{20}\right)^2$$
$$+\left(2(x_1-3x_3+x_4)+20-1.5x_3+\frac{x_3^2}{20}\right)^2$$
$$+\left(2(2x_1+x_4-4x_2)+20-1.5x_2+\frac{x_2^2}{20}\right)^2+\left(2(x_2+2x_3-3x_4)+20-1.5x_4+\frac{x_4^2}{20}\right)^2.$$

We say that $f(x_1,x_2,x_3,x_4)$ is the Heat Conduction Problem function. By using Algorithm 1, we can find the values of x_1,x_2,x_3,x_4 as follows:

$$x_1 = 4.8521, \quad x_2 = 6.0545, \quad x_3 = 6.4042, \quad x_4 = 8.1383.$$

The function value is 1.9631×10^{-7}.

6 Application to Image Restoration

Restoring damaged images is one of the most important applications of the CG method. In this study, we applied Gaussian noise with a standard deviation of 25% to the original images in Table 3. After that, we used Algorithm 1 to restore these images. To express the efficiency of the proposed method, we made a comparison between Algorithm 1, CG-Descent5.3, and DL+ in terms of the number of iterations, CPU time, and root-mean-square error (RMSE).

We utilized the RMSE between the restored image and the original true image to calculate the quality of the restored image:

$$\text{RMSE} = \frac{\|\nu - \nu_k\|_2}{\|\nu\|}.$$

The restored image is denoted by ν_k and the true image by ν. The RMSE determines the quality of the restored image, in which lower values correspond to higher quality. The results in Table 2 show that the new search direction outperforms CG-Descent5.3 and DL+ in terms of the number of iterations, CPU time, and the RMSE value. The criteria for stopping is

$$\frac{\|x_{k+1} - x_k\|_2}{\|x_k\|_2} < \varepsilon.$$

In this context, $\epsilon = 10^{-3}$. Note that if $\epsilon = 10^{-4}$ or $\epsilon = 10^{-6}$, then the RMSE remains fixed, meaning that a fixed RMSE can have a variation in the number of iterations.

Table 3 below shows the outcomes of restoring the destroyed images using Algorithm 1, indicating that it can be regarded as an efficient approach.

Image	Algorithm	Number of Iteration	CPU Time (s)	RMSE
Mandi 128 pixels	DL+	127	1.724e+000	0.1003
	AZHS	126	1.663e+000	0.1002
	CG-Descent5.3	134	1.825e-001	0.1004
Coins 128 pixels	DL+	135	1.542e+000	0.0832
	AZHS	130	1.491e+000	0.0824
	CG-Descent5.3	133	1.491e+000	0.0831
Mandi 256 pixels	DL+	120	1.856e+001	0.0519
	AZHS	111	1.545e+001	0.0510
	CG-Descent5.3	119	1.656e+001	0.0991
Coins 256 pixels	DL+	134	1.447e+001	0.0506
	AZHS	120	1.164e+001	0.0501
	CG-Descent5.3	130	1.564e+001	0.0508
Mandi 512 pixels	DL+	114	7.981e+001	0.0371
	AZHS	105	6.755e+001	0.0360
	CG-Descent5.3	116	7.314e+001	0.0472
Kids 512 pixels	DL+	57	6.955e+001	0.0377
	AZHS	56	5.325e+001	0.0384
	CG-Descent5.3	55	5.634e+001	0.0395
Coins 512 pixels	DL+	129	7.323e+001	0.0326
	AZHS	128	5.248e+001	0.0324
	CG-Descent5.3	127	6.323e+001	0.0503
Coins 1024 pixels	DL+	128	3.441e+002	0.0326
	AZHS	110	2.549e+002	0.0172
	CG-Descent5.3	124	2.897e+002	0.0289

Table 2: Numerical outcomes from the images with Gaussian noise with a 25% standard deviation added to the original images using the Dai-Liao CG method, AZHS, as well as CG-Descent5.3.

7 Conclusion

In this study, we investigate a modified Hestenes–Stiefel (HS) conjugate gradient (CG) method based on the Dai–Liao conjugacy parameter, with the restart property depending on L. The newly modified CG method inherits global convergence properties and a sufficient descent condition through the SWP line search. Moreover, the numerical results are efficient and competitive with CG Descent5.3. Applications to solving the Heat Conduction Problem and image restoration are presented. In future studies, we will focus on the Lipschitz constant because it plays an essential role in the efficiency and robustness of the CG method.

Acknowledgment

We would like to thank Prof. William W. Hager for publishing his code in the implementation of the CG method. Additionally, we extend our gratitude to the editor and the reviewers for their valuable feedback and suggestions.

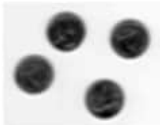
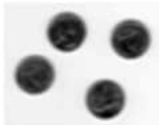
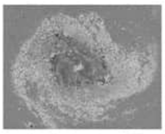
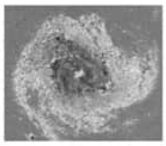
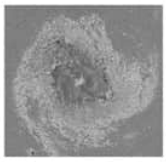
Image	**Original Image**	**Image with Gaussian Noise**	**Restored Image**
Mandi (128 pixels)			
Mandi (256 pixels)			
Coins (256 pixels)			
Kids (512 pixels)			
M.83 (1024 pixels)			

Table 3: Restoration of the destroyed images of Mandi, Coins, Kids, as well as M.83 by reducing z via Algorithm 1.

References

[1] P. Wolfe. Convergence conditions for ascent methods. II: Some corrections. *SIAM Rev.* **13** (2) (1971) 185–188.

[2] S. Yao, L. Ning, H. Tu and J. Xu. A one-parameter class of three-term conjugate gradient methods with an adaptive parameter choice. *Optim. Methods Softw.* **35** (6) (2020) 1051–1064.

[3] M. R. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. *J. Res. Nat. Bur. Stand.* **49** (1952) 409–436.

[4] R. Fletcher and C. M. Reeves. Function minimization by conjugate gradients. *Comput. J.* **7** (1964) 149–154.

[5] P. Elijah and G. Ribiere. Note sur la convergence de méthodes de directions conjuguées. *ESAIM: Math. Model. Numer. Anal.* **3** R1 (1969) 35–43.

[6] M. J. D. Powell. Nonconvex minimization calculations and the conjugate gradient method. In: *Numerical Analysis (Dundee, 1983)*, Lecture Notes in Math. **1066**, Springer, Berlin, 1984, 122–141.

[7] J. C. Gilbert and J. Nocedal. Global convergence properties of conjugate gradient methods for optimization. *SIAM J. Optim.* **2** (1992) 21–42.

[8] M. Al-Baali. Descent property and global convergence of the Fletcher–Reeves method with inexact line search. *IMA J. Numer. Anal.* **5** (1985) 121–124.

[9] Y. H. Dai and L. Z. Liao. New conjugacy conditions and related nonlinear conjugate gradient methods. *Appl. Math. Optim.* **43** (1) (2014) 87–101.

[10] S. Babaie-Kafaki and R. Ghanbari. The Dai–Liao nonlinear conjugate gradient method with optimal parameter choices. *Eur. J. Oper. Res.* **234** (3) (2014) 625–630.

[11] S. Babaie-Kafaki and R. Ghanbari. A descent family of Dai–Liao conjugate gradient methods. *Optim. Methods Softw.* **29** (3) (2014) 583–591.

[12] A. Perry. A modified conjugate gradient algorithm. *Oper. Res.* **26** (1976) 1073–1078.

[13] N. Andrei. A Dai–Liao conjugate gradient algorithm with clustering of eigenvalues. *Numer. Algorithms* **77** (4) (2018) 1273–1282.

[14] W. W. Hager and H. Zhang. A new conjugate gradient method with guaranteed descent and an efficient line search. *SIAM J. Optim.* **16** (2005) 170–192.

[15] W. W. Hager and H. Zhang. The limited memory conjugate gradient method. *SIAM J. Optim.* **23** (2013) 2150–2168.

[16] K. Zhang, H. Liu and Z. Liu. A new Dai–Liao conjugate gradient method with optimal parameter choice. *Numer. Funct. Anal. Optim.* **40** (2) (2019) 194–215.

[17] M. Al-Baali, Y. Narushima and H. Yabe. A family of three-term conjugate gradient methods with sufficient descent property for unconstrained optimization. *Comput. Optim. Appl.* **60** (2015) 89–110.

[18] Z. F. Dai, H. Y. Zhu, and X. Zhang. Dynamic spillover effects and portfolio strategies between crude oil, gold and Chinese stock markets related to new energy vehicle. *Energy Economics* **109** (2022) 105959.

[19] Z. F. Dai, H. Zhuo, J. Kang, and F. Wen. The skewness of oil price returns and equity premium predictability. *Energy Economics* **94** (2021) 105069.

[20] Z. F. Dai and H. Zhu, Forecasting stock market returns by combining sum-of-the-parts and ensemble empirical mode decomposition. *Applied Economics* **52** (2020) 2309–2323.

[21] Z. Dai and F. Wen, Another improved Wei–Yao–Liu nonlinear conjugate gradient method with sufficient descent property. *Applied Mathematics and Computation* **218** (14) (2012) 7421–7430.

[22] Z. Salleh, A. Almarashi and A. Alhawarat. Two efficient modifications of AZPRP conjugate gradient method with sufficient descent property. *Journal of Inequalities and Applications* (1) (2022) 1–21.

[23] A. Alhawarat, Z. Salleh, M. Mamat and M. Rivaie. An efficient modified Polak–Ribière–Polyak conjugate gradient method with global convergence properties. *Optim. Methods Softw.* **32** (6) (2017) 1299–1312.

[24] S. Yao, L. Ning, H. Tu and J. Xu. A one-parameter class of three-term conjugate gradient methods with an adaptive parameter choice. *Optimization Methods and Software* **35** (6) (2020) 1051–1064.

[25] G. Zoutendijk. Nonlinear programming, computational methods, Integer Nonlinear Program. **143** (1970) 37–86.

[26] Y. H. Dai, J. Y. Han, G. H. Liu, D. F. Sun, H. X. Yin and Y. Yuan. Convergence properties of nonlinear conjugate gradient methods. *SIAM J. Optim.* **10** (2) (1999) 348–358.

[27] G. Yuan and Z. Wei. Non-Monotone Backtracking Inexact BFGS Method for Regression Analysis. *Communications in Statistics-Theory and Methods* **42** (2) (2013) 214–238.

[28] I. Bongartz, A.R. Conn, N. Gould and P.L. Toint, CUTE: Constrained and unconstrained testing environment, *ACM Trans. Math. Softw.* **21** (1995) 123–160.

[29] `http://users.clas.ufl.edu/hager/papers/Software/` (2024).

[30] E. D. Dolan and J. J. Moré, Benchmarking optimization software with performance profiles, *Math. Program.* **91** (2002) 201–213.

[31] S. K. Mishra and B. Ram. Introduction to unconstrained optimization with R. Springer Nature, 2019.

[32] A. J. Surkan and C. L. Solution of reaction and heat flow problems by nonlinear estimation. The Canadian Journal of Chemical Engineering **46** (4) (1968) 229–232.

Nonlinear Dynamics and Systems Theory, **25** (3) (2025) 288–298

A New Memristor-Based 4D Hyperchaotic System with Seven Terms and No Equilibrium Points

M. I. Kopp *

Institute for Single Crystals, NAS of Ukraine, Nauky Ave. 60, Kharkiv 61072, Ukraine

Received: October 8, 2024; Revised: May 24, 2025

Abstract: The most pressing challenge in the practical application of chaotic systems is the development of methods for encrypting information. This paper presents a new 4-dimensional (4D) memristive system that is simple, consisting of only seven terms and lacking equilibrium points, which allows it to generate hidden attractors. The paper thoroughly analyzes the system's dynamic properties, including bifurcation diagrams, Lyapunov exponents, Kaplan-York dimensions, and offset boosting analysis. Additionally, the theoretical model is validated through electronic simulation of the new two-winged chaotic system using Multisim.

Keywords: *two-wing attractors; memristor; chaotic behavior; offset boosting control; circuit implementation.*

Mathematics Subject Classification (2020): 34A34, 34D45, 70K43, 93C15.

1 Introduction

A rapidly expanding area within nonlinear circuit theory is the development of chaos generators utilizing memristors. First introduced by Chua [1], the memristor is a device that links electric charge and magnetic flux, functioning as a resistor with memory. Since then, the concept has evolved to include a broader spectrum of memristive systems. HP Laboratories achieved the first successful implementation of a memristor, using a metal-dielectric-metal structure [2]. However, significant technological challenges in memristor fabrication have led to a considerable gap between theoretical models and experimental studies. Memristors have found extensive applications in fields such as image encryption, signal processing, biosystems, and neural networks, particularly in complex neural networks [3]. Their popularity is largely due to the complex dynamics achievable in chaotic

* Corresponding author: mailto:michaelkopp0165@gmail.com

systems based on memristors. Numerous chaotic and hyperchaotic systems have been developed with memristors serving as nonlinear elements. Furthermore, various equivalent circuits for modeling memristor emulators have been proposed in the literature [4]. This brief literature review focuses on 4D chaotic systems that integrate memristors. Chaotic dynamical systems can be categorized into two types: those with self-excited attractors and those with hidden attractors. A self-excited attractor has a basin of attraction that intersects with the vicinity of an equilibrium point, whereas a hidden attractor's basin does not intersect with any equilibrium point's neighborhood. The concept of hidden attractors, first introduced in [5], has since sparked ongoing research in nonlinear science. As noted in [6], hidden attractors in dynamical systems are currently classified into five categories: 1) systems without equilibria, 2) equilibrium curves, 3) planes of curves, 4) equilibrium lines, and 5) stable equilibrium points.

Recently, a new 4D hyperchaos system without equilibrium point was proposed in [7]. Several researchers have developed memristor-based 4D hyperchaotic systems characterized by the absence of equilibrium points. In [8], a 4D memristive system is introduced, consisting of 12 terms, 5 of which are nonlinear. This model notably lacks equilibrium points and exhibits periodic, chaotic, and hyperchaotic behavior within specific parameter ranges. In [9], a 4D memristive system is presented that can display either no equilibrium points or an equilibrium line, depending on the control parameter. The study shows that by adjusting this parameter, the system can transition between chaotic and hyperchaotic dynamics. This nonlinear system comprises 11 terms, including 5 nonlinear ones. A simpler 4D chaotic memristor-based system, consisting of 9 terms with 2 nonlinearities and no equilibrium points, is described in [10]. An even more streamlined 4D memristive two-scroll chaotic system, containing only 7 terms and 3 nonlinearities, is introduced in [11]. This system demonstrates various complex dynamics such as offset boosting, remerging period-doubling bifurcations, and hidden extreme multistability. Furthermore, a 4D hyperchaotic hyperjerk system with a line equilibrium, composed of 7 algebraic terms and a single nonlinearity, is proposed in [12]. Interestingly, the system in [12] is based on intrinsic memristive nonlinearity, a type of nonlinearity that arises naturally from the memristor itself.

In this paper, we present a new hyperchaotic dynamical system (not a jerk system) developed by introducing nonlinearity via a memristor. Our primary motivation is to design a novel 4D hyperchaotic memristor-based system with the fewest possible terms.

2 Derivation and Key Properties of a Novel 4D Memristive Hyperchaotic System

In this section, we introduce a novel 4D memristive two-wing chaotic system comprising only seven terms. This system is based on the one introduced in [13] and is defined as follows:

$$\begin{cases} \dfrac{dx_1}{dt} = a(-x_1 + x_2), \\ \dfrac{dx_2}{dt} = -x_3 \mathrm{sgn}(x_1), \\ \dfrac{dx_3}{dt} = |x_1| - 1, \end{cases} \tag{1}$$

where $|x|$ is the absolute value function and the signum function $\mathrm{sgn}(x)$ of a real number x is a piecewise function. Figure 1 displays typical two-wing butterfly attractors in various

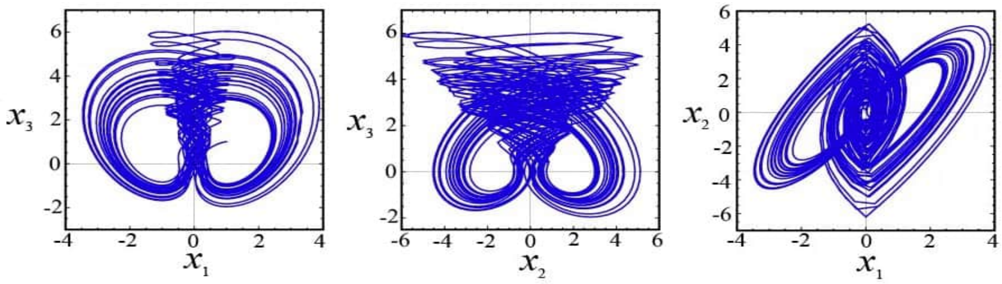


Figure 1: Plots depict the two-wing butterfly attractors of system (1) in the phase planes x_1x_3, x_2x_3, and x_1x_2, respectively.

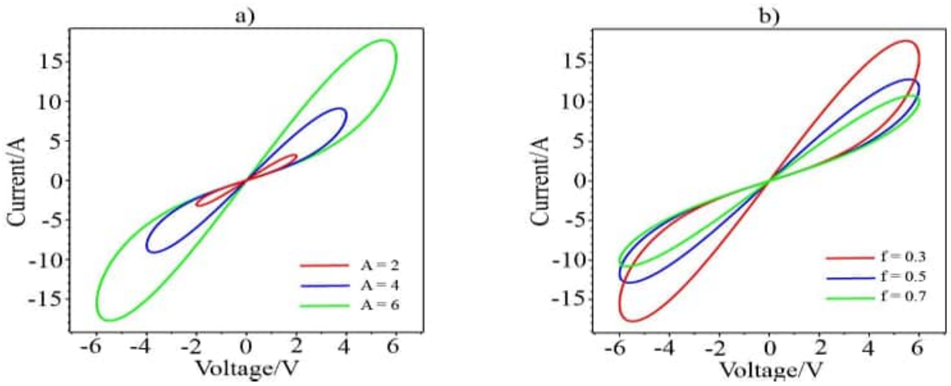


Figure 2: Simulation results of the hysteresis loop for the memductance function $W(\varphi) = 1 + 0.5|\varphi|$: a) different values of amplitude A; b) different values of frequency f.

phase planes for system (1), with parameter set at $a = 0.6$ and the initial conditions $x_1(0) = x_2(0) = x_3(0) = 1$. Using the methodology of Binouse et al. [14], we can compute all LEs:

$$LE_1 = 0.191212, \quad LE_2 \approx 0, \quad LE_3 = -0.799337,$$

and the corresponding Kaplan-Yorke (or Lyapunov) dimension $D_{KY} \approx 2.239$. We see that system (1) demonstrates chaotic behavior with one positive exponent $LE_1 > 0$.

To achieve hyperchaotic behavior, system (1) is extended to 4D by adding a state variable linked to the original system through a memristor. We use the model of an absolute memristor, specifically Bao's magnetically controlled memristor [15], described by the following equations:

$$\begin{cases} i_m = W(\varphi)u_m, \\ \frac{d\varphi}{dt} = u_m, \\ W(\varphi) = \alpha + \beta|\varphi|. \end{cases} \tag{2}$$

In the equations (2), the symbols u_m, i_m, and φ represent the input, output, and state variables of the memory device, respectively. The function φ corresponds to the magnetic flux, while α and β are constant coefficients, set to $\alpha = 1$ and $\beta = 0.5$. The graph for

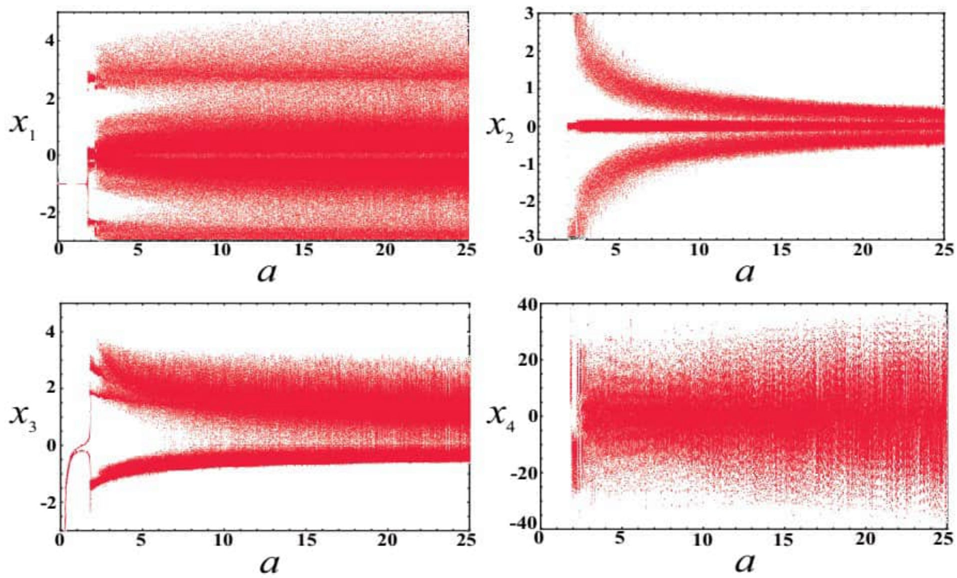


Figure 3: Bifurcation diagrams for the components x_1, x_2, x_3, x_4 of the system (3).

system (2) displays a smooth quadratic nonlinear characteristic curve passing through the origin. Driven by a sinusoidal AC voltage source $u_m = A\sin(2\pi ft)$, where A is the amplitude and f is the frequency, the memristor circuit simulation results, shown in Figure 2, reveal a current-voltage characteristic forming a closed hysteresis loop. As the frequency f increases, the area of the loop decreases, while increasing the amplitude A causes its expansion, consistent with the fundamental properties of memristors.

Integrating the expressions from (2) into the nonlinear dynamic equations (1) yields a novel set of memristor-based 4D equations:

$$\begin{cases} \dfrac{dx_1}{dt} = ax_2 - (\alpha + \beta|x_4|)x_1, \\ \dfrac{dx_2}{dt} = -x_3\mathrm{sgn}(x_1), \\ \dfrac{dx_3}{dt} = |x_1| - 1, \\ \dfrac{dx_4}{dt} = x_1. \end{cases} \tag{3}$$

Here, instead of the notation of flux φ, we introduced a new dynamical variable x_4. As shown in (3), the system includes only seven terms. It represents the minimum number of terms needed for chaotic dynamics in a four-dimensional autonomous system, making it a rare configuration in the literature.

Let us outline some fundamental dynamic properties of the new 4D system. It is readily verifiable that system (3) exhibits symmetry with respect to the x_3-axis and remains invariant under the transformation $(x_1, x_2, x_3, x_4) \to (-x_1, -x_2, x_3, -x_4)$. To further characterize the system's behavior, we calculate its divergence as follows:

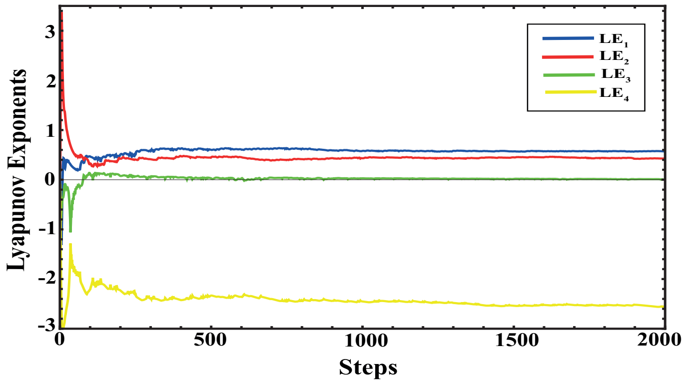


Figure 4: Lyapunov exponents for the system (3) for the parameter value $a = 21$ and initial conditions (4).

$$\frac{\partial \dot{x}_1}{\partial x_1} + \frac{\partial \dot{x}_2}{\partial x_2} + \frac{\partial \dot{x}_3}{\partial x_3} + \frac{\partial \dot{x}_4}{\partial x_4} = -(\alpha + \beta|x_4|).$$

Thus, the system (3) is dissipative for all positive values of parameters. Setting the right-hand side of system (3) to zero $\dot{x}_1 = \dot{x}_2 = \dot{x}_3 = \dot{x}_4 = 0$ yields $x_1 = 0$ from the fourth equation. Substituting this value into the third equation leads to the contradictory result $-1 = 0$, indicating that no equilibrium points exist for the system. This implies that all attractors generated by system (3) are hidden.

2.1 Bifurcation diagrams, Lyapunov exponents, and the calculation of Kaplan-Yorke dimension

In dynamic analysis, a bifurcation diagram visually represents changes in the system's state variables. We use the NDSolve function in Mathematica to solve the equations in (3) under the following initial conditions:

$$x_1(0) = x_2(0) = x_3(0) = x_4(0) = 1. \tag{4}$$

In system (3), the parameter a varies while $\alpha = 1, \beta = 0.1$ remain constant. The bifurcation diagrams in Figure 3 show the components as $a \in [0, 25]$, highlighting stable regions (distinct points), periodic, quasiperiodic, and chaotic behaviors. Examining the Lyapunov exponents provides deeper insight into the system's stability and chaotic characteristics as a changes. The dynamical behaviors of system (3) can be classified into the following categories, as detailed in Table 1. A positive LE indicates instability or chaos within the system, while a negative LE suggests a tendency toward stable equilibrium. Next, we focus on the hyperchaotic behavior of the system (3) at $a = 21$. The sum of all Lyapunov exponents is negative, confirming the dissipative nature of the system. The dynamics are illustrated in Figure 4. To assess the complexity of the attractor, we can calculate the Lyapunov or Kaplan-Yorke dimension:

$$D_{KY} = \xi + \frac{1}{|LE_{\xi+1}|} \sum_{i=1}^{\xi} LE_i = 3 + \frac{1.0041}{2.5617} \approx 3.3919, \tag{5}$$

a	Lyapunov Exponents (LE_1, LE_2, LE_3, LE_4)	Signs	Behavior
$a = 0.01$	$(\mathbf{0.0081}, -0.0148, -0.0081, -11.0423)$	$(0, -, -, -)$	Periodic
$a = 0.45$	$(\mathbf{0.0053}, \mathbf{-0.0012}, \mathbf{-0.0043}, -12.3353)$	$(0, 0, 0, -)$	Quasi − periodic 3 − torus
$a = 1$	$(\mathbf{0.0201}, \mathbf{0.0038}, -0.0165, -8.3231)$	$(0, 0, -, -)$	Quasi − periodic 2 − torus
$a = 1.85$	$(0.0881, \mathbf{-0.0057}, -0.0140, -3.8193)$	$(+, 0, -, -)$	Chaotic
$a = 3$	$(0.0936, \mathbf{0.0080}, \mathbf{0.0038}, -3.8193)$	$(+, 0, 0, -)$	Chaotic 2 − torus
$a = 4$	$(0.1471, \mathbf{0.0520}, -0.0986, -1.9904)$	$(+, 0, -, -)$	Chaotic
$a = 15$	$(0.4646, 0.2873, \mathbf{-0.0077}, -2.2215)$	$(+, +, 0, -)$	Hyperchaotic
$a = 21$	$(0.5702, 0.4303, \mathbf{0.0035}, -2.5617)$	$(+, +, 0, -)$	Hyperchaotic

Table 1: Lyapunov exponents for different values of the parameter a.

where ξ is determined from the conditions

$$\sum_{i=1}^{\xi} LE_i > 0 \;\Rightarrow\; \sum_{i=1}^{3} LE_i = 1.0041, \quad \sum_{i=1}^{\xi+1} LE_i < 0 \;\Rightarrow\; \sum_{i=1}^{4} LE_i = -1.5577 < 0.$$

Here, ξ denotes the number of first non-negative Lyapunov exponents. The Kaplan-Yorke dimension D_{KY} (5) is fractional and is found to be significantly higher for system (3) than for the chaotic system (1), indicating greater dynamic complexity.

2.2 Phase portraits of hidden attractors and offset boosting control

We created phase portraits and time diagrams for the hyperchaotic system (3), shown in Figure 5. Implementing system (3) in an electronic circuit is challenging because the dynamic variable x_4 exceeds the power supply limits of operational amplifiers. To address this, we transform x_4 to $x_4 = 20X_4$ and rename the other variables as $x_1 = X_1$, $x_2 = X_2$, and $x_3 = X_3$. The transformed hyperchaotic system takes the form

$$\begin{cases} \dfrac{dX_1}{dt} = 21X_2 - (1 + 2|X_4|)X_1, \\ \dfrac{dX_2}{dt} = -X_3\mathrm{sgn}(X_1), \\ \dfrac{dX_3}{dt} = |X_1| - 1, \\ \dfrac{dX_4}{dt} = 0.05X_1. \end{cases} \tag{6}$$

The transformed system (6) will be utilized to create an analog chaos generator circuit in the following section.

The offset boosting control method is commonly used in hyperchaotic systems to shift the attractor by introducing a bias. By adding a constant to specific variables, chaotic signals can be manipulated within phase space. In system (3), x_4 appears in the first equation and x_3 in the second. We can control these variables by replacing x_4 with $x_4 + m$ and x_3 with $x_3 + k$, where m and k are constants. As shown in Figure 6, modifying m transforms x_4 from a bipolar to a unipolar signal and shifts the attractor along the x_4-axis. Similar effects for x_3 are illustrated in Figure 7.

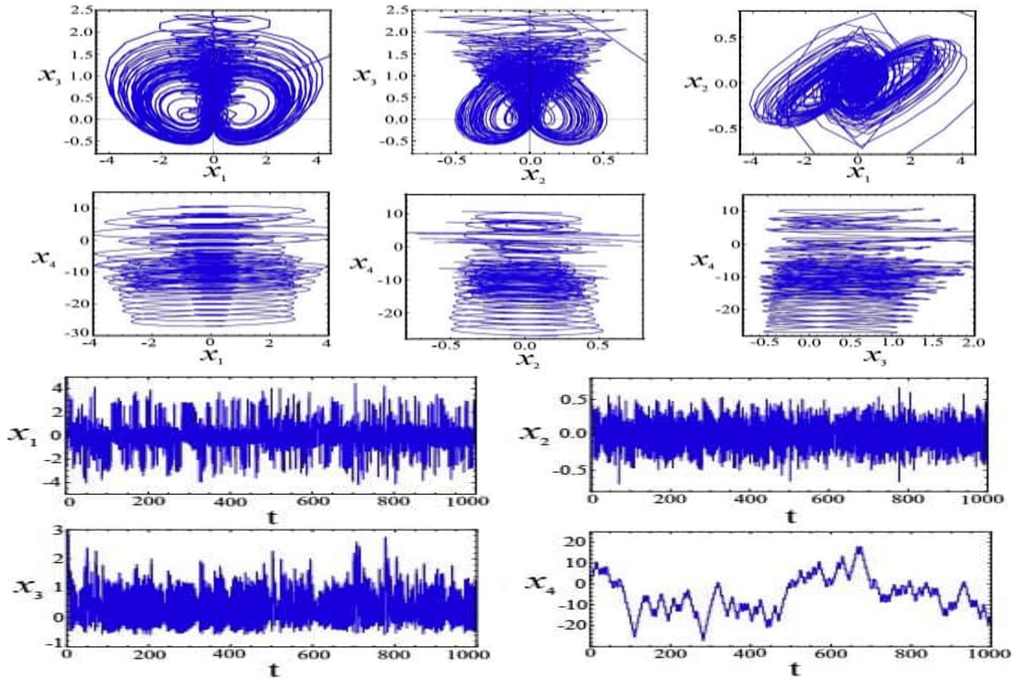


Figure 5: The upper part of the figure displays the hidden attractors of the hyperchaotic system (3) in various planes. In contrast, the lower part of the figure presents the time diagrams for the variables x_1, x_2, x_3, x_4.

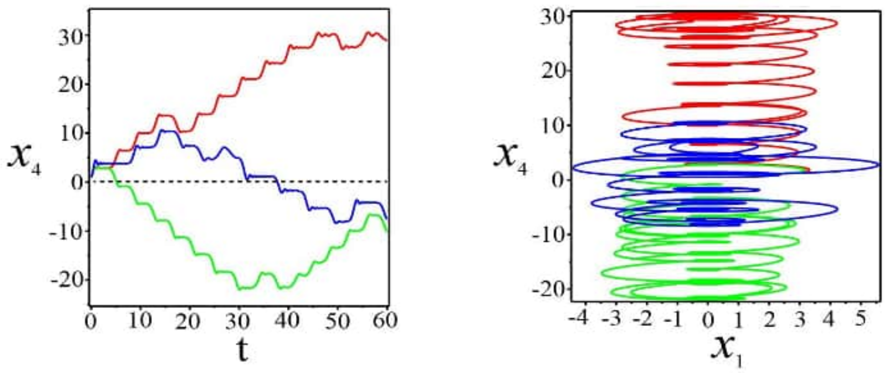


Figure 6: Signal x_4 and phase portrait in the plane x_1x_4 for different values of the offset boosting controller m: $m = 0$ (blue), $m = 30$ (green), $m = -30$ (red).

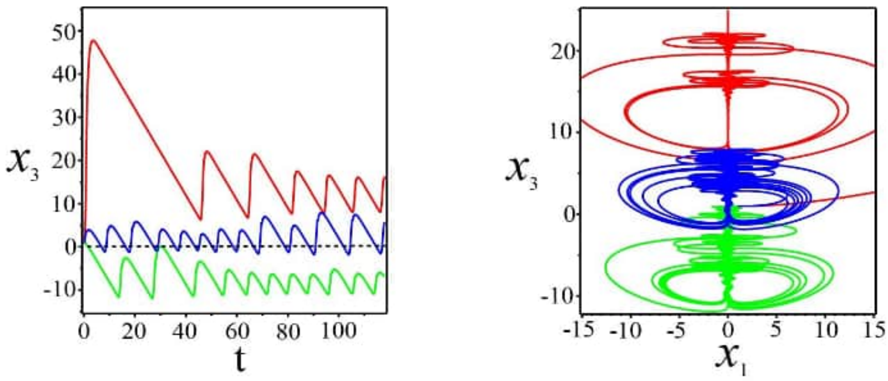


Figure 7: Signal x_3 and phase portrait in the plane x_1x_3 for different values of the offset boosting controller k: $k = 0$ (blue), $k = 10$ (green), $k = -10$ (red).

3 Electronic Circuit Design and Multisim Simulation of the New Hyperchaotic System

Based on Kirchhoff's law for electrical circuits, the electrical analog of the system (6) can be expressed as follows:

$$\begin{cases} \dfrac{d\widetilde{X}_1}{dt} = \dfrac{100\text{k}}{R_2}\widetilde{X}_2 - \left(\dfrac{100\text{k}}{R_1} + \dfrac{100\text{k}}{R_3 \cdot 10}|\widetilde{X}_4|\right)\widetilde{X}_1, \\ \dfrac{d\widetilde{X}_2}{dt} = -\dfrac{100\text{k}}{R_4 \cdot 10}\widetilde{X}_3\text{sgn}(\widetilde{X}_1), \\ \dfrac{d\widetilde{X}_3}{dt} = \dfrac{100\text{k}}{R_5}|\widetilde{X}_1| - \dfrac{100\text{k}}{R_6}V_b, \\ \dfrac{d\widetilde{X}_4}{dt} = \dfrac{100\text{k}}{R_7}\widetilde{X}_1, \end{cases} \tag{7}$$

where $R_1 = R_5 = R_6 = 100\text{k}\Omega$, $R_2 = 4.76\text{k}\Omega$, $R_3 = 20\text{k}\Omega$, $R_4 = 10\text{k}\Omega$, $R_7 = 2\text{M}\Omega$. The analog circuit modules for the equations in system (7) are shown in Figure 8. These circuits utilize standard components such as resistors (R), capacitors (C), diodes D1-D2 (1N4001), multipliers M1-M2 (AD633), operational amplifiers A1-A21 (TL084ACN), and a supply voltage of ±15V. The constant 1 is provided by a voltage source $V_b = 1V$. Figures 8b and 8c illustrate modules that model the signum sgn$(\cdot)$ and absolute value $|\cdot|$ functions. The phase portraits in Figure 10 reveal a remarkable similarity between the Mathematica simulation results (Figure 5) and the Multisim simulation results.

4 Conclusions

In this paper, we obtained a new 4D dynamical system based on a memristor that meets the known criteria for generating hyperchaos: a) it is dissipative; b) it has a four-dimensional phase space; and c) it includes at least one nonlinear term. A new memristive four-dimensional dynamic system, derived from the five-term Lorenz equations, contains only 7 terms. This hyperchaotic system lacks equilibrium points, potentially leading to

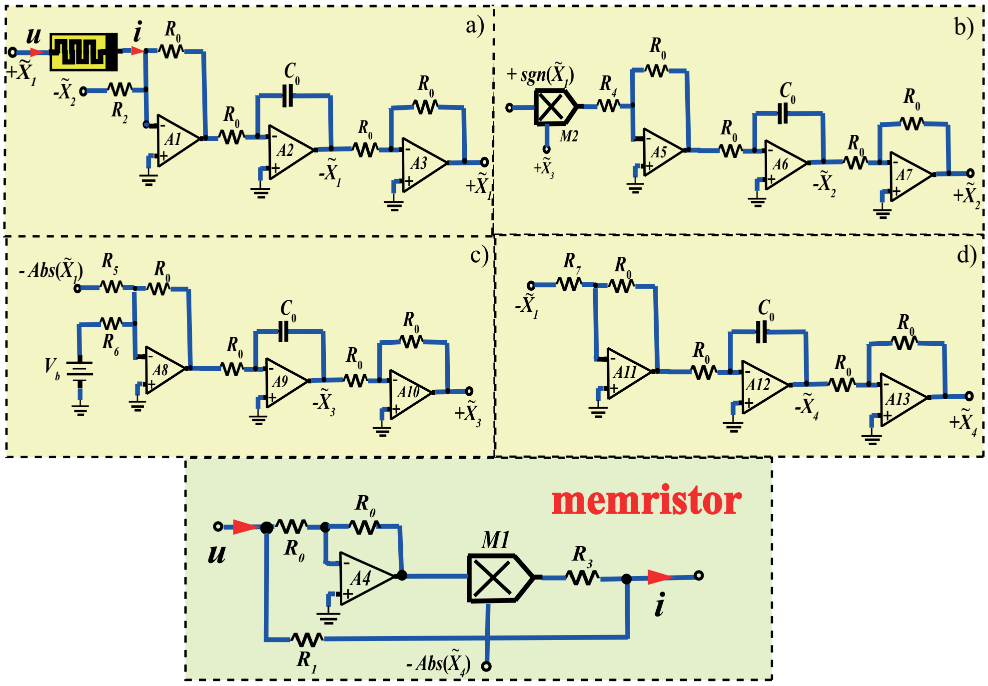


Figure 8: Circuit modules implemented based on the system of equations (5): a) $\widetilde{X}_1$, b) $\widetilde{X}_2$, c) $\widetilde{X}_3$, d) $\widetilde{X}_4$ and the memristor circuit module.

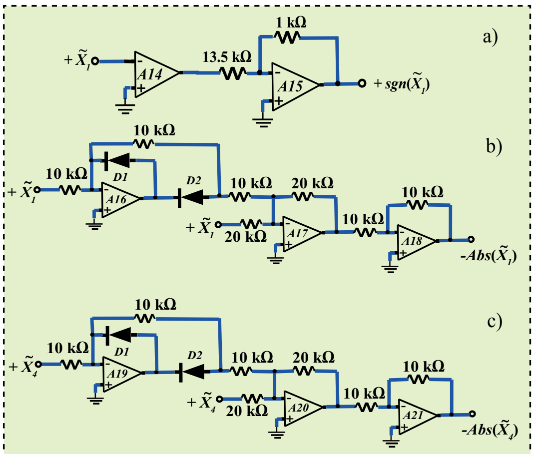


Figure 9: Schematic diagrams for the implementation of functions: a) $\mathrm{sgn}(\widetilde{X}_1)$; b)-c) $|\widetilde{X}_1|$, $|\widetilde{X}_4|$.

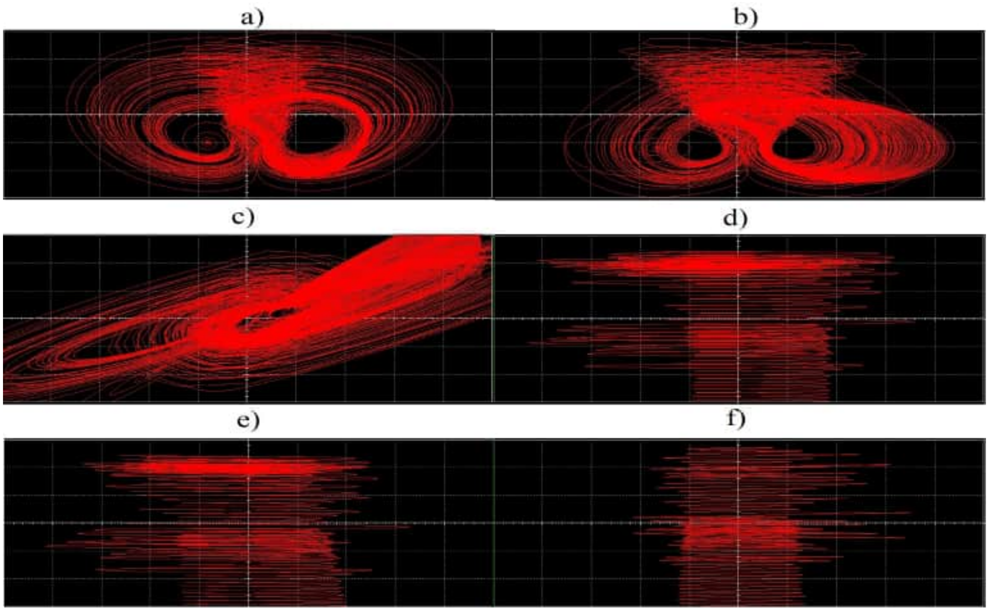


Figure 10: Phase portraits of the new 4D hyperchaotic system as generated in Multisim oscilloscopes: a) $\widetilde{X}_1\widetilde{X}_3$, b) $\widetilde{X}_2\widetilde{X}_3$, c) $\widetilde{X}_1\widetilde{X}_2$, d) $\widetilde{X}_1\widetilde{X}_4$, e) $\widetilde{X}_2\widetilde{X}_4$, f) $\widetilde{X}_3\widetilde{X}_4$.

hidden attractors. With two positive Lyapunov exponents, it is classified as hyperchaotic, and its Kaplan-Yorke dimension ($D_{KY} = 3.3919$) highlights its complexity. Simulation results from the electronic circuit of the proposed 4D system, designed in Multisim 14, align well with those obtained in Mathematica.

The new system shows great potential for applications in encrypting and decrypting information signals, images, data for the Internet of Things, and similar areas.

Acknowledgment

The author is grateful to the editor and reviewers for their interest in the paper and valuable comments.

References

[1] L. O. Chua. Memristor – the missing circuit element. *IEEE Transactions on Circuit Theory* **18** (5) (1971) 507–519.

[2] D. B. Strukov, G. S. Snider, D. R. Stewart, and R. S. Williams. The missing memristor found. *Nature* **453** (2008) 80–83.

[3] O. Krestinskaya, A. James, L. O. Chua. Neuro-memristive circuits for edge computing: a review. *IEEE Transactions on Neural Networks and Learning Systems* **31** (2020) 4–23.

[4] A. G. Radwan, M. E. Fouda. *On the Mathematical Modeling of Memristor, Memcapacitor and Meminductor*. Springer Cham, 2015. https://doi.org/10.1007/978-3-319-17491-4

[5] G. A. Leonov, N. V. Kuznetsov and V. I. Vagaitsev. Localization of hidden Chia's attractors. *Physics Letters A* **375** (23) (2011) 2230–2233.

[6] S. F. Al-Azzawi, Anmar M. Hasan. A New 4D Hidden Hyperchaotic System with Higher Largest Lyapunov exponent and its Synchronization. *International Journal of Mathematics, Statistics and Computer Science* **2** (2024) 63–74.

[7] A. Sambas, S. Vaidyanathan, S. F. Al-Azzawi, M. K. M. Nawawi, M. A. Mohamed, Z. A. Zakaria, S. S. Abas and M. Mamat. Modelling and MultiSim Simulation of a New Hyperchaos System with No Equilibrium Point. *Nonlinear Dynamics and Systems Theory* **23** (4) (2023) 422–433.

[8] V. T. Pham, C. Volos, L. V. Gambuzza. A memristive hyperchaotic system without equilibrium. *Scientific World Journal* **2014** (2014) 368986.

[9] K. Rajagopal, A. Bayani, A. J. M. Khalaf, H. Namazi, S. Jafari and V. T. Pham. A no-equilibrium memristive system with four-wing hyperchaotic attractor. *International Journal of Electronics and Communications (AEU)* **95** (2018) 207–215.

[10] P. P. Singh, B. K. Roy, C. Volos. Chapter 9 – Memristor-based novel 4D chaotic system without equilibria. Editors: Christos Volos, Viet-Thanh Pham. In: Advances in Nonlinear Dynamics and Chaos (ANDC), Mem-elements for Neuromorphic Circuits with Artificial Intelligence Applications. Academic Press, 2021, Pages 183–205. https://doi.org/10.1016/B978-0-12-821184-7.00017-7.

[11] V. K. Tamba, F. K. Tagne, A. L. M. Biamou, M. C. Nkeing and A. N. Takougang. Chapter 7 – Hidden extreme multistability generated from a novel memristive two-scroll chaotic system. Editors: Christos Volos, Viet-Thanh Pham. In: Advances in Nonlinear Dynamics and Chaos (ANDC), Mem-elements for Neuromorphic Circuits with Artificial Intelligence Applications, Academic Press, 2021, Pages 147–164. https://doi.org/10.1016/B978-0-12-821184-7.00015-3.

[12] A. Irfan. Simple chaotic jerk, hyperjerk, and hyperchaotic hyperjerk circuits and systems: families of self-excited and hidden attractors. Thesis PhD. 2020. https://doi.org/10.14457/TU.the.2020.75.

[13] A. S. Elwakil, S. Ozoguz and M. P. Kennedy. Creation of a complex butterfly attractor using a novel Lorenz-Type system. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications* **49** (4) (2002) 527–530.

[14] H. Binous and N. Zakia. An Improved Method for Lyapunov Exponents Computation. https://library.wolfram.com/infocenter/MathSource/7109/, March 09, 2008.

[15] X. Liu, X. Y. Zhang, P. Li and H. Yan. A new Lorentz chaotic system and its circuit implementation. Proceedings of the 8th EAI International Conference on Green Energy and Networking, GreeNets 2021, June 6–7, 2021, Dalian, People's Republic of China, 2021, https://api.semanticscholar.org/CorpusID:239146374.

Nonlinear Dynamics and Systems Theory, **25** (3) (2025) 299–314

Exponential Decay of Timoshenko System with Fractional Delays and Source Terms

C. Messikh [1*] N. Bellal [2] S. Labidi [1] and Kh. Zennir [3]

[1] *Department of Mathematics, Applied Mathematics Laboratory, Badji Mokhtar University, B.O. Box 12, El Hadjar, 23000, Annaba, Algeria.*
[2] *Department Mathematics, Numerical Analysis, Optimisation and Statistics Laboratory, Badji Mokhtar University, B.O. Box 12, El Hadjar, 23000, Annaba, Algeria.*
[3] *Department of Mathematics, College of Science, Qassim University, Saudi Arabia.*

Received: June 27, 2024; Revised: May 29, 2025

Abstract: The objective of this paper is to analyse the asymptotic behavior of a Timoshenko beam system with fractional delays and nonlinear external sources. Under suitable conditions on the damping, delay and initial data, we obtain exponential decay rate of the solution.

Keywords: *Timoshenko system; energy decay; nonlinear systems; fractional delay.*

Mathematics Subject Classification (2020): 35B40, 47D03, 74D05, 93D23, 93D15.

1 Introduction

In this work, we study the following Timoshenko system with fractional delays:

$$\begin{cases} \rho_1\varphi_{tt} - k\left(\varphi_x + \psi\right)_x + a_1\partial_t^{\alpha,\beta}\varphi\left(t-s\right) + \alpha_1\varphi_t = \left|\varphi\right|^{p-2}\varphi, \\ \rho_2\psi_{tt} - b\psi_{xx} + k\left(\varphi_x + \psi\right) + a_2\partial_t^{\alpha,\beta}\psi\left(t-s\right) + \alpha_2\psi_t = \left|\psi\right|^{q-2}\psi, \\ \varphi(x=0,t) = \psi(x=0,t) = \varphi(x=L,t) = \psi(x=L,t) = 0, \\ \varphi(x,t=0) = \varphi_0(x),\ \psi(x,t=0) = \psi_0(x), \\ \varphi_t(x,t=0) = \varphi_1(x),\ \psi_t\left(x,0\right) = \psi_1(x), \\ \varphi_t\left(x,t-s\right) = f_0\left(x,t-s\right), t\in\left(0,s\right), \\ \psi_t\left(x,t-s\right) = g_0\left(x,t-s\right), t\in\left(0,s\right), \end{cases} \tag{1}$$

where $x \in \Omega = (0,L), L > 0, t \in \mathbb{R}_+^\star, \rho_1, \rho_2, a_1, a_2, \alpha_1, \alpha_2, b$ and k are positive real constants. The constant $s > 0$ is the time delay and the exponents p and q satisfy $p > 2$

* Corresponding author: mailto:chahrazed.messikh@univ-annaba.dz

and $q > 2$. The functions $\varphi_0, \varphi_1, \psi_0, \psi_1, f_0, g_0$ are the initial data belonging to suitable spaces. The well known notation $\partial_t^{\alpha,\beta}$ stands for the generalized Caputo's fractional derivative, see [17, 18], it is defined as

$$\partial_t^{\alpha,\beta} u(t) = \frac{1}{\Gamma(1-\alpha)} \int_0^t (t-s)^{-\alpha} \ e^{-\beta(t-s)} \ u_s(s) \ ds, \quad 0 < \alpha < 1, \ \beta > 0.$$

The problem (1) is considered without internal and external forces, this type of systems has been introduced in [19]. It describes the transverse vibration of a thick beam of length L, where φ is the transverse displacement of the beam, ψ is the rotation angle of the filament of the beam, and ρ_1, ρ_2, k and b account for some physical properties of the beam, see [11]. In our case, the Timoshenko beam is subject to internal forces given by fractional delay terms and frictional damping, and to external forces represented by source terms. Physically, the occurrence of fractional delay terms in many systems can lead to undesirable dynamics such as degraded performance, reduced robustness, or even instability. Generally, these harmful effects are controlled by various dissipation terms; for more results, see [1, 2].

In the last decades, the study of the well-posedness and stability/instability of evolution equations with time delay has received considerable attention of researchers. Many authors have shown that the time delay can be a source of instability that is asymptotically stable in the absence of time delay, see in this direction [3, 15]. More results in this context can be found in [4, 5, 8, 10, 20].

For the Timoshenko system with time delay, we mention the work [7], in which the following problem is considered:

$$\begin{cases} \rho_1 \varphi_{tt}(x,t) - k(\varphi_x + \psi)_x(x,t) + a_1 \varphi(x, t-\tau_1) + \alpha_1 \varphi_t(x,t) = 0, \\ \rho_2 \psi_{tt}(x,t) - b\psi_{xx}(x,t) + k(\varphi_x + \psi)(x,t) + a_2 \psi(x, t-\tau_2) + \alpha_2 \psi_t(x,t) = 0. \end{cases} \tag{2}$$

The authors obtained the exponential decay rate when the weights of time delays are smaller than the corresponding damping. By adopting the spectral analysis approach, A. Adnane et al. [1] showed the same result by considering the time delay of fractional type rather than the time delay in the system (2) without sources.

In the absence of delay, the problem of existence and energy decay for a single wave equation with damping and/or source terms has been extensively studied by several authors. They showed the damping term assures global existence in the absence of source term, whereas without the damping term, the source term causes finite time blow-up of the solution. Hence, it is valuable to study the asymptotic behavior of a single wave equation with terms having opposite effects, see [6, 12, 13]. For more results about systems with various other damping and source terms, we refer the reader to [9, 14, 16].

The purpose of this paper is to analyse the influence of the damping terms, delay terms and source terms on the solutions to (1). Under suitable assumptions, we establish local existence, global existence and asymptotic behavior of solutions to (1). As far as we know, this type of problems has never been considered before in the literature.

This paper is structured as follows. In Section 2, we state some assumptions, the augmented problem (8), and lemmas for this analysis. Section 3 is devoted to the proof of the local and global existence results by using the semi-group approach. In Section 4, we state and prove the exponential decay rate result by using the multiplier method and appropriate Lyapunov functional.

2 Preliminaries and Tools

Here, we shall reformulate the initial problem (1) into the augmented system (8). To this end, we need the following results.

Lemma 2.1 *(see [2], p. 286) Let ϖ be a function defined for $\alpha \in (0,1)$ as*

$$\varpi(\nu) = |\nu|^{\frac{2\alpha-1}{2}}, \quad \nu \in \mathbb{R}.$$

Then the relationship between the "input" U and "output" O of the system

$$\begin{cases} \phi_t(x,\nu,t) + \left(\nu^2+\beta\right)\phi(x,\nu,t) - U(x,t)\varpi(\nu) = 0, \\ \phi(x,\nu,t=0) = 0, \\ O(t) = (\pi)^{-1}\sin(\alpha\pi)\int_{-\infty}^{+\infty}\phi(x,\nu,t)\varpi(\nu)\,d\nu, \end{cases} \tag{3}$$

where $\nu \in \mathbb{R},\ t>0,\ \beta>0,$ is given by

$$O = I^{1-\alpha,\beta}U,$$

here,

$$I^{\alpha,\beta}w(t) = \frac{1}{\Gamma(\alpha)}\int_0^t (t-\tau)^{\alpha-1}\ e^{-\beta(t-\tau)}w(\tau)\ d\tau.$$

Lemma 2.2 *([2], p. 286) If*

$$\lambda \in D_\beta = \mathbb{C}\setminus\,]-\infty,-\beta[\,,$$

then

$$\int_{-\infty}^{+\infty}\frac{\varpi^2(\nu)}{\lambda+\beta+\nu^2}\ d\nu = \frac{\pi}{\sin(\alpha\pi)}(\lambda+\beta)^{\alpha-1}.$$

The constants a_i, α_i are supposed to satisfy

$$a_i\beta^{\alpha-1} < \alpha_i \quad \text{for } i=1,2. \tag{4}$$

As in ([1], p. 1063), we can introduce new variables

$$z_1(x,\rho,t) = \varphi_t(x,t-s\rho),\ \rho\in(0,1),\ t>0, \tag{5}$$

$$z_2(x,\rho,t) = \psi_t(x,t-s\rho),\ \rho\in(0,1),\ t>0. \tag{6}$$

Then

$$z_{it}(x,\rho,t) = \frac{-1}{s}z_{i\rho}(x,\rho,t),\ \rho\in(0,1),\ t>0, \tag{7}$$

with $i=1,2$. For $\nu\in\mathbb{R},\ \rho\in(0,1)$, we denote $z_{it} = \frac{\partial}{\partial t}(z_i)$ and $z_{i\rho} = \frac{\partial}{\partial\rho}(z_i)$, then by (7) and Lemma 2.1, the initial system (1) is equivalent to

$$\begin{cases} \rho_1\varphi_{tt} - k(\varphi_x+\psi)_x + b_1\phi_1*\varpi + \alpha_1\varphi_t = |\varphi|^{p-2}\varphi, \\ \rho_2\psi_{tt} - b\psi_{xx} + k(\varphi_x+\psi) + b_2\phi_2*\varpi + \alpha_2\psi_t = |\psi|^{p-2}\psi, \\ \phi_{1t}(x,\nu,t) + \left(\nu^2+\beta\right)\phi_1(x,\nu,t) - z_1(x,1,t)\varpi(\nu) = 0, \\ sz_{1t}(x,\rho,t) + z_{1\rho}(x,\rho,t) = 0, \\ \phi_{2t}(x,\nu,t) + \left(\nu^2+\beta\right)\phi_2(x,\nu,t) - z_2(x,1,t)\varpi(\nu) = 0, \\ sz_{2t}(x,\rho,t) + z_{2\rho}(x,\rho,t) = 0, \\ \varphi(x=L,t) = \varphi(x=0,t) = \psi(x=L,t) = \psi(x=0,t) = 0, \\ z_1(x,\rho=0,t) = \varphi_t(x,t), z_2(x,\rho=0,t) = \psi_t(x,t), \\ \varphi(x,t=0) = \varphi_0,\ \varphi_t(x,t=0) = \varphi_1, \\ \psi(x,t=0) = \psi_0,\ \psi_t(x,t=0) = \psi_1, \\ z_1(x,\rho,0) = f_0(x,-s\rho), z_2(x,\rho,0) = g_0(x,-s\rho), \\ \phi_1(x,\nu,t=0) = \phi_2(x,\nu,t=0) = 0, \end{cases} \tag{8}$$

where $\nu \in \mathbb{R}$ and

$$\phi_i * \varpi = \int_{-\infty}^{+\infty} \phi_i(x,\nu,t)\,\varpi(\nu)\ d\nu,$$

and

$$b_i = (\pi)^{-1}\sin(\alpha\pi)\,a_i, \quad i=1,2.$$

To prove the dissipativity of the energy $\mathcal{E}$, we need the following lemma.

Lemma 2.3 *(See [2], p. 286) For $z \in L^2(\Omega)$ and $\nu\phi \in L^2(\Omega\times(-\infty,+\infty))$, we have*

$$\begin{aligned}\left|\int_\Omega z(x,\rho,t)\int_{-\infty}^{+\infty}\varpi(\nu)\,\phi(x,\nu,t)\ d\nu\ dx\right| &\le A_0\int_\Omega |z(x,\rho,t)|^2\ dx \\ &+\frac{1}{4}\int_\Omega\int_{-\infty}^{+\infty}\left(\nu^2+\beta\right)|\phi(x,\nu,t)|^2\ d\nu\ \ dx,\end{aligned}$$

where

$$A_0 = \int_{-\infty}^{+\infty}\frac{\varpi^2(\nu)}{\nu^2+\beta}\ d\nu.$$

The energy associated to (8) is defined by

$$\begin{aligned}&\mathcal{E}(t) = \tfrac{1}{2}\left[\rho_1\|\varphi_t\|^2 + k\|\varphi_x+\psi\|^2 + \rho_2\|\psi_t\|^2 + b\|\psi_x\|^2\right] \\ &+\sum_{i=1}^{2}\tfrac{b_i}{2}\int_0^L\int_{-\infty}^{+\infty}|\phi_i(x,\nu,t)|^2\ d\nu\ dx + \sum_{i=1}^{2} v_i s\int_0^L\int_0^1|z_i(x,\rho,t)|^2\ d\rho\ dx \\ &-\tfrac{1}{p}\|\varphi\|^p - \tfrac{1}{q}\|\psi\|^q,\end{aligned} \tag{9}$$

where v_i satisfies

$$A_0 b_i < v_i < \alpha_i - b_i A_0, \quad i=1,2. \tag{10}$$

Lemma 2.4 *Let (4) hold. Then the energy (9) satisfies*

$$\begin{aligned}&\tfrac{d\mathcal{E}(t)}{dt} \le -C\sum_{i=1}^{2}\int_\Omega\left(|z_i(x,1,t)|^2 + |z_i(x,0,t)|^2\right)\ dx \\ &-\sum_{i=1}^{2}\tfrac{b_i}{2}\int_0^L\int_{-\infty}^{+\infty}\left(\nu^2+\beta\right)|\phi_i(x,\nu,t)|^2\ d\nu\ dx \le 0\end{aligned} \tag{11}$$

for $C>0$ and $b_i = (\pi)^{-1}\sin(\alpha\pi)\,a_i, i=1,2$.

Proof. By multiplying $(8)_1$ by φ_t and integrating over $(0,L)$, integrating by parts and using the boundary conditions, we find

$$\begin{aligned}&\tfrac{d}{dt}\left[\tfrac{\rho_1}{2}\|\varphi_t\|^2 - \tfrac{1}{p}\|\varphi\|^p\right] + k\int_0^L(\varphi_x+\psi)\,\varphi_{xt}\ dx + \alpha_1\|\varphi_t\|^2 \\ &+b_1\int_0^L\left(\int_{-\infty}^{+\infty}\phi_1(x,\nu,t)\,\varpi(s)\ d\nu\right)\varphi_t\ dx = 0.\end{aligned} \tag{12}$$

Multiplying $(8)_2$ by ψ_t and integrating over $(0,L)$, we have

$$\begin{aligned}&\tfrac{d}{dt}\left[\tfrac{\rho_2}{2}\|\psi_t\|^2 + \tfrac{b}{2}\|\psi_x\|^2 - \tfrac{1}{q}\|\psi\|^q\right] + \alpha_2\|\psi_t\|^2 + k\int_0^L(\varphi_x+\psi)\,\psi_t\ dx \\ &+b_2\int_0^L\left(\int_{-\infty}^{+\infty}\phi_2(x,\nu,t)\,\varpi(\nu)\ d\nu\right)\psi_t\ dx = 0.\end{aligned} \tag{13}$$

Multiplying $(8)_j$ by $b_i\phi_i$ with $(i,j)=(1,3)$, respectively $(i,j)=(2,5)$, and integrating over $(0,L)\times\mathbb{R}$, we obtain

$$\begin{aligned} &b_i\int_0^L\int_{-\infty}^{+\infty}\left(\frac{d}{2dt}\left|\phi_i\left(x,\nu,t\right)\right|^2+\left(\nu^2+\beta\right)\left|\phi_i\left(x,\nu,t\right)\right|^2\right)\,d\nu\,dx \\ &-b_i\int_0^L z_i\left(x,1,t\right)\int_{-\infty}^{+\infty}\varpi\left(\nu\right)\phi_i\left(x,\nu,t\right)\,d\nu\,dx=0. \end{aligned}\tag{14}$$

Multiplying $(8)_j$ by $2v_iz_i$ with $(i,j)=(1,4)$, respectively $(i,j)=(2,6)$, and integrating over $(0,L)\times(0,1)$, we have

$$\begin{aligned} &\frac{d}{dt}\left\{sv_i\int_0^L\int_0^1\left|z_i\left(x,\rho,t\right)\right|^2\,d\rho\,dx\right\} \\ &+v_i\int_0^L\left[\left|z_i\left(x,1,t\right)\right|^2-\left|z_i\left(x,0,t\right)\right|^2\right]\,dx=0. \end{aligned}\tag{15}$$

Summing (12), (13), (14) and (15) and due to the fact that $\varphi_t\left(x,t\right)=z_1\left(x,0,t\right)$, $\psi_t\left(x,t\right)=z_2\left(x,0,t\right)$, we have

$$\begin{aligned} \frac{d\mathcal{E}(t)}{dt}&=-\sum_{i=1}^{2}\left(\alpha_i-v_i\right)\int_0^L\left|z_i\left(x,0,t\right)\right|^2\,dt \\ &-\sum_{i=1}^{2}b_i\int_0^L z_i\left(x,0,t\right)\int_{-\infty}^{+\infty}\phi_i\left(x,\nu,t\right)\varpi\left(\nu\right)\,d\nu\,dx \\ &-\sum_{i=1}^{2}b_i\int_0^L\int_{-\infty}^{+\infty}\left(\nu^2+\beta\right)\left|\phi_i\left(\nu\right)\right|^2\,d\nu\,dx \\ &+\sum_{i=1}^{2}b_i\int_0^L z_i\left(x,1,t\right)\int_{-\infty}^{+\infty}\phi_i\left(x,\nu,t\right)\varpi\left(\nu\right)\,d\nu\,dx \\ &-\sum_{i=1}^{2}v_i\int_0^L\left|z_i\left(x,1,t\right)\right|^2\,dx. \end{aligned}$$

Thanks to Lemma 2.2 and putting $C=\min_{i=1,2}\left(v_i-A_0b_i,\alpha_i-v_i-b_iA_0\right)>0,\ i=1,2,$ the estimate (11) is established.

3 Unique Local and Global Weak Solution

Set $u=\varphi_t$ and $v=\psi_t$ and denote $U=\left(\varphi,u,\psi,v,\phi_1,\phi_2,z_1,z_2\right)^T$, then (8) takes the abstract form

$$\left\{\begin{array}{l} U_t(t)=AU(t)+\mathbb{F}\left(U(t)\right), \\ U_0=\left(\varphi_0,\varphi_1,\psi_0,\psi_1,0,0,f_0\left(-\rho s\right),g_0\left(-\rho s\right)\right)^T,\quad \text{for } \rho\in(0,1), \end{array}\right.\tag{16}$$

where the operator A is defined by

$$\begin{aligned} AU=\Bigg(&u,\frac{k}{\rho_1}\left(\varphi_x+\psi\right)_x-\frac{b_1}{\rho_1}\phi_1\star\varpi-\frac{\alpha_1}{\rho_1}u,v,\frac{b}{\rho_2}\psi_{xx}-\frac{k}{\rho_2}\left(\varphi_x+\psi\right)-\frac{b_2}{\rho_2}\phi_2\star\varpi-\frac{\alpha_2}{\rho_2}v, \\ &-\left(\nu^2+\beta\right)\phi_1+z_1\left(x,1\right)\varpi\left(\nu\right),-\left(\nu^2+\beta\right)\phi_2+z_2\left(x,1\right)\varpi\left(\nu\right),-\frac{1}{s}z_{1\rho}\left(x,\rho\right),-\frac{1}{s}z_{2\rho}\left(x,\rho\right)\Bigg)^T, \end{aligned}$$

where

$$\phi_i\star\varpi=\int_{-\infty}^{+\infty}\phi_i\left(x,\nu\right)\varpi\left(\nu\right)\,d\nu,\quad i=1,2,$$

for $i=1,2$, the domain is given by

$$D\left(A\right)=\left\{\begin{array}{l} U\in\mathcal{H}\ :\ \left(\varphi,\psi\right)\in\left(H^{2}\left(\Omega\right)\right)^{2},\ \left(u,v\right)\in\left(H_{0}^{1}\left(\Omega\right)\right)^{2},\\ z_{i}\in L^{2}\left(\Omega\times H^{1}\left(0,1\right)\right)\quad\text{for } i=1,2,\quad u=z_{1}\left(.,0\right),v=z_{2}\left(.,0\right),\\ \nu\phi_{i}\in L^{2}\left(\Omega\times\left(-\infty,+\infty\right)\right)\quad\text{for } i=1,2,\\ \left(\nu^{2}+\beta\right)\phi_{i}-z_{i}\left(x,1\right)\varpi\left(\nu\right)\in L^{2}\left(\Omega\times\left(-\infty,+\infty\right)\right), \end{array}\right\}$$

where $\mathcal{H}$ is given as

$$\mathcal{H}=\left(H_{0}^{1}\left(\Omega\right)\times L^{2}\left(\Omega\right)\right)^{2}\times\left(L^{2}\left(\Omega\times\left(-\infty,+\infty\right)\right)\right)^{2}\times\left(L^{2}\left(\Omega\times\left(0,1\right)\right)\right)^{2}$$

and equipped with the inner product

$$\begin{aligned}\left\langle U,\bar{U}\right\rangle_{\mathcal{H}}&=k\int_{\Omega}\left(\varphi_{x}+\psi\right)\left(\bar{\varphi}_{x}+\bar{\psi}\right)\ dx+b\int_{\Omega}\psi_{x}\bar{\psi}_{x}\ dx+\rho_{1}\int_{\Omega}u\bar{u}+\rho_{2}\int_{\Omega}v\bar{v}\ dx\\ &+\sum_{i=1}^{2}b_{i}\int_{\Omega}\int_{-\infty}^{+\infty}\phi_{i}\left(x,\nu\right)\bar{\phi}_{i}\left(x,\nu\right)\ d\nu\ dx+2\sum_{i=1}^{2}v_{i}s\int_{\Omega}\int_{0}^{1}z_{i}\left(x,\rho\right)\bar{z}_{i}\left(x,\rho\right)\ d\rho\ dx,\end{aligned}$$

for all $\bar{U}=\left(\bar{\varphi},\bar{u},\bar{\psi},\bar{v},\bar{\phi}_{1},\bar{\phi}_{2},\bar{z}_{1},\bar{z}_{2}\right).$

Theorem 3.1 *(Unique local weak solution) Assume that $p>2$ and $q>2$. Let (10) hold. Then, for any $U_0\in\mathcal{H}$, the system (16) has a unique local weak solution*

$$U\in\mathrm{C}\left(\left[0,T\right],\mathcal{H}\right).$$

Moreover, if $U_0\in D\left(A\right)$, then

$$U\in\mathrm{C}\left(\left[0,T\right],D\left(A\right)\right)\cap\mathrm{C}^{1}\left(\left[0,T\right],\mathcal{H}\right).$$

Proof. It will be proved that A is a maximal dissipative operator. We have

$$\begin{aligned}\frac{d\mathcal{E}(t)}{dt}&=\frac{1}{2}\frac{d}{dt}\|U\|^{2}=\langle AU,U\rangle\leq-C\sum_{i=1}^{2}\int_{\Omega}\left|z_{i}\left(x,1,t\right)\right|^{2}\ dx-C\sum_{i=1}^{2}\int_{\Omega}\left|z_{i}\left(x,0,t\right)\right|^{2}\ dx\\ &-\sum_{i=1}^{2}\frac{b_{i}}{2}\int_{0}^{L}\int_{-\infty}^{+\infty}\left(\nu^{2}+\beta\right)\left|\phi_{i}\left(x,\nu,t\right)\right|^{2}\ d\nu\ dx\leq0,\end{aligned}$$

therefore A is dissipative.

Now, it will be shown that $I-A$ is surjective. Indeed, let $F=\left(f_{1},f_{2},f_{3},f_{4},f_{5},f_{6},f_{7},f_{8}\right)^{T}\in\mathcal{H}$, and look for $U\in D\left(A\right)$ such that $\left(I-A\right)U=F$. This is equivalent to

$$\left\{\begin{array}{l}\varphi-u=f_{1}\left(x\right),\\ \left(1+\frac{\alpha_{1}}{\rho_{1}}\right)u-\frac{k}{\rho_{1}}\left(\varphi_{x}+\psi\right)_{x}+\frac{b_{1}}{\rho_{1}}\phi_{1}\star\varpi=f_{2}\left(x\right),\\ \psi-v=f_{3}\left(x\right),\\ \left(1+\frac{\alpha_{2}}{\rho_{2}}\right)v-\frac{b}{\rho_{2}}\psi_{xx}+\frac{k}{\rho_{2}}\left(\varphi_{x}+\psi\right)+\frac{b_{2}}{\rho_{2}}\phi_{2}\star\varpi=f_{4}\left(x\right),\\ \left(1+\nu^{2}+\beta\right)\phi_{1}-z_{1}\left(x,1\right)\varpi\left(\nu\right)=f_{5}\left(x,\nu\right),\\ \left(1+\nu^{2}+\beta\right)\phi_{2}-z_{2}\left(x,1\right)\varpi\left(\nu\right)=f_{6}\left(x,\nu\right),\\ z_{1}+\ \frac{1}{s}z_{1\rho}=f_{7}\left(x,\rho\right)\quad\rho\in\left(0,1\right),\\ z_{2}+\frac{1}{s}z_{2\rho}=f_{8}\left(x,\rho\right)\quad\rho\in\left(0,1\right).\end{array}\right.\tag{17}$$

Suppose $\left(\varphi,\psi\right)\in\left(H_{0}^{1}\left(\Omega\right)\right)^{2}$, then by $(17)_1$ and $(17)_3$, we obtain

$$u=\varphi-f_{1}\in H_{0}^{1}\left(\Omega\right),\tag{18}$$

$$v = \psi - f_3 \in H_0^1(\Omega), \tag{19}$$

and from $(17)_{7,8}$, we get

$$z_1(x,\rho) = e^{-s\rho} z_1(x,0) + s e^{-s\rho} \int_0^\rho e^{\tau s} f_7(x,\tau)\ d\tau, \tag{20}$$

$$z_2(x,\rho) = e^{-s\rho} z_2(x,0) + s e^{-s\rho} \int_0^\rho e^{\tau s} f_8(x,\tau)\ d\tau. \tag{21}$$

Using (18) and (19), we have

$$u(x) = z_1(x,0) = \varphi - f_1(x), \tag{22}$$

$$v(x) = z_2(x,0) = \psi - f_3(x). \tag{23}$$

Substituting (22) and (23) respectively in (20) and (21), we get, for all $x \in (\Omega)$, $\rho \in (0,1)$,

$$\begin{aligned} z_1(x,\rho) &= e^{-s\rho}[\varphi - f_1(x)] + s e^{-s\rho} \textstyle\int_0^\rho e^{-s\tau} f_7(x,\tau)\ d\tau \ \in L^2(\Omega \times (0,1)), \\ z_2(x,\rho) &= e^{-s\rho}[\psi - f_3(x)] + s e^{-s\rho} \textstyle\int_0^\rho e^{s\tau} f_8(x,\tau)\ d\tau \in L^2(\Omega \times (0,1)). \end{aligned} \tag{24}$$

Returning back to $(17)_{7,8}$, we find that

$$z_{1\rho} = s f_7(x,\rho) - s z_1 \in L^2(\Omega \times (0,1)), z_{2\rho} = s f_8(x,\rho) - s z_2 \in L^2(\Omega \times (0,1)).$$

Using $(17)_5$ and $(17)_6$, we obtain

$$\phi_1 = \frac{f_5 + z_1(x,1)\varpi(\nu)}{1+\nu^2+\beta} \in L^2(\Omega \times (-\infty,+\infty)), \tag{25}$$

$$\phi_2 = \frac{f_6 + z_2(x,1)\varpi(\nu)}{1+\nu^2+\beta} \in L^2(\Omega, \times (-\infty,+\infty)). \tag{26}$$

Therefore

$$\begin{aligned} \nu\phi_1 &= \tfrac{\nu}{1+\nu^2+\beta}[f_5 + z_1(x,1)\varpi(\nu)] \in L^2((0,L) \times (-\infty,+\infty)), \\ \nu\phi_2 &= \tfrac{\nu}{1+\nu^2+\beta}[f_6 + z_2(x,1)\varpi(\nu)] \in L^2(\Omega \times (-\infty,+\infty)). \end{aligned}$$

Inserting $(17)_1$ and (25) in $(17)_2$, respectively $(17)_3$, and (26) in $(17)_4$, we have

$$\left\{ \begin{array}{l} \left(1+\frac{\alpha_1}{\rho_1}\right)\varphi - \frac{k}{\rho_1}(\varphi_x+\psi)_x = f_2 - \frac{b_1}{\rho_1}\left[\frac{f_5+z_1(x,1)\varpi(\nu)}{1+\nu^2+\beta}\right] \star \varpi + \left(1+\frac{\alpha_1}{\rho_1}\right) f_1, \\ \left(1+\frac{\alpha_2}{\rho_2}\right)\psi - \frac{b}{\rho_2}\psi_{xx} + \frac{k}{\rho_2}(\varphi_x+\psi) = f_4 - \frac{b_2}{\rho_2}\left[\frac{f_6+z_2(x,1)\varpi(\nu)}{1+\nu^2+\beta}\right] \star \varpi \\ + \left(1+\frac{\alpha_2}{\rho_2}\right) f_3. \end{array} \right. \tag{27}$$

By replacing (20) and (21) for $\rho = 1$ in (27), we get

$$\left\{ \begin{array}{l} \left(1+\frac{b_{11}e^{-s}}{\rho_1}+\frac{\alpha_1}{\rho_1}\right)\varphi - \frac{k}{\rho_1}(\varphi_x+\psi)_x = f_2 + \left(1+\frac{\alpha_1}{\rho_1}\right) f_1 \\ -\frac{b_1}{\rho_1}\int_{-\infty}^{+\infty} \frac{\varpi(\nu) f_5(\nu)}{1+\nu^2+\beta}\ d\nu, + \frac{b_1}{\rho_1} f_{1,7} \int_{-\infty}^{+\infty} \frac{\varpi^2(\nu)}{1+\nu^2+\beta} d\nu, \\ \left(1+\frac{b_{22}}{\rho_2}e^{-s}+\frac{\alpha_2}{\rho_2}\right)\psi - \frac{b}{\rho_2}\psi_{xx} + \frac{k}{\rho_2}(\varphi_x+\psi) = f_4 + \left(1+\frac{\alpha_2}{\rho_2}\right) f_3 \\ -\frac{b_2}{\rho_2}\int_{-\infty}^{+\infty} \frac{\varpi(\nu) f_6(\nu)}{1+\nu^2+\beta}\ d\nu + \frac{b_2}{\rho_2} f_{3,8} \int_{-\infty}^{+\infty} \frac{\varpi^2(\nu)}{1+\nu^2+\beta} d\nu, \end{array} \right. \tag{28}$$

where, for i=1, 2,

$$b_{ii}=b_i\int_{-\infty}^{+\infty}\frac{\varpi^2(\nu)}{1+\nu^2+\beta}d\nu,\quad f_{1,7}=f_1-se^{-s}\int_0^1 e^{\tau s}f_7\left(x,\tau\right)\,d\tau,$$

and

$$f_{3,8}=f_3-se^{-s}\int_0^1 e^{\tau s}f_8\left(x,\tau\right)\,d\tau.$$

Let $\left(\bar{\varphi},\bar{\psi}\right)\in\left(H_0^1\left((0,L)\right)\right)^2$. Multiply $(28)_1$ by $\rho_1\bar{\varphi}$ and $(28)_2$ by $\rho_2\bar{\psi}$. Integrating by parts, and summing the obtained result, we get

$$M\left(\varphi,\psi;\bar{\varphi},\bar{\psi}\right)=L\left(\bar{\varphi},\bar{\psi}\right),\tag{29}$$

here, the bilinear form

$$M:\left(H_0^1((0,L))\times H_0^1((0,L))\right)^2\to\mathbb{R}$$

is defined by

$$\begin{array}{l}M\left(\varphi,\psi;\bar{\varphi},\bar{\psi}\right)=\left(\rho_1+b_{11}e^{-s}+\alpha_1\right)\int_0^L\varphi\bar{\varphi}\,dx+\left(\rho_2+b_{22}e^{-s}+\alpha_2\right)\int_0^L\psi\bar{\psi}\,dx\\ +k\int_0^L\left(\varphi_x+\psi\right)\left(\bar{\varphi}_x+\bar{\psi}\right)\,dx+b\int_0^L\psi\bar{\psi}_x\,dx,\end{array}$$

and the linear form

$$L:\left(H_0^1((0,L))\right)^2\to\mathbb{R}$$

by

$$\begin{array}{l}L\left(\bar{\varphi},\bar{\psi}\right)=\rho_1\int_0^L f_2\bar{\varphi}\,dx+\left(\rho_1+\alpha_1\rho_1\right)\int_0^L f_1\bar{\varphi}dx-b_1\int_0^L\left\{\int_{-\infty}^{+\infty}\frac{\varpi(\nu)f_5(x,\nu)}{1+\nu^2+\beta}\,d\nu\right\}\bar{\varphi}dx\\ -b_2\int_0^L\left\{\int_{-\infty}^{+\infty}\frac{\varpi(\nu)f_6(x,\nu)}{1+\nu^2+\beta}\,d\nu\right\}\bar{\psi}\,dx+\rho_2\int_0^L f_4\bar{\psi}\,dx+\left(\rho_2+\alpha_2\right)\int_0^L f_3\bar{\psi}\,dx\\ +b_1\left(\int_{-\infty}^{+\infty}\frac{\varpi^2(\nu)}{1+\nu^2+\beta}d\nu\right)\int_0^L\left\{f_1-se^{-s}\int_0^1 e^{\tau s}f_7\left(x,\tau\right)\,d\tau\right\}\bar{\varphi}dx\\ +b_2\left(\int_{-\infty}^{+\infty}\frac{\varpi^2(\nu)}{1+\nu^2+\beta}d\nu\right)\int_0^L\left\{f_3-se^{-s}\int_0^1 e^{\tau s}f_8\left(x,\tau\right)\,d\tau\right\}\bar{\psi}dx.\end{array}$$

It is not hard to see the bilinear operator M is coercive and continuous and L is continuous. Then, applying the Lax-Milgram Theorem to find $\forall\left(\bar{\varphi},\bar{\psi}\right)\in\left(H_0^1\left((0,L)\right)\right)^2$, we see that the system (29) has a unique weak solution $(\varphi,\psi)\in\left(H_0^1\left((0,L)\right)\right)^2$. Owing to the classical elliptic regularity, we find by (29) that

$$(\varphi,\psi)\in\left(H^2\left((0,L)\right)\right)^2.$$

It remains only to prove

$$\nu^2+\beta\phi_i-z_i\left(x,1\right)\varpi\left(\nu\right)\in L^2\left((0,L)\times(-\infty,+\infty)\right),i=1,2.$$

Indeed, we have from $(17)_5$ and $(17)_6$,

$$\begin{array}{l}\left(\nu^2+\beta\right)\phi_1-z_1\left(x,1\right)\varpi\left(\nu\right)=f_5-\phi_1\in L^2\left((0,L)\times(-\infty,+\infty)\right),\\ \left(\nu^2+\beta\right)\phi_2-z_2\left(x,1\right)\varpi\left(\nu\right)=f_6-\phi_2\in L^2\left((0,L)\times(-\infty,+\infty)\right).\end{array}$$

Therefore, $U\in D\left(A\right)$. Thus, the operator $I-A$ is surjective. Now, we prove that

$$F:\ \mathcal{H}\to\mathcal{H}$$

is locally Lipschitz. For $U, \bar{U} \in \mathcal{H}$, we have

$$\|\mathbb{F}(U) - \mathbb{F}(\bar{U})\|_{\mathcal{H}}^2 \quad \leq C\left[\|\varphi - \bar{\varphi}\|_{H_0^1(\Omega)}^2 + \|\psi - \bar{\psi}\|_{H_0^1(\Omega)}^2\right]. \tag{30}$$

Thus, $\mathbb{F}$ is locally Lipschitz. This completes the proof.

We show the global existence result. First, we introduce the following useful functionals:

$$\begin{array}{l} I_1(t) = b_1 \int_0^L \int_{-\infty}^{+\infty} |\phi_1(x,\nu,t)|^2 \ d\nu \ dx + k\|\varphi_x + \psi\|^2 \\ +\frac{b}{2}\|\psi_x\|^2 - \|\varphi\|_p^p + sv_1 \int_\Omega \int_0^1 |z_1(x,\rho,t)|^2 \ d\rho \ , \end{array} \tag{31}$$

$$\begin{array}{l} I_2(t) = b_2 \int_\Omega \int_{-\infty}^{+\infty} |\phi_2(x,\nu,t)|^2 \ d\nu \ dx + \frac{b}{2}\|\psi_x\|^2 \\ -\|\psi\|_q^q + sv_2 \int_\Omega \int_0^1 |z_2(x,\rho,t)|^2 \ d\rho \ dx, \end{array} \tag{32}$$

$$\begin{array}{l} J_1(t) = \frac{b_1}{2} \int_\Omega \int_{-\infty}^{+\infty} |\phi_1(x,\nu,t) \, d\nu \ dx|^2 + \frac{k}{2}\|\varphi_x + \psi\|^2 + \frac{b}{4}\|\psi_x\|^2 \\ -\frac{1}{p}\|\varphi\|_p^p + sv_1 \int_\Omega \int_0^1 |z_1(x,\rho,t)|^2 \ d\rho \ dx, \end{array} \tag{33}$$

and

$$\begin{array}{l} J_2(t) = \frac{b_2}{2} \int_\Omega \int_{-\infty}^{+\infty} |\phi_2(x,\nu,t)|^2 \ d\nu \ dx + \frac{b}{4}\|\psi_x\|^2 \\ - \frac{1}{q}\|\psi\|_q^q + sv_2 \int_\Omega \int_0^1 |z_2(x,\rho,t)|^2 \ d\rho \ dx. \end{array} \tag{34}$$

We easily see that

$$\mathcal{E}(t) = \frac{1}{2}\|\varphi_t\|^2 + \frac{1}{2}\|\psi_t\|^2 + J_1(t) + J_2(t). \tag{35}$$

Lemma 3.1 *Suppose that conditions (4), $p > 2$ and $q > 2$ hold. Then, for $U_0 \in \mathcal{H}$ satisfying*

$$\begin{cases} \widetilde{B} = \max\left(C_{\star\star}^p \left(\frac{2p}{p-2}\mathcal{E}(0)\right)^{\frac{p-2}{2}}, \ C_{\star\star}^q \left(\frac{2q}{q-2}\mathcal{E}(0)\right)^{\frac{q-2}{2}}\right) < 1, \\ I_i(0) > 0 \quad \textit{for } i = 1,2, \end{cases} \tag{36}$$

we have for all $t > 0$,

$$I_i(t) > 0, \quad \textit{for } i = 1,2.$$

Proof. As $I_i(0) > 0$ for $i = 1,2$, by continuity of φ and ψ, there exists $T^\star < T$ such that

$$I_i(t) \geq 0 \quad \text{for all } t \in [0, t^\star], \quad i = 1,2, \tag{37}$$

and with a straight forward calculation, we can find

$$\begin{array}{l} \frac{2p}{p-2}J_1(t) = k\|\varphi_x + \psi\|^2 + b_1 \int_\Omega \int_{-\infty}^{+\infty} |\phi_1(x,\nu,t)|^2 \ d\nu dx + \frac{b}{2}\|\psi_x\|^2 \\ +\frac{2(p-1)}{p-2} sv_1 \int_\Omega \int_0^1 |z_1(x,\rho,t)|^2 \ d\rho \ dx + \frac{2}{p-2}I_1(t) \geq k\|\varphi_x + \psi\|^2 + \frac{b}{2}\|\psi_x\|^2, \end{array} \tag{38}$$

$$\begin{array}{l} \frac{2q}{q-2}J_2(t) = b_2 \int_\Omega \int_{-\infty}^{+\infty} |\phi_2(x,\nu,t)|^2 \ d\nu \ dx + \frac{b}{2}\|\psi_x\|^2 \\ + \frac{2(q-1)}{q-2} sv_2 \int_\Omega \int_0^1 |z_2(x,\nu,t)|^2 \ d\nu \ dx + \frac{2}{q-2}I_2(t) \geq \frac{b}{2}\|\psi_x\|^2. \end{array} \tag{39}$$

Exploiting (35), (38), (39) and Lemma 2.4, we find

$$\frac{b}{2}\|\psi_x\|^2 + k\|\varphi_x + \psi\|^2 \leq \frac{2p}{p-2}\mathcal{E}(t) \leq \frac{2p}{p-2}\mathcal{E}(0) \quad \text{for all } t \in [0, t^\star], \tag{40}$$

and

$$\frac{b}{2}\|\psi_x\|^2 \leq \frac{2q}{q-2}\mathcal{E}(t) \leq \frac{2q}{q-2}\mathcal{E}\left(0\right) \quad \text{for all } t \in [0, t^\star]\,. \tag{41}$$

Applying Sobolev-Poincaré's inequality and taking into consideration (36), (40) and (41), we get

$$\begin{aligned} \|\varphi\|_p^p &\leq C_\star^p \|\varphi_x\|^p \leq C_{1\star}^p \left[\sqrt{k}\|\varphi_x + \psi\| + \sqrt{\tfrac{b}{2}}\|\psi_x\|\right]^p \\ &\leq C_{\star\star}^p \left(\tfrac{2p}{p-2}\mathcal{E}\left(0\right)\right)^{\frac{q-2}{2}} \left[k\|\varphi_x + \psi\|^2 + \tfrac{b}{2}\|\psi_x\|^2\right] \leq k\|\varphi_x + \psi\|^2 + \tfrac{b}{2}\|\psi_x\|^2, \end{aligned} \tag{42}$$

and

$$\|\psi\|_q^q \leq C_\star^q \|\psi_x\|_2^q = C_{\star\star}^q \left[\tfrac{2q}{q-2}\mathcal{E}\left(0\right)\right]^{\frac{p-2}{2}} \tfrac{b}{2}\|\psi_x\|_2^2 \leq \tfrac{b}{2}\|\psi_x\|_2^2. \tag{43}$$

This implies that

$$I_i(t) > 0 \quad \text{for } i = 1, 2 \ \forall\ t \in [0, t^\star]\,.$$

By repeating this procedure and using the fact that

$$\lim_{t \to T^\star} \max\left(C_{\star\star}^p \left(\frac{2p}{p-2}\mathcal{E}\left(0\right)\right)^{\frac{p-2}{2}},\ C_{\star\star}^q \left(\frac{2q}{q-2}\mathcal{E}\left(0\right)\right)^{\frac{q-2}{2}}\right) < 1,$$

we can take $T^\star = T$.

Theorem 3.2 *(Global existence) Assume that condition (10), $p > 2$ and $q > 2$ are satisfied. Then, for $U_0 \in D(A)$ satisfying (36), the solution of system* (8) *is global in time.*

Proof. It suffices to show that $\|\varphi_x + \psi\|^2 + \|\psi_x\|^2 + \|\psi_t\|^2 + \|\varphi_t\|^2$ is bounded independently of t.
Indeed, by (35), (38), (39), we get

$$\begin{aligned} &\mathcal{E}\left(0\right) \geq \mathcal{E}(t) = \tfrac{1}{2}\left[\|\varphi_t\|^2 + \|\psi_t\|^2\right] + J_1(t) + J_2(t) \\ &\geq \min\left(\tfrac{1}{2}, \tfrac{(p-2)}{2p}k, \tfrac{q-2}{2q}\tfrac{b}{2}\right)\left[\|\varphi_t\|^2 + \|\psi_t\|^2 + \|\varphi_x + \psi\|^2 + \|\psi_x\|^2\right], \end{aligned}$$

which implies that

$$\|\varphi_t\|^2 + \|\psi_t\|^2 + \|\varphi_x + \psi\|^2 + \|\psi\|^2 \leq CE\left(0\right),$$

where C is a constant depending only on p, q, k and b.

4 Decay Rate Result

Our next step is devoted to the proof of the decay result to the problem (8). For this purpose, we prepare some Lemmas and present some appropriate functionals. Firstly, we define

$$k_1(t) = \sum_{i=1}^{2}\int_\Omega \rho_i \varphi_t^i \varphi^i \ dx + \sum_{i=1}^{2}\frac{bi}{2}\int_\Omega\int_{-\infty}^{+\infty}\left(\nu^2 + \beta\right)\left|M_i\left(x, \nu, t\right)\right|^2 \ d\nu \ dx, \tag{44}$$

and

$$k_2(t) = s\sum_{i=1}^{2}\int_{\Omega}\int_{0}^{1} e^{-s\rho}\left|z_i\left(x,\rho,t\right)\right|^2\ d\rho\ dx, \tag{45}$$

where

$$M_i\left(x,\nu,t\right) = \int_0^t \phi_i\left(x,\nu,z\right)\ dz - \frac{s\varpi\left(\nu\right)}{\nu^2+\beta}\int_0^1 f_0^i\left(x,-\rho s\right)\ d\rho + \frac{\varphi_0^i\varpi\left(\nu\right)}{\nu^2+\beta} \tag{46}$$

with

$$\left(f_0^i\left(x,-\rho s\right),\varphi_0^i\left(x\right),\varphi^i\left(x,t\right)\right) = \begin{cases} \left(f_0\left(x,-\rho s\right),\varphi_0\left(x\right),\varphi\left(x,t\right)\right) & i=1, \\ \left(g_0\left(x,-\rho s\right),\psi_0\left(x\right),\psi\left(x,t\right)\right) & i=2. \end{cases}$$

Lemma 4.1 *[1] Let $(\varphi,\phi_1,z_1,\psi,\phi_2,z_2)$ be a regular solution of problem (8), then we have*

$$\left(\nu^2+\beta\right)M_i\left(x,\nu,t\right) = -s\varpi\left(\nu\right)\int_0^1 z_i\left(x,\rho,t\right)\ d\rho + \varphi^i\left(x,t\right)\varpi\left(\nu\right) - \phi_i\left(x,\nu,t\right),$$

and

$$\begin{gathered}\int_\Omega\int_{-\infty}^{+\infty}\left(\nu^2+\beta\right)\phi_i\left(x,\nu,t\right)M_i\left(x,\nu,t\right)\ d\nu\ dx = \int_\Omega\varphi^i\left(x,t\right)\int_{-\infty}^{+\infty}\phi_i\left(x,\nu,t\right)\varpi\left(\nu\right)\ d\nu\ dx \\ -s\int_\Omega\int_0^1 z_i\left(x,\rho,t\right)\int_{-\infty}^{+\infty}\varpi\left(\nu\right)\phi_i\left(x,\nu,t\right)d\nu\ d\rho\ dx - \int_\Omega\int_{-\infty}^{+\infty}\left|\phi_i\left(x,\nu,t\right)\right|^2 d\nu\ dx, i=1,2.\end{gathered}$$

Lemma 4.2 *[1] Let $(\varphi,\phi_1,z_1,\psi,\phi_2,z_2)$ be a regular solution of the problem (8), then we have*

$$\begin{aligned}\left|\int_\Omega\int_{-\infty}^{+\infty}\left(\nu^2+\beta\right)\left|M_i\left(x,\nu,t\right)\right|^2\ d\nu\ dx\right| &\le 3s^2A_0\int_\Omega\int_0^1\left|z_i\left(x,\rho,t\right)\right|^2\ d\rho\ dx \\ &+ 3A_0C_\star^2\|\varphi_x^i\|_2^2 + \tfrac{3}{\beta}\int_\Omega\int_{-\infty}^{+\infty}\left|\phi_i\left(x,\nu,t\right)\right|^2\ d\nu\ dx,\ i=1,2.\end{aligned}$$

Lemma 4.3 *Assume (4) with $p>2$ and $q>2$ hold. The functional k_1 defined in (44) satisfies*

$$\begin{aligned}&k_1'(t) \le -C_1\|\varphi_x+\psi\|^2 - C_2\|\psi_x\|^2 + C\|\varphi_t\|^2 + C\|\psi_t\|^2 \\ &-\sum_{i=1}^{2} b_i\int_\Omega\int_{-\infty}^{+\infty}\left|\phi_i\left(x,\nu,t\right)\right|^2\ d\nu\ dx + \|\varphi\|_p^p + \|\psi\|_q^q \\ &+s^2\sum_{i=1}^{2} v_i\int_\Omega\int_0^1\left|z_i\left(x,\rho,t\right)\right|^2\ d\rho\ dx + \sum_{i=1}^{2}\tfrac{b_i}{4}\int_\Omega\int_{-\infty}^{+\infty}\left(\nu^2+\beta\right)\left|\phi_i\left(x,\nu,t\right)\right|^2\ d\nu\ dx,\end{aligned} \tag{47}$$

where C_1, C_2, C are positive constants.

Proof. Differentiating k_1 with respect to t, using $(8)_1$ and $(8)_2$, by integration by parts and using Lemma 4.1, we obtain

$$\begin{aligned}&k_1'(t) = -k\|\varphi_x+\psi\|^2 - b\|\psi_x\|^2 + \rho_1\|\varphi_t\|^2 + \rho_2\|\psi_t\|^2 \\ &-\sum_{i=1}^{2} b_i\int_\Omega\int_{-\infty}^{+\infty}\left|\phi_i\left(x,\nu,t\right)\right|^2\ d\nu\ dx - \alpha_1\int_\Omega\varphi\varphi_t\ dx - \alpha_2\int_\Omega\psi\psi_t\ dx \\ &-s\sum_{i=1}^{2} b_i\int_\Omega\int_0^1 z_i\int_{-\infty}^{+\infty}\varpi\phi_i\left(x,\nu,t\right)\ d\nu\ d\rho\ dx + \|\varphi\|_p^p + \|\psi\|_q^q.\end{aligned} \tag{48}$$

Now, we will estimate the last three terms of the RHS as follows. Using Lemma 2.3 and due to the fact that $b_iA_0 < v_i,\ i = 1, 2$, and then integrating over $(0, 1)$ with respect to ρ, we can write

$$\begin{aligned}&-\sum_{i=1}^{2} b_i \int_\Omega sz_i \int_{-\infty}^{+\infty} \varpi\phi\left(x,\nu,t\right)\ d\nu\ dx \leq s^2 \sum_{i=1}^{2} v_i \int_\Omega \int_0^1 \left|z_i\left(x,\rho,t\right)\right|^2\ d\rho\ dx \\ &+\sum_{i=1}^{2} \tfrac{b_i}{4} \int_\Omega \int_{-\infty}^{+\infty} \left(\nu^2+\beta\right) \left|\phi_i\left(x,\nu,t\right)\right|^2\ d\nu\ dx.\end{aligned} \tag{49}$$

By Young and Poincaré's inequality, we have

$$\begin{aligned}-\alpha_1 \int_\Omega \varphi\varphi_t\ dx + \alpha_2 \int_\Omega \psi\psi_t\ dx \quad &\leq \tfrac{\alpha_1}{4\delta}\|\varphi_t\|^2 + \tfrac{\alpha_2}{4\delta}\|\psi_t\|^2 + C\delta\alpha_1\|\varphi_x+\psi\|^2 \\ &+C\delta\left(\alpha_2+\alpha_1 C\right)\|\psi_x\|^2.\end{aligned} \tag{50}$$

Inserting (49) and (50) in (48), we arrive at

$$\begin{aligned}&k_1'(t) \leq -(k-C\delta\alpha_1)\|\varphi_x+\psi\|^2-(b-C\delta\left(\alpha_2+\alpha_1C\right))\|\psi_x\|^2+\left(\tfrac{\alpha_1}{4\delta}+\rho_1\right)\|\varphi_t\|^2 \\ &+\left(\tfrac{\alpha_2}{4\delta}+\rho_2\right)\|\psi_t\|^2 - \sum_{i=1}^{2} b_i \int_\Omega \int_{-\infty}^{+\infty} \left|\phi_i\left(x,\nu,t\right)\right|^2\ d\nu\ dx + s^2\sum_{i=1}^{2} v_i \int_\Omega \int_0^1 \left|z_i\left(x,\nu,t\right)\right|^2\ d\rho\ dx \\ &+\sum_{i=1}^{2} \tfrac{b_i}{4} \int_\Omega \int_{\infty}^{+\infty} \left(\nu^2+\beta\right) \left|\phi_i\left(x,\nu,t\right)\right|^2\ d\nu\ dx + \|\varphi\|_p^p + \|\psi\|_q^q,\end{aligned}$$

we choose $\delta = \min\left(\frac{b}{2C(\alpha_2+\alpha_1C)}, \frac{k}{2C\alpha_1}\right)$, then setting $C_1 = k - C\delta\alpha_1$ and $C_2 = b - C\delta\left(\alpha_2+\alpha_1C\right)$, we get (47).

Lemma 4.4 *With the same hypotheses as in Lemma 4.3, the functional k_2 defined in (45) satisfies*

$$k_2'(t) \leq -se^{-s}\sum_{i=1}^{2}\int_\Omega\int_0^1 \left|z_i\left(x,\rho,t\right)\right|^2\ d\rho\ dx + \|\varphi_t\|^2 + \|\psi_t\|^2. \tag{51}$$

Proof. We take the derivative of k_2 with respect to t, and using $(8)_4$ and $(8)_6$, we get

$$k_2'(t) = \sum_{i=1}^{2}\int_\Omega \left|z_i\left(x,0,t\right)\right|^2 - \sum_{i=1}^{2}\int_\Omega e^{-s}\left|z_i\left(x,1,t\right)\right|^2 d\rho\, dx - s\sum_{i=1}^{2}\int_\Omega\int_0^1 e^{-s\rho}\left|z_i\left(x,\rho,t\right)\right|^2 d\rho\, dx.$$

We have $z_i\left(x,0,t\right) = \varphi_t^i\left(x,t\right)$, and since $e^{-s\rho} \geq e^{-s}$, we obtain (51).

Now, we introduce the perturbed modified energy, named Lyapunov function, as

$$\mathcal{L}(t) = N\mathcal{E}(t) + \varepsilon k_1(t) + k_2(t)$$

for $\varepsilon > 0$ and $N > 0$.

Lemma 4.5 *For ε_1 small and N large enough, we have*

$$\frac{N}{2}\mathcal{E}(t) \leq \mathcal{L}(t) \leq 2N\mathcal{E}(t), \quad \forall\ t \geq 0. \tag{52}$$

Proof. The application of Young and Poincaré's inequalities gives

$$\mathcal{L}(t) \le N\mathcal{E}(t) + \tfrac{\varepsilon}{2}\left[\rho_1\|\varphi_t\|^2 + \rho_1 C_\star^2\|\varphi_x+\psi\|^2\right] + \tfrac{\varepsilon}{2}\left[\rho_2\|\psi_t\|^2 + C_{\star\star}^2\left\{\rho_2 + C_\star^2\rho_1\right\}\|\psi_x\|^2\right]$$
$$+\sum_{i=1}^{2}\tfrac{b_i}{2}\varepsilon\int_\Omega\int_{-\infty}^{+\infty}\left(\nu^2+\beta\right)|M_i(x,\nu,t)|^2\,d\nu\,dx + s\sum_{i=1}^{2}\int_\Omega\int_0^1 e^{-s\rho}|z_i(x,\rho,t)|^2\,d\rho\,dx.$$

Using $\mathcal{E}(t)$, I_1, I_2, Lemma 4.2 and the fact that $b_iA_0 < v_i$ for $i=1,2$, we get

$$2N\mathcal{E}(t) - \mathcal{L}(t) \ge \tfrac{\rho_1}{2}[N-\varepsilon]\|\varphi_t\|_2^2 + \tfrac{\rho_2}{2}[N-\varepsilon]\|\psi_t\|_2^2$$
$$+\tfrac{N}{p}I_1 + \tfrac{N}{q}I_2 + \tfrac{1}{2}\left[\tfrac{Nk(p-2)}{p} - \varepsilon C_\star^2[3v_1+\rho_1]\right]\|\varphi_x+\psi\|^2$$
$$+\tfrac{1}{2}\left[\tfrac{bN(pq-q-p)}{pq} - \varepsilon C_{\star\star}^2\left\{C_\star^2(3v_1+\rho_1)+3v_2+\rho_2\right\}\right]\|\psi_x\|^2$$
$$+s\int_\Omega\int_0^1\left(\left[\tfrac{Nv_1(p-1)}{p}-1-\tfrac{3}{2}s\varepsilon v_1\right]|z_1(x,\rho,t)|^2+\left[\tfrac{Nv_1(q-1)}{q}-1-\tfrac{3}{2}s\varepsilon v_2\right]|z_2(x,\rho,t)|^2\right)d\rho\,dx$$
$$+\tfrac{b_1}{2}\left[\tfrac{N(p-2)}{p}-\tfrac{3\varepsilon}{\beta}\right]\int_\Omega\int_{-\infty}^{+\infty}|\phi_1(x,\nu,t)|^2\,d\nu\,dx+\tfrac{b_2}{2}\left[\tfrac{N(q-2)}{q}-\tfrac{3\varepsilon}{\beta}\right]\int_\Omega\int_{-\infty}^{+\infty}|\phi_2(x,\nu,t)|^2\,d\nu\,dx.$$

On the other hand, we can estimate the following:

$$\mathcal{L}(t) - \tfrac{N}{2}\mathcal{E}(t) \ge \tfrac{N}{2}\mathcal{E}(t) - \tfrac{\varepsilon}{2}\left[\rho_1\|\varphi_t\|^2 + C_\star^2\rho_1\|\varphi_x+\psi\|^2\right]$$
$$+\tfrac{\varepsilon}{2}\left[\rho_2\|\psi_t\|^2 + C_{\star\star}^2\left\{C_\star^2\rho_1+\rho_2\right\}\|\psi_x\|^2\right] + s\sum_{i=1}^{2}e^{-s}\int_\Omega\int_0^1 z_i(x,\rho,t)\,d\rho\,dx$$
$$+\sum_{i=1}^{2}\tfrac{b_i\varepsilon}{2}\int_\Omega\int_{-\infty}^{+\infty}\left(\nu^2+\beta\right)|M_i(x,\nu,t)|^2\,d\nu\,dx.$$

Using Lemma 4.2 and the fact that $b_iA_0 < v_i,\ i=1,2$, we obtain

$$\mathcal{L}(t) - \tfrac{N}{2}\mathcal{E}(t) \ge \tfrac{\rho_1}{2}\left[\tfrac{N}{2}-\varepsilon\right]\|\varphi_t\|^2 + \tfrac{\rho_2}{2}\left[\tfrac{N}{2}-\varepsilon\right]\|\psi_t\|^2 + \tfrac{N}{2}pI_1 + \tfrac{N}{2}qI_2$$
$$+\tfrac{1}{2}\left[\tfrac{kN(p-2)}{2p} - C_\star^2\varepsilon(\rho_1+3v_1)\right]\|\varphi_x+\psi\|^2$$
$$+\tfrac{1}{2}\left[\tfrac{Nb(qp-p-q)}{2pq} - \varepsilon C_{\star\star}^2\left\{\rho_2+3v_2+C_\star^2(\rho_1+3v_1)\right\}\right]\|\psi_x\|^2$$
$$+\tfrac{b_1}{2}\left(\tfrac{N(p-2)}{2p}-\tfrac{3\varepsilon}{\beta}\right)\int_\Omega\int_{-\infty}^{+\infty}|\phi_1(x,\nu,t)|^2\,d\nu\,dx$$
$$+\tfrac{b_2}{2}\left(\tfrac{N(q-2)}{2q}-\tfrac{3\varepsilon}{\beta}\right)\int_\Omega\int_{-\infty}^{+\infty}|\phi_2(x,\nu,t)|^2\,d\nu\,dx$$
$$+s\left[\tfrac{v_1N(p-1)}{2p}+e^{-s}-\tfrac{3}{2}\varepsilon v_1 s\right]\int_\Omega\int_0^1|z_1(x,\rho,t)|^2\,d\rho\,dx$$
$$+s\left[\tfrac{v_2N(q-1)}{2q}+e^{-s}-\tfrac{3}{2}sv_2\varepsilon\right]\int_\Omega\int_0^1|z_2(x,\rho,t)|^2\,d\rho\,dx.$$

Finally, if we pick ε small and N large enough, we deduce that

$$\mathcal{L}(t) - \frac{N}{2}\mathcal{E}(t) \ge 0 \quad\text{and}\quad 2N\mathcal{E}(t) - \mathcal{L}(t) \ge 0.$$

Hence, we conclude that

$$\mathcal{E}(t) \sim \mathcal{L}(t) \quad \forall t > 0.$$

Theorem 4.1 *(Exponential decay rate) Let $p > 2$ and $q > 2$. Assume that (4) holds for $i = 1, 2$, and $U_0 \in \mathcal{H}$ satisfying (36), then the unique solution of* (8) *satisfies*

$$\mathcal{E}(t) \le ke^{-mt} \quad \forall t \ge 0,$$

for some positive constants k and m independent of t.

Proof. We remember that

$$\mathcal{L}(t) = N\mathcal{E}(t) + \varepsilon k_1(t) + k_2(t).$$

By means of Lemma 4.3 and Lemma 4.4, we get for all $t \geq 0$,

$$\begin{aligned}
&\mathcal{L}'(t) \leq -\left(NC - \varepsilon C - 1\right)\|\varphi_t\|^2 - \left(NC - \varepsilon C - 1\right)\|\psi_t\|^2 \\
&-\sum_{i=1}^{2} \tfrac{b_i}{2}\left[N - \tfrac{\varepsilon}{2}\right] \int_0^L \int_{-\infty}^{\infty} \left(\nu^2 + \beta\right) \left|\phi_i\left(x, \nu, t\right)\right|^2 \, d\nu \, dx - C_1\varepsilon\|\varphi_x + \psi\|^2 - C_2\varepsilon\|\psi_x\|^2 \\
&-\sum_{i=1}^{2} \varepsilon b_i \int_\Omega \int_{-\infty}^{\infty} \left|\phi_i\left(x, \nu, t\right)\right|^2 \, d\nu \, dx - \sum_{i=1}^{2} s\left(e^{-s} - v_i s\varepsilon\right) \int_\Omega \int_0^1 \left|z_i\left(x, \rho, t\right)\right|^2 \, d\rho \, dx \\
&+\varepsilon\left[\|\varphi\|_p^p + \|\psi\|_q^q\right].
\end{aligned}$$

We now choose ε small enough such that $e^{-s} - v_i s\varepsilon > 0, i = 1, 2$. Pick N large enough such that $N > max\left(\frac{C\varepsilon+1}{C}, \frac{\varepsilon}{2}\right)$. Thus, $\exists m_1 > 0$ so that

$$\mathcal{L}'(t) \leq -m_1\mathcal{E}(t) \quad \forall t \geq 0.$$

By Lemma 4.5, it follows that $\mathcal{E}(t)$ and $\mathcal{L}(t)$ are equivalent $\forall t > 0$. Then, $\exists m > 0$ such that

$$\mathcal{L}'(t) \leq -m\mathcal{L}(t) \quad \forall t \geq 0. \tag{53}$$

Hence, the solution of (53) is given by

$$\mathcal{L}'(t) \leq \mathcal{L}\left(0\right) e^{-mt} \quad \forall t \geq 0,$$

so, we have

$$\mathcal{E}(t) \leq ke^{-mt} \quad \forall t \geq 0,$$

with $k > 0$. This completes the proof.

Example

Consider the problem (1) with $\Omega = (0, 2\pi), \rho_1 = \rho_2 = 1$,
$p = q = 3 > 2, b = 1, K = \frac{1}{2}, \varphi_0(x) = \psi_0(x) = \frac{1}{\sqrt{24\pi C}} \sin x$,
$\varphi_1(x) = \psi_1(x) = -\frac{1}{\sqrt{24\pi C}} \sin x$, where C is the maximal value between two constants denoted by the same notation C_{**} and they are given by (42) and (43). The initial delays $f_0\left(x, t - s\right) = g_0\left(x, t - s\right) = 0$ for $t \in (0, s)$. We set $v_i = 2b_i A_0$ and $\alpha_i = 4b_i A_0$ for i=1,2. Then we have

1. The initial condition $U_0 = \frac{1}{\sqrt{24\pi C}}\left(sinx, -sinx, sinx, -sinx, 0, 0, 0, 0\right) \in D(A)$.

2. By Lemma 2.2, we have $A_0 = \frac{\pi\beta^{\alpha-1}}{sin(\pi\alpha)}$, from the definition of b_i, it follows that $\alpha_i = 4a_i\beta^{\alpha-1}$. Then the condition (4) is satisfied.

3. It easy to notice that the relation (10) holds.

4. From the expression of the energy (9), we get $\mathcal{E}(0) = \frac{1}{12C^3}$. Thus, $\widetilde{B} = \frac{1}{\sqrt{2}} < 1$. By a simple and direct calculation, we find $I_1(0) = 3I_2(0) = \frac{1}{48C^3} > 0$. Then we deduce that the conditions (36) are verified.

So, by Theorem 3.1 and Theorem 3.2, the problem (1) has a unique local and global solution. Furthermore, by Theorem 4.1, we get the decay result.

Conclusion

In this paper, we prove the well-posedness result of problem (1) using the semi-group theories. Then, we prove that the solution decay exponentially by means of the multiplier approach. Finally, we provide an example in which our results can be applied. The main contribution of this work is the extension of the previous results from [2]. It will be interesting to extend our results to the following system:

$$\begin{cases} \rho_1\varphi_{tt} - k\left(\varphi_x + \psi\right)_x + a_1\partial_t^{\alpha,\beta}\varphi\left(0t - s\right) = \left|\varphi\right|^{p-2}\varphi, \\ \rho_2\psi_{tt} - b\psi_{xx} + k\left(\varphi_x + \psi\right) + a_2\partial_t^{\alpha,\beta}\psi\left(t - s\right) = \left|\psi\right|^{q-2}\psi, \\ \varphi(x, t=0) = \varphi_0(x),\ \psi(x, t=0) = \psi_0(x), \\ \varphi_t(x, t=0) = \varphi_1(x),\ \psi_t\left(x, 0\right) = \psi_1(x), \\ \varphi_t\left(x, t-s\right) = f_0\left(x, t-s\right), t \in \left(0, s\right), \\ \psi_t\left(x, t-s\right) = g_0\left(x, t-s\right), t \in \left(0, s\right) \end{cases}$$

under the following boundary conditions:

$$\begin{cases} (\varphi_x + \psi)(L, t) + \alpha_1\varphi_t(L, t) = 0, \\ \psi_x(L, t) + \alpha_2\phi_t(L, t) = 0, \end{cases}$$

which will be an open problem.

Acknowledgment

The authors would like to thank the anonymous referees who provided useful and detailed comments on a previous/earlier version of the manuscript.

References

[1] A. Adnane, A. Benaissa and K. Benomar. Uniform stabilization for a Timoshenko beam system with delays in fractional order internal damping. SeMA Journal (2022) 283–302.

[2] R. Aounallah, A. Benaissa and A. Zaraï. Blow-up and asymptotic behavior for a wave equation with a time delay condition of fractional type. *Rendiconti del Circolo Matematico di Palermo Series 2* **70** (2) (2021) 1061–1081.

[3] R. Datko, J. Lagnese and M. Polis. An example on the effect of time delays in boundary feedback stabilization of wave equations. *SIAM Journal on Control and Optimization* **24** (1) (1986) 152–156.

[4] C. Duque, H. Leiva and A. Tridane. *Existence and uniqueness of solutions for a semilinear functional dynamic equation with infinite delay and impulses.* Nonlinear Dynamics and Systems Theory **22** (2) (2022) 155–168.

[5] J. Ferreira, E. Piskin, C. Raposo, M. Shahrouzi and H. Yuksekkaya. Stability result for a Kirchhoff beam equation with variable exponent and time delay. *Universal Journal of Mathematics and Applications* **5** (1) (2022) 1–9.

[6] V. Georgiev and G. Todorova. Existence of a solution of the wave equation with nonlinear damping and source terms. *Journal of differential equations* **109** (2) (1994) 295–308.

[7] A. Guesmia and A. Soufyane. On the stability of Timoshenko-type systems with internal frictional damping and discrete time delays. *Applicable Analysis* **96** (12) (2017) 2075–2101.

[8] J. Hao and F. Wang. Energy decay in a Timoshenko-type system for thermoelasticity of type III with distributed delay and past history. *Electronic Journal of Differential Equations* **75** (2018) 1–27.

[9] L. He. On decay and blow-up of solutions for a system of viscoelastic equations with weak damping and source terms. *Journal of Inequalities and Applications* **2019** (1) (2019) 1–27.

[10] O. D. Kichmarenko, I. V. Chepovskyi, Y. Platonova and S. Dashkovskiy. *Analysis of solutions to equations with a generalized derivative and delay.* Nonlinear Dynamics and Systems Theory **23** (2) (2023) 195–206.

[11] J. E. Lagnese. *Boundary Stabilization of This Plates.* SIAM, Philadelphia, 1989.

[12] H. A Levine. Instability and nonexistence of global solutions to nonlinear wave equations of the form $\mathrm{P}u_{tt} = -\mathrm{A}u + \mathcal{F}(u)$. *Transactions of the American Mathematical Society* **192** (1974) 1–21.

[13] H. A. Levine. Some additional remarks on the nonexistence of global solutions to nonlinear wave equations. *SIAM Journal on Mathematical Analysis* **5** (1) (1947) 138–146.

[14] M. Lv and J. Hao. General decay and blow-up for coupled kirchhoff wave equations with dynamic boundary conditions. *Mathematical Control and Related Fields* (2021).

[15] S. Nicaise and J. Valein. Stabilization of second order evolution equations with unbounded feedback with delay. *ESAIM: Control, Optimisation and Calculus of Variations* **16** (2) (2010) 420–456.

[16] A. Peyravi. General decay and blow up of solutions for a system of viscoelastic wave equations with nonlinear boundary source terms. *Journal of Mathematical Analysis and Applications* **451** (2) (2017) 1056–1076.

[17] I. Podlubny. Fractional differential equations. *Mathematics in science and engineering* **198** (1999) 41–119.

[18] S. G. Samko, A. A. Kilbas, O. I. Marichev, et al. *Fractional Integrals and Derivatives.* Gordon and breach science, Yverdon Yverdon-les-Bains, Switzerland **1** (1993).

[19] S. Timonchenko, On the correction for shear of the differential equation for transverse vibrations of prismatic bars. *Pilos. Mag.* **41** (1921) 744–746.

[20] K. Zennir. Stabilization for solutions of plate equation with time-varying delay and weak-viscoelasticity in $\mathbb{R}^n$. *Russian Mathematics* **64** (2020) 21–33.

Nonlinear Dynamics and Systems Theory, **25** (3) (2025) 315–337

Formation Flight of UAVs for Search and Detection Missions by Tracking Time-Variable Trajectories

Rolando Díaz-Castillo [1], Rosa Martha Lopéz-Gutiérrez [1*], Juan José Cetina-Denis [2] and César Cruz-Hernández [2]

[1] *Engineering, Architecture and Design Faculty, Baja California Autonomous University (UABC), Ensenada 22860, BC, México.*
[2] *Electronics and Telecommunication Department, Scientific Research and Advanced Studies Center of Ensenada (CICESE), Ensenada 22860, BC, México.*

Received: November 6, 2024; Revised: June 7, 2025

Abstract: In this paper, the formation problem of multiple unmanned aerial vehicles (UAVs) is addressed. In particular, the formation of UAVs is achieved by using complex systems theory and backstepping nonlinear control. We apply the obtained formation of multiple UAVs to search for and detect a target of interest within an exploration area. In addition, a coverage study of the formation of UAVs for search and detection by tracking time-variable trajectories is reported.

Keywords: *complex systems; formation control; backstepping control; multiple UAVs; search mission.*

Mathematics Subject Classification (2020): 70K42, 93-08, 93-16.

1 Introduction

From the formation of some kinds of birds to extend their flight time, to the grouping of fish to avoid attacks of predators, different groups of animals often associate naturally to achieve a common goal or benefit, which they individually could not achieve, and therefore could not survive [6], [18], [19].

The exchange of information due to the interactions between the members of these groups gives rise to a set of collective behaviors that are different from an isolated individual behavior. It is called emergent collective behavior [17], [21].

* Corresponding author: mailto:roslopez@uabc.edu.mx

In the literature, there are many works dealing with this problem, one of them is the work by R. Abas and Wu [1], in which the dynamic model of a quadcopter is studied by using the Newton-Euler method, and the synchronization of three quadcopters is achieved in the simulation using sliding modes as a control algorithm.

In the work done by P. Flores [9], the author faces the problem of the formation of a group of unmanned aerial vehicles (UVAs), for this a dynamic model of a quadcopter is consedered using the Newton-Euler method, the control algorithm for the formation is backstepping control, and the formation of a group of three quadcopters is achieved.

In the work reported by A. Toledo [2], the dynamic model of a quadcopter is considered using the Newton-Euler method and an integral backstepping control algorithm with sliding modes is proposed for an unmanned aerial vehicle. The experimental results are obtained by using a Qball-X4 quadcopter.

In the work by N. Koksal [14], the dynamic model of the Qball-X4 quadcopter is considered. A PID control algorithm is used for the translation system and another algorithm is applied for the rotation system, the simulations results are obtained for a group of 3 quadcopters, and experimental tests with two Qball-X4 type quadcopters are carried out.

In the work done by X. Dong [7], a dynamic model for a small UAV type mini helicopter is considered assuming that there is a leading quadcopter and the other are followers, they use a PID control algorithm and obtain the formation of the group of quadcopters in simulation and experimental results.

The main goals of this paper are: (i) to obtain network synchronization and formation flight of coupled UAVs in star topology, considering a single master UAV with four slave UAVs. This objective is achieved by using recent results from complex systems theory. In addition, (ii) to apply the network formation to object detection, and (iii) to carry out a coverage study of the formation of UAVs for search and detection by tracking time-variable trajectories. To our knowledge, the results have not been reported.

The organization of the paper is as follows. In Section 2, the problem statement is presented. Section 3 describes the mathematical model of the UAV quadrotor used in this work. Section 4 contains the designed control algorithm for synchronization and formation of UAVs. Section 5 presents the obtained numerical results. In Section 6, an application to object detection is provided. Finally, some conclusions are given in Section 7.

2 Problem Statement

In recent decades, many control proposals have emerged in order to achieve formations in mobile robots. Particularly, formations in Unmanned Aerial Vehicles (UAVs) have received considerable interest due to their wide potential applications in the military, civil and industrial fields, and agriculture [10], [13]. The purpose of this study is to preserve mobility and compact groups at the same time, which generates advantages such as reduced implementation costs, increased robustness, system efficiency, etc.

The quadcopter is used to access hostile environments, where the safety of the pilots can not be guaranteed. The quadcopter´s configuration makes it capable of taking off vertically, controlled landing, as well as great maneuverability. These advantages have attracted many researchers' interest in recent years.

Different control techniques can be applied to a quadcopter, for example, a nonlinear controller, PID control, backstepping, dynamic feedback linearization, and sliding modes, among others. See for example [3], [7], [12].

The study of collective behaviors seen in nature and their representation in mathematical equations opens the door for multiple applications in robotics and, in our particular case, the formation of multiple unmanned aerial vehicles (UVAs).

The purpose of this paper is to reproduce collective behaviors observed in animals, namely synchronization and formation, and apply them to the networks of unmanned aerial vehicles (UVAs) for applications in search, rescue, and patrol task. Fig. 1 illustrates a group of quadcopters searching for the target of interest T.

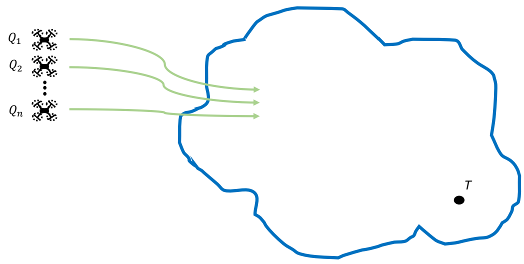


Figure 1: Group of quadrotors searching for the target T.

We will solve the stated problem on the network formation of five UAVs with a single master and four slaves by using complex systems theory and nonlinear backstepping control, providing an analytical stability proof based on the Lyapunov theory, and we will also analyze the search and detection coverage of the object in the area. In addition, for a particular type of UAV, we will use the mathematical model of the quadrotor described in Section 3.

3 Quadrotor Dynamic Model

The complete quadrotor dynamic model, with the $x, y, z-$plane position and orientation angles ($roll, pitch$, and yaw), is described by [2], [3]- [5], [20]

$$\begin{array}{l}
\ddot{\phi} = \dot{\theta}\dot{\psi}\left(\frac{I_y - I_z}{I_x}\right) - \frac{J_r}{I_x}\dot{\theta}\Omega + \frac{l}{Ix}U_2, \\
\ddot{\theta} = \dot{\phi}\dot{\psi}\left(\frac{I_z - I_x}{I_y}\right) + \frac{J_r}{I_y}\dot{\phi}\Omega + \frac{l}{Iy}U_3, \\
\ddot{\psi} = \dot{\phi}\dot{\theta}\left(\frac{I_x - I_y}{I_z}\right) + \frac{l}{I_z}U_4, \\
\ddot{x} = (cos\phi\ sin\theta \cos\psi + sin\phi\ sin\psi)\frac{1}{m}U_1, \\
\ddot{y} = (cos\phi\ sin\theta\ sin\psi - sin\phi\ cos\psi)\frac{1}{m}U_1, \\
\ddot{z} = -g + cos\phi\ cos\theta\frac{1}{m}U_1.
\end{array} \tag{1}$$

The first three differential equations correspond to the quadrotor orientation $(\phi, \theta, \psi)^T$, and the last three differential equations represent the position of the quadrotor with respect to the original inertial frame $(x, y, z)^T$, see Figure 2.

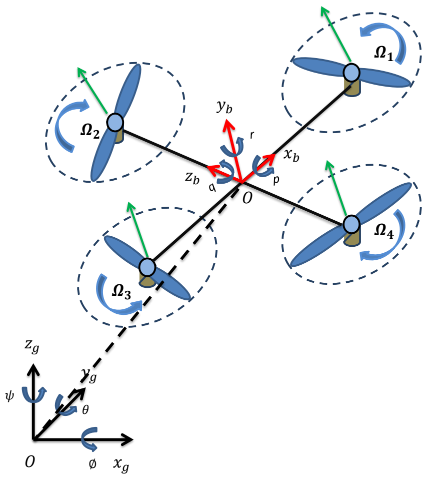


Figure 2: Quadcopter representation with respect to the inertial frame.

The angular velocity due to the propellers in each engine is represented by Ω_i, for $i = 1, 2, 3, 4$, respectively. The control inputs of the quadrotor are denoted by U_i, $i = 1, 2, 3, 4$, and Ω is a disturbance, which correspond to

$$\begin{aligned}
&U_1 = b(\Omega_1^2 + \Omega_2^2 + \Omega_3^2 + \Omega_4^2),\\
&U_2 = b(\Omega_4^2 - \Omega_2^2),\\
&U_3 = b(\Omega_3^2 - \Omega_1^2),\\
&U_4 = d(\Omega_2^2 + \Omega_4^2 - \Omega_1^2 - \Omega_3^2),\\
&\Omega = \Omega_2 + \Omega_4 - \Omega_1 - \Omega_3.
\end{aligned} \tag{2}$$

The quadrotor dynamic model described in Eq. (1) can be rewritten in a state space as $\dot{\boldsymbol{X}} = f(\boldsymbol{X}, \boldsymbol{U})$, introducing the following state vector:

$$\boldsymbol{X} = [\phi, \dot{\phi}, \theta, \dot{\theta}, \psi, \dot{\psi}, x, \dot{x}, y, \dot{y}, z, \dot{z}]^T, \tag{3}$$

where

$$\begin{aligned}
x_1 &= \phi, & x_2 &= \dot{x}_1 = \dot{\phi},\\
x_3 &= \theta, & x_4 &= \dot{x}_3 = \dot{\theta},\\
x_5 &= \psi, & x_6 &= \dot{x}_5 = \dot{\psi},\\
x_7 &= z, & x_8 &= \dot{x}_7 = \dot{z},\\
x_9 &= x, & x_{10} &= \dot{x}_9 = \dot{x},\\
x_{11} &= y, & x_{12} &= \dot{x}_{11} = \dot{y}.
\end{aligned} \tag{4}$$

From Equations (1) and (4), the quadrotor mathematical model can be described in the state space as follows:

$$\dot{\boldsymbol{X}} = f(\boldsymbol{X}, \boldsymbol{U}) = \begin{pmatrix} x_2 \\ x_4 x_6 a_1 + x_4 a_2 \Omega + b_1 U_2 \\ x_4 \\ x_2 x_6 a_3 + x_2 a_4 \Omega + b_2 U_3 \\ x_6 \\ x_4 x_2 a_5 + b_3 U_4 \\ x_8 \\ -g + (cosx_1 cosx_3)\frac{1}{m} U_1 \\ x_{10} \\ u_x \frac{1}{m} U_1 \\ x_{12} \\ u_y \frac{1}{m} U_1 \end{pmatrix}, \tag{5}$$

where
$a_1 = \frac{I_y - I_z}{I_x}, \quad a_2 = -\frac{J_r}{I_x}, \quad a_3 = \frac{I_z - I_x}{I_y}, \quad a_4 = \frac{J_r}{I_y}, \quad a_5 = \frac{I_x - I_y}{I_z},$
$b_1 = \frac{l}{I_x}, \quad b_2 = \frac{l}{I_y}, \quad b_3 = \frac{l}{I_z},$
$u_x = cosx_1 \; sinx_3 \; cosx_5 + sinx_1 \; sinx_5,$
$u_y = cosx_1 \; sinx_3 \; sinx_5 - sinx_1 \; cosx_5.$

The quadrotor mathematical model (5) can be divided into two subsystems: orientation and translation. The first one is given by

$$\dot{\boldsymbol{x}}_o = \boldsymbol{f}_o(\boldsymbol{x}_i) + \boldsymbol{B}_i \boldsymbol{U}_{oi}, \tag{6}$$

where $\boldsymbol{U}_{oi} = \begin{pmatrix} U_2 & U_3 & U_4 \end{pmatrix}^T$,

$$\boldsymbol{f}_o = \begin{pmatrix} x_2 \\ x_4 x_6 a_1 + x_4 a_2 \Omega \\ x_4 \\ x_2 x_6 a_3 + x_2 a_4 \Omega \\ x_6 \\ x_4 x_2 a_5 \end{pmatrix}, \quad \boldsymbol{B}_i = \begin{pmatrix} 0 & 0 & 0 \\ b_1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & b_2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & b_3 \end{pmatrix}. \tag{7}$$

On the other hand, the translation subsystem is described by the following expression:

$$\dot{\boldsymbol{x}}_{ti} = \boldsymbol{f}_t(\boldsymbol{x}_i) + \boldsymbol{G} \boldsymbol{U}_{ti}, \tag{8}$$

where $\boldsymbol{U}_{ti} = \begin{pmatrix} U_1 & U_x & U_y \end{pmatrix}^T$,

$$\boldsymbol{f}_t = \begin{pmatrix} x_8 \\ -g \\ x_{10} \\ 0 \\ x_{11} \\ 0 \end{pmatrix}, \quad \boldsymbol{G} = \begin{pmatrix} 0 & 0 & 0 \\ \frac{cosx_1 cosx_3}{m} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & \frac{1}{m} & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \frac{1}{m} \end{pmatrix}. \tag{9}$$

The physical parameters of the quadrotor mathematical model (5) are given in Table 1.

Parameter	**Definition**	**Value**
m	Mass	$0.650\ kg$
I_x	x-axis inertia	$7.5e^{-3}\ kgm^2$
I_y	y-axis inertia	$7.5e^{-3}\ kgm^2$
I_z	z-axis inertia	$1.3e^{-2}\ kgm^2$
b	Thrust coefficient	$3.13e^{-3}\ Ns^2$
d	Drag coefficient	$7.5e^{-7}\ Nms^2$
J_r	Rotor inertia	$6e^{-5}\ kgm^2$
l	Arm length	$0.23\ m$
g	Gravity	$9.8\ N/kg$

Table 1: Physical parameters of the quadrotor mathematical model (5).

4 Control Design for Trajectory Tracking

In nonlinear control theory, backstepping is a technique developed around 1990 by Petar V. Kokotovic, Miroslav Krstiv, and Ioannis Kanellakopoulos [15] to design stabilizing controls for a special class of nonlinear dynamical systems. These systems are built from subsystems that originate from an irreducible subsystem that can be stabilized using some other method.

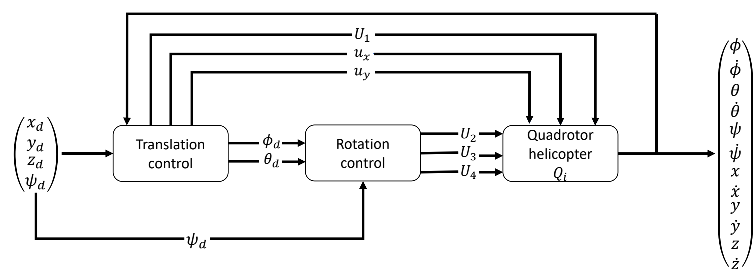


Figure 3: Block diagram of quadrotor helicopter Q and its controller.

In the backstepping approach, the control law is designed so that the system can follow the desired trajectory. For this, it is considered that the quadrotor mathematical model (5) can be divided into two subsystems, one is the orientation and the other is the position [5], as previously discussed in Section 3. Figure 3 shows the block diagram of quadcopter and its controller for tracking a desired path.

Due to its complete independence with respect to the other subsystem (Eqs.(6) and (8)), the control input for the angular rotations of the subsystem is considered first and then the position control input is derived. A desired trajectory x_{1d} is defined, in which the following error is given by

$$z_1 := x_{1d} - x_1, \tag{10}$$

from expression (10), we have

$$\dot{z}_1 = \dot{x}_{1d} - \dot{x}_1. \tag{11}$$

From the quadrotor mathematical model (5), $\dot{x}_1 = x_2$ is known. Substituting into (11), we have

$$\dot{z}_1 = \dot{x}_{1d} - x_2. \tag{12}$$

Now, consider the following Lyapunov candidate function in terms of z_1:

$$V(z_1) = \frac{1}{2} z_1^2. \tag{13}$$

Differentiating the candidate Lyapunov function with respect to time gives

$$\dot{V}(z_1) = z_1 \dot{z}_1. \tag{14}$$

Substituting Equation (12) in (14) gives

$$\dot{V}(z_1) = z_1(\dot{x}_{1d} - x_2). \tag{15}$$

x_2 is considered as a virtual control to stabilize z_1, thus we have

$$x_2 = \dot{x}_{1d} + \alpha_1 z_1, \tag{16}$$

we make $\alpha_1 > 0$ so that the derivative of the Lyapunov function is negative definite. Solving for (16) in (15), we have

$$\begin{aligned} \dot{V}(z_1) &= z_1(\dot{x}_{1d} - x_2) \\ &= z_1(\dot{x}_{1d} - \dot{x}_{1d} - \alpha_1 z_1) \\ &= -\alpha_1 z_1^2. \end{aligned} \tag{17}$$

After the variable change

$$z_2 = x_2 - \dot{x}_{1d} - \alpha_1 z_1, \tag{18}$$

differentiating Equation (18), we have

$$\dot{z}_2 = \dot{x}_2 - \ddot{x}_{1d} - \alpha_1 \dot{z}_1. \tag{19}$$

The following Lyapunov candidate function is proposed as a function of (z_1, z_2):

$$V(z_1, z_2) = \frac{1}{2}(z_1^2 + z_2^2). \tag{20}$$

Differentiating the candidate Lyapunov function and solving, we have

$$\begin{aligned}\dot{V}(z_1, z_2) &= z_2\dot{z}_2 + z_1\dot{z}_1 \\ &= z_2(\dot{x}_2 - \ddot{x}_{1d} - \alpha_1\dot{z}_1) + z_1(\dot{x}_{1d} - x_2) \\ &= z_2(\dot{x}_2 - \ddot{x}_{1d} - \alpha_1(\dot{x}_{1d} - x_2)) + z_1(\dot{x}_{1d} - x_2).\end{aligned} \tag{21}$$

Solve x_2 from Equation (18): $x_2 = z_2 + \dot{x}_{1d} + \alpha_1 z_1$, thus

$$\begin{aligned}\dot{V}(z_1, z_2) &= z_2(\dot{x}_2 - \ddot{x}_{1d} - \alpha_1(\dot{x}_{1d} - x_2)) + z_1(\dot{x}_{1d} - x_2) \\ &= z_2(\dot{x}_2 - \ddot{x}_{1d} - \alpha_1(\dot{x}_{1d} - (z_2 + \dot{x}_{1d} \\ &\quad + \alpha_1 z_1)) + z_1(\dot{x}_{1d} - (z_2 + \dot{x}_{1d} + \alpha_1 z_1)) \\ &= z_2\dot{x}_2 - z_2(\ddot{x}_{1d} - \alpha_1(z_2 + \alpha_1 z_1)) - z_1 z_2 - \alpha_1 z_1^2 \\ &= z_2(a_1 x_4 x_6 + a_2 x_4 \Omega + b_1 U_2) - z_2(\ddot{x}_{1d} - \alpha_1(z_2 + \alpha_1 z_1)) \\ &\quad - z_1 z_2 - \alpha_1 z_1^2.\end{aligned} \tag{22}$$

Considering $\ddot{x}_{1d,2d,3d} = 0$ and given $\dot{V}(z_1, z_2) < 0$, the virtual controller U_2 is designed as

$$U_2 = \frac{1}{b_1}(z_1 - a_1 x_4 x_6 - a_2 x_4 \Omega - \alpha_1(z_2 + \alpha_1 z_1) - \alpha_2 z_2). \tag{23}$$

The remaining control inputs U_3, U_4, and U_1 can be solved by a similar approach, obtaining the corresponding virtual controllers for each control input:

$$U_3 = \frac{1}{b_2}(z_3 - a_3 x_2 x_6 - a_4 x_2 \Omega - \alpha_3(z_4 + \alpha_3 z_3) - \alpha_4 z_4), \tag{24}$$

$$U_4 = \frac{1}{b_3}(z_5 - a_5 x_2 x_4 - \alpha_5(z_2 + \alpha_1 z_1) - \alpha_6 z_6). \tag{25}$$

The control input for the positioning subsystem is given by

$$U_1 = \frac{m}{cos x_1 cos x_3}(z_7 + g - \alpha_7(z_8 + \alpha_7 z_7) - \alpha_8 z_8 + \ddot{x}_7), \tag{26}$$

$$u_x = \frac{m}{U_1}(z_9 - \alpha_9(z_{10} + \alpha_9 z_9) - \alpha_{10} z_{10} + \ddot{x}_9), \tag{27}$$

$$u_y = \frac{m}{U_1}(z_{11} - \alpha_{11}(z_{12} + \alpha_{11} z_{11}) - \alpha_{12} z_{12} + \ddot{x}_{11}), \tag{28}$$

where

$$\begin{aligned}z_1 &= x_{1d} - x_1, \\ z_2 &= x_2 - \dot{x}_{1d} - \alpha_1 z_1, \\ z_3 &= x_{3d} - x_3, \\ z_4 &= x_4 - \dot{x}_{3d} - \alpha_3 z_3, \\ z_5 &= x_{5d} - x_5, \\ z_6 &= x_6 - \dot{x}_{5d} - \alpha_5 z_5, \\ z_7 &= x_{7d} - x_7, \\ z_8 &= x_8 - \dot{x}_{7d} - \alpha_7 z_7, \\ z_9 &= x_{9d} - x_9, \\ z_{10} &= x_{10} - \dot{x}_{9d} - \alpha_9 z_9, \\ z_{11} &= x_{11d} - x_{11}, \\ z_{12} &= x_{12} - \dot{x}_{11d} - \alpha_{11} z_{11}.\end{aligned} \tag{29}$$

Communication in a multi-agent topology can be represented directly or indirectly by a graph, where each node is an agent and the edges are the communication media that exist between them [11]. A group of 5 quadcopters (Eq.(5)) is considered, one of which is the master quadrotor (M) and the rest are slave quadrotors (S_1, S_2, S_3, S_4). The network of quadrotors can be represented by the following graph shown in Figure 4.

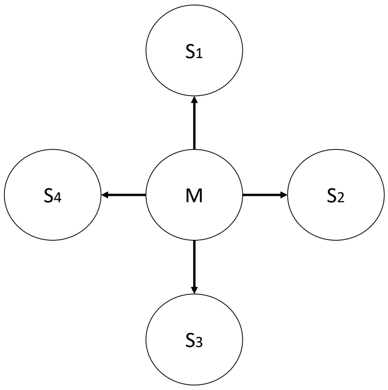


Figure 4: Connection graph of 5 quadrotors.

The corresponding adjacency matrix associated with this graph is

$$A(G) = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{30}$$

The network synchronization was achieved through the variables z_i of the control law, given by Eq. (29), so each slave quadrotor $S_i, i = 1, 2, 3, 4$, follows the master quadrotor M, and their states will have to follow the states of the master quadrotor. The auxiliary variables for each quadrotor are given as follows:

$$\begin{aligned} z_{1Si} &= x_{1M} - x_{1Si}, \\ z_{2Si} &= x_{2Si} - \dot{x}_{1M} - \alpha_{1Si} z_{1Si}, \\ z_{3Si} &= x_{3M} - x_{3Si}, \\ z_{4Si} &= x_{4Si} - \dot{x}_{3M} - \alpha_{3Si} z_{3Si}, \\ z_{5Si} &= x_{5M} - x_{5Si}, \\ z_{6Si} &= x_{6Si} - \dot{x}_{5M} - \alpha_{5Si} z_{5Si}, \\ z_{7Si} &= x_{7M} - x_{7Si}, \\ z_{8Si} &= x_{8Si} - \dot{x}_{7M} - \alpha_{7Si} z_{7Si}, \\ z_{9Si} &= x_{9M} - x_{9Si}, \\ z_{10Si} &= x_{10Si} - \dot{x}_{9M} - \alpha_{9Si} z_{9Si}, \\ z_{11Si} &= x_{11M} - x_{11Si}, \\ z_{12Si} &= x_{12Si} - \dot{x}_{11M} - \alpha_{11Si} z_{11Si}. \end{aligned} \tag{31}$$

When performing the calculations from Equation (5) to Equation (9), with the help of the auxiliary variables (31), the following control inputs for network synchronization of the quadrotors in star connection are obtained:

$$\begin{aligned} U_{1l} = \frac{m}{cos\phi cos\theta}(\dot{x}_8 + g) + \frac{m}{cos\phi cos\theta}\sum_{j=1}^{N=4} a_{ij}[(1+\alpha_7\alpha_8)(x_{7j} - x_{7i}) \\ + (\alpha_7 + \alpha_8)(x_{8j} - x_{8i})], \end{aligned} \tag{32}$$

$$\begin{aligned} u_{xl} = \frac{m}{U_{1l}}\dot{x}_{10} + \frac{m}{U_{1l}}\sum_{j=1}^{N=4} a_{ij}[(1+\alpha_9\alpha_{10})(x_{9j} - x_{9i}) \\ + (\alpha_9 + \alpha_{10})(x_{10j} - x_{10i})], \end{aligned} \tag{33}$$

$$\begin{aligned} u_{yl} = \frac{m}{U_{1l}}\dot{x}_{12} + \frac{m}{U_{1l}}\sum_{j=1}^{N=4} a_{ij}[(1+\alpha_{11}\alpha_{12})(x_{11j} - x_{11i}) \\ + (\alpha_{11} + \alpha_{12})(x_{12j} - x_{12i})], \end{aligned} \tag{34}$$

where $(l = 1, 2, 3, 4)$ is the *i-th* quadrotor, $(\dot{x}_8, \dot{x}_{10}, \dot{x}_{12})$ are the states of the master quadrotor M, a_{ij} are the entries of the adjacency matrix (30) associated with the graph used for communication between the quadrotors.

In order to achieve quadrotor formation, the use of a vector Δ with components corresponding to each axis is proposed and then used in the controller of each slave quadrotor $(S_i,\ i = 1, 2, 3, 4)$. This vector separates each slave from the master M by a distance corresponding to each axis in the plane (x, y, z). The auxiliary variables for the formation of each quadrotor respectively are given as follows:

$$\begin{aligned} z_{1Si} &= x_{1M} - x_{1Si}, \\ z_{2Si} &= x_{2Si} - \dot{x}_{1M} - \alpha_{1Si}z_{1Si}, \\ z_{3Si} &= x_{3M} - x_{3Si}, \\ z_{4Si} &= x_{4Si} - \dot{x}_{3M} - \alpha_{3Si}z_{3Si}, \\ z_{5Si} &= x_{5M} - x_{5Si}, \\ z_{6Si} &= x_{6Si} - \dot{x}_{5M} - \alpha_{5Si}z_{5Si}, \\ z_{7Si} &= x_{7M} - x_{7Si} + \Delta_{zSi}, \\ z_{8Si} &= x_{8Si} - \dot{x}_{7M} - \alpha_{7Si}z_{7Si}, \\ z_{9Si} &= x_{9M} - x_{9Si} + \Delta_{xSi}, \\ z_{10Si} &= x_{10Si} - \dot{x}_{9M} - \alpha_{9Si}z_{9Si}, \\ z_{11Si} &= x_{11M} - x_{11Si} + \Delta_{ySi}, \\ z_{12Si} &= x_{12Si} - \dot{x}_{11M} - \alpha_{11Si}z_{11Si}. \end{aligned} \tag{35}$$

5 Numerical Results

When considering a star-shaped topology connection between five quadrotors, a numerical simulation is carried out, in which the group of quadrotors will follow a desired circular trajectory with a radius of 3 meters at a height of 2 meters. These trajectories start after the quadrotors takes off. The initial conditions for the quadrotors are: Master $(x_1(0), y_1(0), z_1(0)) = (0, 0, 0)$, slave 1 $(x_2(0), y_2(0), z_2(0)) = (0.5, 0, 0)$, slave 2 $(x_3(0), y_3(0), z_3(0)) = (-0.5, 0, 0)$, slave 3 $(x_4(0), y_4(0), z_4(0)) = (1.5, 0, 0)$, and slave 4 $(x_5(0), y_5(0), z_5(0)) = (-1.5, 0, 0)$. The physical parameters of the five quadrotors are taken from Table 1 and the alphas are definitely positive with a value of 50.

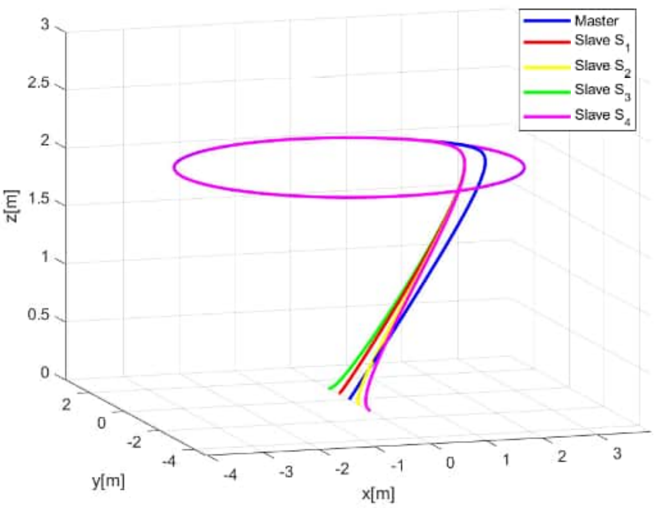


Figure 5: Synchronization of five quadrotors to a desired circular trajectory.

In Figure 5, it can be seen that the network of five quadrotors is synchronized in a desired circular trajectory. Next, to achieve the network formation of five quadrotors, we proceed to give a net constant separation between the master quadrotor M and the slave quadrotors S_i by using the vector Δ for the separation. The initial conditions for quadrotors are the same as considered in the previous simulation. The quadrotors are separated by a distance Δ on the x-axis for a horizontal line formation, where the slave quadrotors S_1 and S_2 are desired to be separated by a distance $D1 = 0.5\ m$ from the master M, while slave quadrotors S_3 and S_4 are separated from the master quadrotor M by a distance $D2 = 1\ m$. This can be seen in Figure 6.

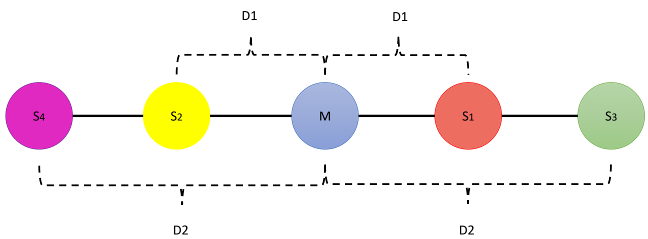


Figure 6: Separation Δ for the formation of five quadrotors.

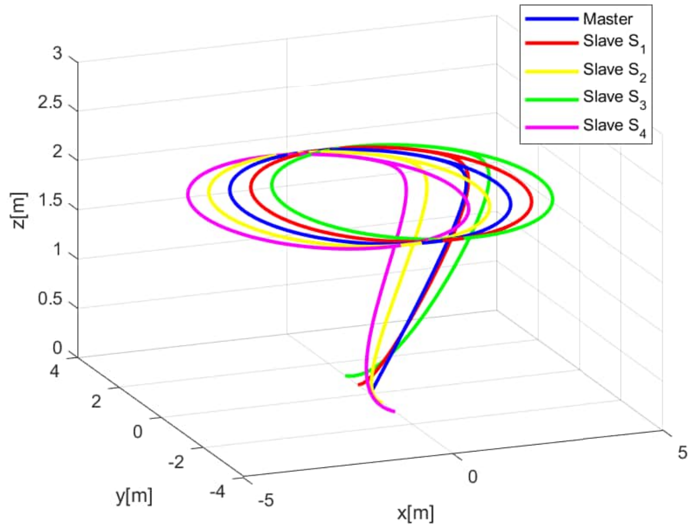


Figure 7: Flight of the formation of five quadrotors describing a circular trajectory.

Figure 7 shows the five quadrotors are already in formation describing a circular trajectory. Next, we will apply the formation of quadrotors to the search for an object on a surface. For this, a rectangular search surface is considered, but it can be deduced that using a circular trajectory to explore a rectangular surface would have a great disadvantage since there are unexplored areas since the object, when placed randomly on the surface, could not be detected. For this reason, it is decided to use the Lissajous trajectories, of which the denser trajectory is used.

6 Application to Object Detection

The target of interest to be detected within the search area is a circle object, which varies in size with respect to the total percentage of the search area. The tarjet to be searched and detected is randomly placed within the exploration area. The target can be considered detected by using the coordinates of the quadrotors and the coordinates where the center of the target is located, considering the radius of the target. The detection of the target is determined by using the coordinates of the trajectories of each quadrotor and the coordinates of the center of the target calculating their distances by means of the expression

$$d(A,B) = |\overrightarrow{AB}|. \tag{36}$$

Substituting the coordinates of the center of the target and the trajectory of the quadrotor, the elements of vector A and B are obtained. The distance is calculated as follows:

$$d(A,B) = |\overrightarrow{AB}| = \sqrt{(X_2 - X_1)^2 + (Y_2 - Y_1)^2}, \tag{37}$$

where X_2 and Y_2 are the quadrotor coordinates, and X_1 and Y_1 are the position of the center of the target. The following Figure 8 shows how the detection is done.

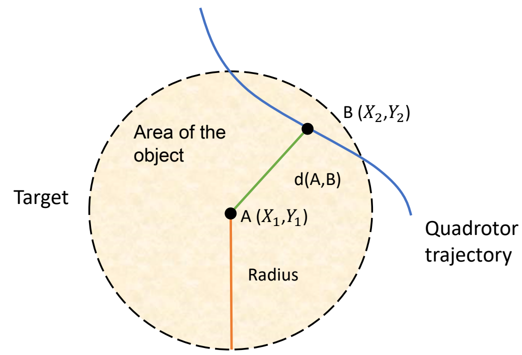


Figure 8: Target detection process.

For target detection, if any point of the trajectories of the quadrotors is within the radius of the target, it indicates that this point is within the area of the target, concluding that it was detected and, given that more than one quadrotor trajectory passes through this area, we can determine which quadrotor passed first and met this condition before the others.

Lissajous curves

The Lissajous curve, also known as the Lissajous figure or Bowditch curve, is the graph of the parametric equation system corresponding to the superposition of two simple harmonic motions in perpendicular directions [8], [16], [22] defined by

$$x = Asin(w_x t + \alpha), \quad y = Bsin(w_y t + \beta), \quad \delta = \alpha - \beta. \tag{38}$$

The Lissajous curves for different parameter values are shown in Figure 9.

Now, we will proceed to carry out some numerical simulations in MatLab, where quadrotors formation explores a rectangular area. This search surface, or the area of exploration, has a dimension of 8x6 m^2, the Lissajous curve used has the parameter values $5:6$ and $\frac{\pi}{2}$ with the values of $A = 3$, $B = 3$, $w_x = 6$, and $w_y = 5$. Quadrotors will be considered to start outside of this area. The search for the randomly placed target will begin, and then the area of this target will be varied as shown in Table 2.

The first simulation is carried out with a target size of 10% with respect to the search area and in the network, the quadrotors are separated by a distance Δ on the x-axis for a horizontal line formation, where the slave quadrotors S_1 and S_2 are desired to be separated by a distance $D1 = 0.5\ m$ from the master M, while the slave quadrotors S_3 and S_4 are separated from the master quadrotor M by a distance $D2 = 1\ m$, as shown in Figure 6. The initial conditions for the quadrotors are: Master $(x_1(0), y_1(0), z_1(0)) = (4.5, 0, 0)$, slave 1 $(x_2(0), y_2(0), z_2(0)) = (4.5, 0.6, 0)$, slave 2 $(x_3(0), y_3(0), z_3(0)) = (4.5, 1.2, 0)$, slave 3 $(x_4(0), y_4(0), z_4(0)) = (4.5, -0.6, 0)$, and slave 4 $(x_5(0), y_5(0), z_5(0)) = (4.5, -1.2, 0)$.

As seen in Figure 10, several trajectories of the quadrotor formation pass over the target, so it can be deduced that it was easily detected within the search surface. It can

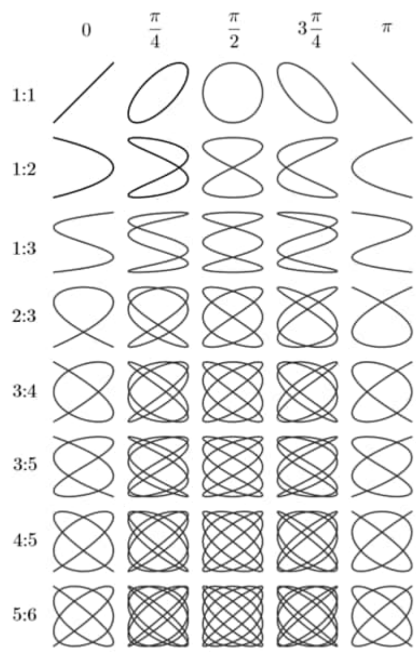


Figure 9: Lissajous curves described by (38) for different parameter values.

Surface percentage	Object area	Object radius
10 %	4.8 m^2	1.23 m
7.5 %	3.6 m^2	1.07 m
5 %	2.4 m^2	0.87 m
1 %	0.48 m^2	0.39 m
0.5 %	0.24 m^2	0.27 m
0.3 %	0.144 m^2	0.21 m
0.1 %	0.048 m^2	0.12 m
0.01 %	0.0048 m^2	0.039 m

Table 2: Object area table.

be considered that the searched target was detected due to its big size, so it is decided to reduce it, according to Table 2. Next, a new numerical simulation is performed using a target with a size of 1% with respect to the rectangular search surface. The initial conditions for quadrotors are the same as considered in the previous simulation.

Figure 11 shows that the searched target was detected by at least two quadrotors. Considering that its size makes it easier to be found, its dimensions can be reduced further with respect to the rectangular search surface.

In addition, another simulation is performed now using a target sized 0.1% with respect to the rectangular search surface. The initial conditions for quadrotors are the

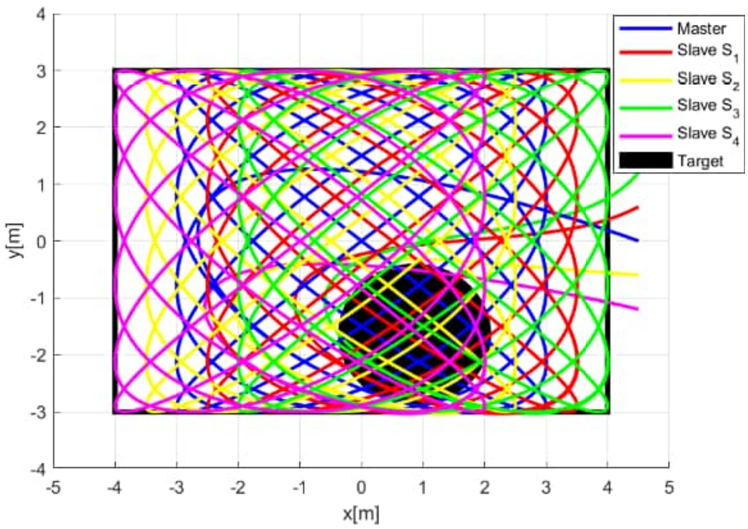


Figure 10: Search for a target sized 10% of the exploration area.

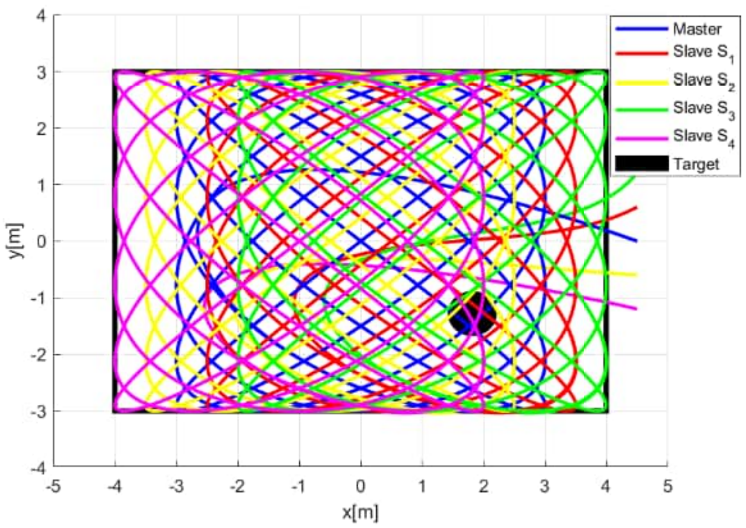


Figure 11: Search for a target sized 1% of the exploration area.

same as considered in the previous simlation.

Figure 12 shows that the target was detected by two of the slave quadrotors. In addition, it can be considered that it already has a suitable size with respect to the rectangular search surface, therefore we will use this target size. It was decided to use this target size to run 10 tests, to determine if some of the quadrotors can detect it. The

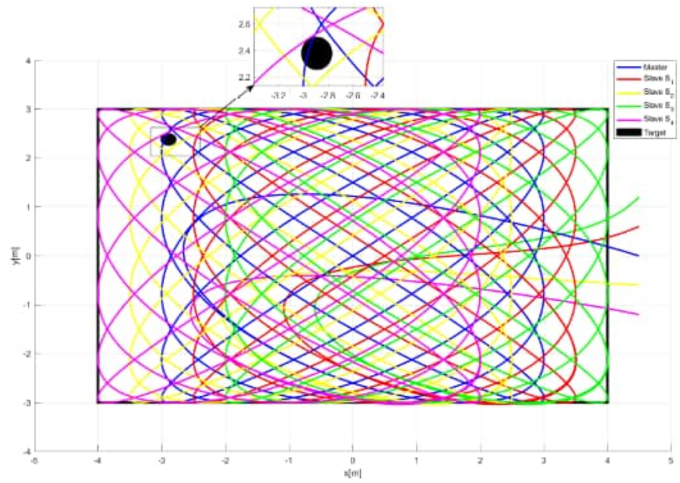


Figure 12: Search for a target sized 0.1% of the exploration area.

obtained numerical results are shown in Table 3.

Test	Run time	Detection time	x-axis position	y-axis position	Target detected by
1	0.622 s	966.648 s	2.439 m	2.334 m	S_1
2	0.624 s	389.48 s	-2.891 m	2.334 m	M
3	0.620 s	309.201 s	1.026 m	-2.315 m	M
4	0.620 s	309.201 s	-1.717 m	0.269 m	S_3
5	0.631 s	958.754 s	3.546 m	2.674 m	S_1
6	0.626 s	147.690 s	-2.654 m	-2.707 m	S_2
7	0.629 s	250.066 s	2.328 m	-2.060 m	M
8	0.628 s	128.67 s	-0.606 m	2.391 m	S_3
9	0.630 s	1047.094 s	2.265 m	2.643 m	S_2
10	0.736 s	72.773 s	1.207 m	-2.670 m	S_3

Table 3: 10 tests to search for a target sized 0.1 % of the rectangular search area.

As shown in Table 3, the target was detected in each of the ten tests that were carried out. However, considering the size of the target in Figure 12, it can be seen that there are zones through which no trajectory passes and the target could be located in any of them. For this reason, it was decided to combine some of the Lissajous curves of the Figure 9 to get a suitable trajectory for the formation of quadrotors that do pass through these zones.

6.1 Lissajous curves as desired trajectories

It is desired to explore most of the search area, for which some of the Lissajous curves will be used. First, the curve with the parameters 2 : 3 and $\frac{\pi}{2}$ will be used. Taking into account that, due to the shape of the curve, there will be unexplored zones, it is decided to combine it with another Lissajous curve with the parameters 3 : 4 and $\frac{\pi}{2}$. These curves can be seen in Figure 9.

Considering the combined trajectory 1 with two mentioned Lissajous curves, one after another, we have

$$Trajectory_1 = \begin{cases} t_0 \quad to \quad t_1; \quad x = 3sin(2\pi t + \pi), \quad y = 3sin(3\pi t + 0.5\pi), \\ t_1 \quad to \quad t_2; \quad x = 3sin(4\pi t + \pi), \quad y = 3sin(3\pi t + 0.5\pi). \end{cases} \tag{39}$$

To explore the rectangular search area with the combined $trajectory_1$, the final trajectory is taken into consideration. With the combination of both trajectories, a target with the size 0.1 % of the total search area will be detected. This is shown in Figure 13.

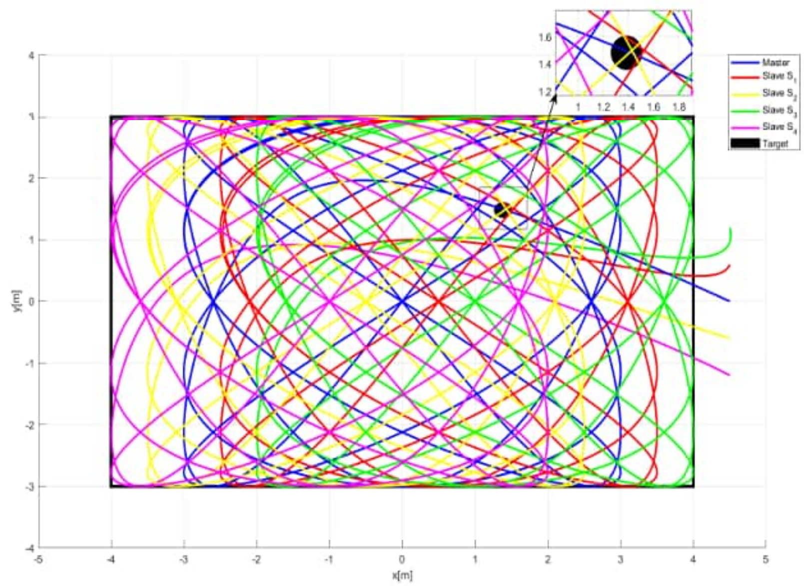


Figure 13: Target search with the combined trajectory 1 for the target sized 0.1% of the rectangular search area.

Ten tests are performed with this combined trajectory 1 for target detection, the numerical results are shown in Table 4.

According to the obtained results (Table 4), the target was also found with this combined trajectory 1. However, the unexplored zones did not decrease. Therefore, it was decided to use another combination of two other Lissajous curves.

We explore the search area again, making use of a different combination of the Lissajous curves. The curve with the parameters 3 : 4 and $\frac{\pi}{2}$ will be first used, and then another curve with parameters 6 : 5 and $\frac{\pi}{2}$. These curves can be seen in Figure 9.

The resulting combined trajectory 2 is described as follows.

$$Trajectory_2 = \begin{cases} t_0 \quad to \quad t_1; \quad x = 3sin(4\pi t + \pi), \quad y = 3sin(3\pi t + 0.5\pi), \\ t_1 \quad to \quad t_2; \quad x = 3sin(6\pi t + \pi), \quad y = 3sin(5\pi t + 0.5\pi). \end{cases} \tag{40}$$

The resulting search $trajectory_2$ is shown in Figure 14.

Test	Run time	Detection time	x-axis position	y-axis position	Target detected by
1	1.044 s	865.115 s	2.439 m	2.334 m	S_2
2	0.519 s	792.091 s	-2.891 m	2.334 m	M
3	0.157 s	1075.385 s	1.026 m	-2.315 m	S_2
4	0.145 s	60.001 s	-1.717 m	0.269 m	S_1
5	0.146 s	286.197 s	3.546 m	2.674 m	S_1
6	0.148 s	834.347 s	-2.654 m	-2.707 m	S_2
7	0.155 s	1142.108 s	2.328 m	-2.060 m	S_3
8	0.136 s	888.934 s	-0.606 m	2.391 m	S_3
9	0.141 s	2215.229 s	2.265 m	2.643 m	S_3
10	0.128 s	454.929 s	1.207 m	-2.670 m	S_1

Table 4: Test results for combined trajectory 1 for the search of the target sized 0.1% of the rectangular search area.

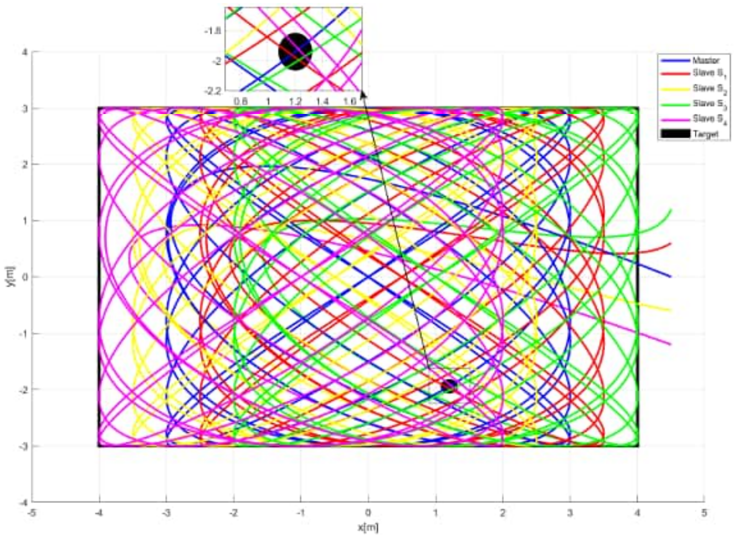


Figure 14: Target search with the combined trajectory 2 for a target sized 0.1% of the rectangular search area.

We perform ten tests with this combination for target detection and the results are shown in Table 5.

With the combined trajectory 2 of these two Lissajous curves, it is observed that some of the trajectories pass through the zones that they did not pass before (Figure 12); and the target was successfully detected in all ten tests, as shown in Table 5.

In the last table, only 10 of the 100 tests that were carried out were recorded, of which only in one case, the randomly placed target could not be detected.

The target search time is measured from the moment all five quadrotors start to fly

Test	Run time	Detection time	x-axis position	y-axis position	Target detected by
1	0.472 s	865.11 s	2.439 m	2.334 m	S_2
2	0.293 s	792.091 s	-2.891 m	2.334 m	M
3	0.154 s	1075.385 s	1.026 m	-2.315 m	S_2
4	0.140 s	60 s	-1.717 m	0.269 m	S_1
5	0.155 s	286.197 s	3.546 m	2.674 m	S_1
6	0.139 s	834.347 s	-2.654 m	-2.707 m	S_2
7	0.138 s	1142.108 s	2.328 m	-2.060 m	S_3
8	0.148 s	888.934 s	-0.606 m	2.391 m	S_3
9	0.152 s	2215.229 s	2.265 m	2.643 m	S_3
10	0.173 s	189.443 s	1.207 m	-2.670 m	M

Table 5: Test results for combined trajectory 2 for the search of the target sized 0.1% of the rectangular area.

and till the moment when any of them finds the object within the search area, as shown in Figure 15.

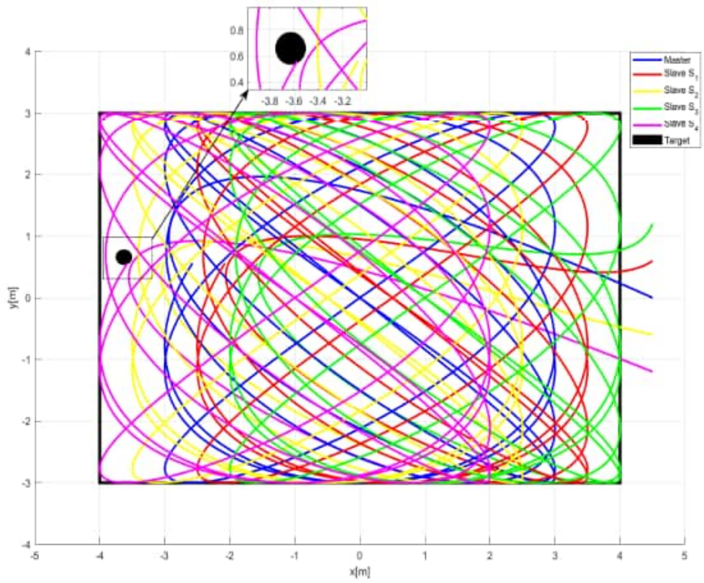


Figure 15: Target search and detection time for the formation of quadrotors.

Figure 15 shows that the target was detected by a slave quadrotor within the rectangular search area. The position of the center of the target was $(x, y) = (-3.628$ m, 0.659 m) and it was found in 12,877.043 seconds.

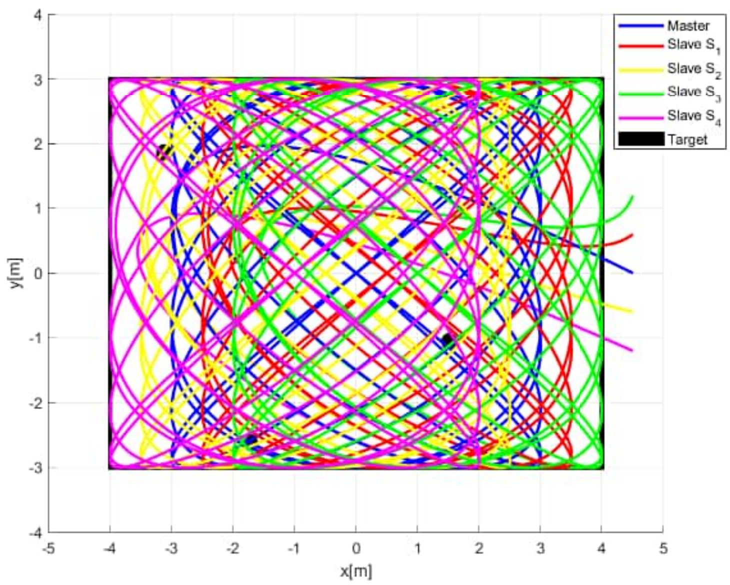


Figure 16: Three targets detected by the formation of quadrotors.

6.2 Three targets detection

In Figure 16, the detection of three targets placed randomly within the rectangular search surface is simulated and it is observed that at least one of the trajectories of the quadrotors passes through any of them, so it is assumed that they were detected.

In order to determine the percentage of the exploration area that is covered by the trajectories of the quadrotors formation, it is divided into 3136 squares of equal size and those that are visited by some quadrotor are noted, see Figure 17. The size of each square was determined to be smaller than the target to be detected. Using this method for both combinations of trajectories, a percentage of 84.27% of the covered area was obtained for the first case, and 92.12% for the second, demonstrating that the latter is more suitable for finding targets. With these results, it is concluded that the trajectories could easily find almost any targets in the rectangular search area, including small ones.

6.3 Discussion

The used exploration area is a rectangle which is 6 meters wide and 8 meters long. It is explored with the formation of five quadrotors. At first, they formed a vertical line and followed a desired circular trajectory, where it was observed that there were large unexplored zones. If none of the trajectories of the quadrotors passes over that zone, it is considered as not explored. Approximately 48% of the total search area remained unexplored, predominantly the central zone.

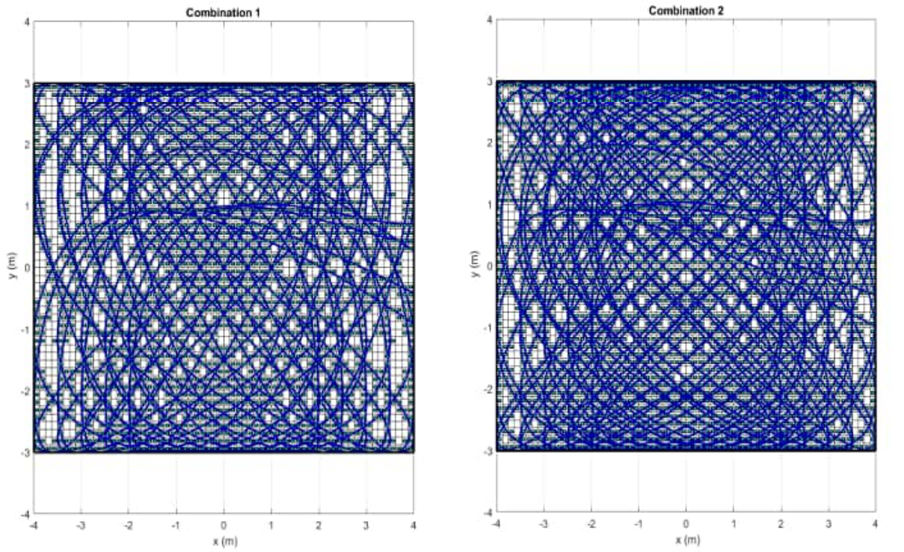


Figure 17: Quadcopters formation training trajectories coverage: a) by combined trajectory 1 and b) by combined trajectory 2.

Some tests were also carried out with other trajectories by using different parameters for the Lissajous curves, and it was observed that the unexplored zones were reduced. Different trajectories were made, taking the densest one, thus decreasing the unexplored zones between the quadrotors, with 4 % of the total unexplored surface.

Various formations were applied in the search and detection of the target. The first formation was a vertical line in which the slave quadrotors were observed to leave the boundaries of the search surface. The "V" formation was also used and it was observed that there were unexplored zones, mainly in the upper corners of the search surface. Therefore it was decided to use the formation in a horizontal line to explore this surface.

Next, the search surface was explored by using now two combinations of the Lissajous curves with different parameters. Ten tests for each combination were carried out. Because the target was found every time, another 90 tests were made (100 in total for each one). In the first case, comparing with the trajectory in Figure 13, an efficiency of 96% was reached. Exploring the search surface with the second combination, an efficiency of 99% was obtained.

7 Conclusions

In this paper, we have presented the formation problem of multiple UAVs for applications to search and detection by tracking time-variable trajectories. The main contributions of this work are: the mentioned formation was obtained by using complex systems theory and backstepping nonlinear control. We made a comprehensive study of UAVs formation coverage for search and detection of a random target within the search zone by tracking time-variable trajectories. The reported numerical results show that the methodology employed meets the purpose of detecting the target within the search zone and reducing the unexplored zones by the combination of time-varying trajectories.Thus, in future work, we plan the physical implementation of the proposed formation scheme, we will use switching trajectories and switching topology connection, as well as chaotic trajectories.

Acknowledment

This work was supported by CONAHCyT, México under Research Grand A1-S-31628.

References

[1] R. Abbas and Q. Wu. Formation tracking for multiple quadrotor based on sliding mode and fixed communication topology. In: *2013 5th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)* **2** (2013) 233–238.

[2] M. E. Antonio-Toledo and E. N. Sanchez, A. Y. Alanis, J. A. Flórez and M. A. Perez-Cisneros. Real-time integral backstepping with sliding mode control for a quadrotor UAV. *IFAC-PapersOnLine* **13** (2018) 549–554.

[3] S. Bouabdallah, P. Murrieri and R. Siegwart Design and control of an indoor micro quadrotor. In: *Robotics and Automation, 2004. Proceedings. ICRA'04 2004 IEEE International Conference on Robotics and Automation* **5** (2004) 4393–4398.

[4] S. Bouabdallah and R. Siegwart. Backstepping and sliding-mode techniques applied to an indoor micro quadrotor. In: *Robotics and Automation, 2005. ICRA 2005 Proceedings of the 2005 IEEE International Conference on Robotics and Automation* (2005) 2247–2252.

[5] S. Bouabdallah. *Design and Control of Quadrotors with Application to Autonomous Flying.* Ecole Polytechnique Federale de Lausanne, Lausana, 2007.

[6] I.D. Couzin, J. Krause, R. James, G. D. Ruxton and N. R. Franks. Collective memory and spatial sorting in animal groups. *Journal of Theoretical Biology* **218** (2002) 1–11.

[7] Xi. Dong and Y. Zhou, Z. Ren and Y. Zhong. Time-varying formation tracking for second-order multi-agent systems subjected to switching topologies with application to quadrotor formation flying. *IEEE Transactions on Industrial Electronics* **64** (2017) 5014–5024.

[8] J. Flemming and A. Hornes. Lissajous-like figures with triangular and square waves. *Revista Brasileira de Ensino de Física* **35** (2013) 3702.

[9] P. P. Flores, P. Castillo and F. Castanos. Backstepping-Based Controller for Flight Formation. In: *2019 International Conference on Unmanned Aircraft Systems (ICUAS)* (20179) 263–269.

[10] P. K. Freeman and R. S. Freeland. Agricultural UAVs in the US: potential, policy and hype. *Remote Sensing Applications: Society and Environment* **2** (2015) 35–43.

[11] C. Godsil and G. F. Royle. *Algebraic Graph Theory. Springer Science & Business Media*, **207**, 2001.

[12] D. Ivanov, I. Kalyaev and S. Kapustyan. Method of circles for solving formation task in a group of quadrotor UAVs. In: *Systems and Informatics (ICSAI), 2014 2nd International Conference on Systems and Informatics* (2014) 236–240.

[13] A. King. Technology: The future of agriculture. *Nature* **544** (2017) S21–S23.

[14] N. Koksal, B. Fidan and K. Buyukkabasakal. Real-time implementation of decentralized adaptive formation control on multi-quadrotor systems. In: *European Control Conference (ECC), IEEE* (2015) 3162–3167.

[15] M. Krstic, I. Kanellakopoulos and P. V. Kokotovic. *Nonlinear and Adaptive Control Design. Wiley*, 1995.

[16] J. D. Lawrence. *A Catalog of Special Plane Curves*. Courier Corporation, 2013.

[17] Q. Ouyang, Z. Wu, Y. Cong and Z. Wang. Formation control of unmanned aerial vehicle swarms: A comprehensive review. *Asian Journal of Control* (2023) **25** 570–593.

[18] D.J. Sumpter. The principles of collective animal behaviour. *Philosophical Transactions of the Royal Society B: Biological Sciences* (2006) **361** 5–22.

[19] D.J. Sumpter. *Collective Animal Behavior*. Princeton University Press, 2010.

[20] A. Swarup and Sudhir. Comparison of Quadrotor Performance Using Backstepping and Sliding Mode Control. In: *International Conference on Circuits, Systems and Control* **1** (2014) 79–82.

[21] T. Han, M. Chi, Z. Guan, B. Hu, J. Xiao and Y. Huang. Distributed three-dimensional formation containment control of multiple unmanned aerial vehicle systems. *Asian Journal of Control* **19** (2017) 1103–1113.

[22] S. T. Thornton and J. B. Marion. *Classical Dynamics of Particles and Systems*. Cengage Learning, 2021.

Nonlinear Dynamics and Systems Theory, **25** (3) (2025) 338–348

Estimation of Pitch Angle and Heave Position of Remotely Operated Vehicle Using Linear Quadratic Gaussian

A. Suryowinoto [1], T. Herlambang [2*], P. Triwinanto [3], Y. A. Prabowo [1], D. Rahmalia [4] and M. S. Baital [5]

[1] *Electrical Engineering Department, Adhi Tama Institute of Technology, Indonesia.*
[2] *Department of Information System, Universitas Nahdlatul Ulama Surabaya, Indonesia.*
[3] *National Research and Innovation Agency, Indonesia.*
[4] *Department of Mathematics, University of Islam Darul Ulum, Indonesia.*
[5] *Shipbuilding Construction Engineering Technology, Vocational School, Diponegoro University.*

Received: July 21, 2024; Revised: May 20, 2025

Abstract: In a Remotely Operated Vehicle (ROV), there are six degrees of freedom: surge, sway, heave, roll, pitch, and yaw. In linear motions, there are surge, sway and heave. In angular rotations, there are roll, pitch and yaw. In practice, there are disturbances and noise in the linear motion and angular rotation in their measurements. Therefore, in this research, the estimation of pitch angle and heave position of a ROV will be carried out by Linear Quadratic Gaussian (LQG). LQG is used for optimal control when there are disturbance input and measurement noise in the plant model. From the simulation, the estimation of the state solution and optimal control with various noise can be compared by LQG.

Keywords: *ROV; pitch angle; heave position; LQG.*

Mathematics Subject Classification (2020): 37N35, 49N10.

* Corresponding author: mailto:teguh@unusa.ac.id

1 Introduction

Seventy percent of Indonesia is covered by sea, so Indonesia has significant marine resource. For exploring marine resource, a Remotely Operated Vehicle (ROV) with its control is required. In a ROV, there are six degrees of freedom: surge, sway, heave, roll, pitch, and yaw. In linear motions, there are surge, sway and heave. Surge, sway, and heave are linear motions on the x-axis, y-axis, and z-axis, respectively. In angular rotations, there are roll, pitch and yaw. Roll, pitch, and yaw are angular rotations on the x-axis, y-axis, and z-axis, respectively [1].

In practice, there are disturbances and noise in the linear motion and angular rotation in their measurements. Therefore, in this research, the estimation of pitch angle and heave position of ROV will be carried out by Linear Quadratic Gaussian (LQG). Linear Quadratic Gaussian (LQG) control refers to an optimal control problem where the plant model is linear with a white noise disturbance input and white measurement noise. LQG is used for optimal control when there are a disturbance input and measurement noise in the plant model. Generally, optimal control is used to find the control minimizing the performance index and result in the trajectory of state solution [2].

From the previous research, an AUV has been developed by the method of Linear Quadratic Regulator (LQR) minimizing the control with no reference [3], Linear Quadratic Tracking (LQT) minimizing the control included to the tracking path [4], PID control for determining the response [5]. The estimation method is also established by the Kalman Filter and H-infinity method [6]. Besides the AUV, the LQR and LQT are also applied in the inverted pendulum [7], missile and projectile [8], mobile robot [9] and human arm models [10].

The dynamical model of LQG is almost similar to the LQR, however in the LQG model, there are distubance input and measurement noise. To determine the estimation of the state solution, we need a Riccati solution and feedback gain from the LQR and Kalman Gain. LQG is also used for optimizing the given performance index. To run the simulation, the calculation of LQR should be made before. Then the Riccati solution and feeback gain of LQR are brought to the LQT as variables. From the simulation, the estimation of state solution of pitch and heave and optimal control with various noise can be compared by LQG with small mean square error (MSE).

2 Mathematical Model of ROV

There are some motions and rotations in a ROV which will be constructed as a mathematical model. In linear motions such as surge, sway, and heave, there are motions on the x-axis, y-axis, and z-axis, respectively. In angular rotations such as roll, pitch, and yaw are rotations on the x-axis, y-axis, and z-axis, respectively. The mathematical model of ROV is shown in Figure 1. According to Newton's Law, the mathematical model of ROV [5] is

$$X = m\left[i - vr + wq - x_G\left(q^2 + r^2\right) + y_G\left(pq - \dot{r}\right) + z_G\left(pr + \dot{q}\right)\right], \tag{1}$$

$$Y = m\left[\dot{v} + ur - wp + x_G\left(pq + \dot{r}\right) - y_G\left(p^2 + r^2\right) + z_G\left(qr - \dot{p}\right)\right], \tag{2}$$

$$Z = m\left[\dot{w} - uq + vp + x_G(vp - \dot{q}) + y_G(qr + \dot{p}) - z_G(p^2 + q^2)\right], \tag{3}$$

$$\begin{aligned} K = {} & I_x\dot{p} + (I_x - I_y)qr + I_xy(pr - \dot{q}) - I_yz(q^2 - r^2) - I_xz(pq + \dot{r}) \\ & + m\left(y_G(\dot{w} - uq + vp) - z_G(\dot{v} + ur - wp)\right), \end{aligned} \tag{4}$$

$$M = I_y\dot{q} + (I_x - I_z)pr - I_x y(qr - \dot{p}) - I_y z(p1 - \dot{r}) - I_x z(p^2 - r^2) \\ - m\left(x_G(\dot{w} - uq + vp) - z_G(\dot{u} - vr + w1)\right), \tag{5}$$

$$N = I_z\dot{r} + (I_y - I_x)pq - I_x y(p^2 - q^2) - I_y z(pr - \dot{q}) - I_x z(qr - \dot{p}) \\ - m\left(x_G(\dot{v} + ur - wp) - y_G(\dot{u} - vr + wq)\right). \tag{6}$$

The model parameters are: m is the ROV mass, I_x, I_y, I_z are the inertia moments on the x-axis, y-axis, and z-axis, respectively, x_G, y_G, z_G are the positions of gravity center on the x-axis, y-axis, and z-axis, respectively. The other notations are:

X : surge force	x : surge position	u : surge velocity
Y : sway force	y : sway position	v : sway velocity
Z : heave force	z : heave position	w : heave velocity
K : roll moment	ϕ : roll angle	p : roll rate
M : pitch moment	θ : pitch angle	q : pitch rate
N : yaw moment	ψ : yaw angle	r : yaw rate

The problem is only limited for linear motion and angular rotation. By removing surge, sway, roll and yaw, the variables are $v = r = p = \varphi = \psi = y = 0$. The state solutions are the pitch angle θ, pitch velocity q, heave position z, and heave velocity w.

The simplicity of the model of ROV can be explained as follows.

As the pitch moment equation, take equation (5): $M = I_y\dot{q} + (I_x - I_z)pr - I_x y(qr + \dot{p}) - I_y z(pq - \dot{r}) - I_x z(p^2 - r^2) - m\left(x_G(\dot{w} - uq + vp) - z_G(\dot{u} - vr + wq)\right)$.

By removing surge, sway, roll and yaw, the variables are $v = r = p = \varphi = \psi = y = 0$. The pitch moment M can be elaborated as $M = M_q\dot{q} + M_q q + M_{\dot{w}}\dot{w} + M_w w$, then the pitch moment equation is

$$M_{\dot{q}}\dot{q} + M_q q + M_{\dot{w}}\dot{w} + M_w w = I_y\dot{q} - mx_G\dot{w} + mx_G uq + mz_G wq, \tag{7}$$

where $M_{\dot{q}}$ is the added inertia mass moment related to the pitch rate, $M_{\dot{w}}$ is the added inertia mass moment related to the heave velocity, M_q is the pitch moment coefficient induced by the pitch rate, M_w is the pitch moment coefficient induced by the heave velocity.

As the heave force equation, take equation (3):
$Z = m\left[\dot{w} - uq + vp + x_G(vp - \dot{q}) + y_G(qr + \dot{p}) - z_G(p^2 + q^2)\right]$. By removing surge, sway, roll and yaw, the variables are $v = r = p = \varphi = \psi = y = 0$. The heave force Z can be elaborated as $Z = Z_{\dot{q}\dot{q} + Z_q q + Z_{\dot{w}}\dot{w} + Z_w w}$, then the heave force equation is

$$Z_{\dot{q}}\dot{q} + Z_q q + Z_{\dot{w}}\dot{w} + Z_w w = m\dot{w} - muq - mx_G\dot{q} - mz_G q^2, \tag{8}$$

where $Z_{\dot{q}}$ is the added mass related to the pitch rate, $Z_{\dot{w}}$ is the added mass related to the heave velocity, Z_q is the heave force coefficient induced by pitch rate, Z_w is heave force coefficient induced by the heave velocity.

From equations (7) and (8), the nonlinear system of ROV is

$$\dot{\theta} = q, \tag{9}$$

$$(M_{\dot{q}} - I_y)\dot{q} + (M_{\dot{w}} + mx_G)\dot{w} = (z_G W - z_B)\theta + mx_G uq + mz_G wq - M_q q - M_w w + M_\delta\delta, \tag{10}$$

$$\dot{z} = w\cos\theta - u\sin\theta, \tag{11}$$

$$(Z_{\dot{q}} + mx_G)\dot{q} + (Z_{\dot{w}} - m)\dot{w} = -muq - mz_G q^2 - Z_q q - Z_w w - Z_\delta \delta. \tag{12}$$

By the linearization $\sin\theta \approx, \cos\theta \approx 1$ and the Jacobian around the equilibrium point $(\theta^*, q^*, z^*, w^*) = (0, 0, 0, 0)$, the linear form is

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & (M_{\dot{q}} & 0 & (M_{\dot{w}} + mx_G) \\ 0 & 0 & 1 & 0 \\ 0 & (Z_{\dot{q}} + mx_G) & 0 & (Z_{\dot{w}} - m) \end{bmatrix} \begin{bmatrix} \dot{\theta} \\ \dot{q} \\ \dot{z} \\ \dot{w} \end{bmatrix} =$$
$$\begin{bmatrix} 0 & (z_G W - z_B) & -u & 0 \\ 1 & -(M_q - mx_G u) & 0 & -M_w \\ -u & 0 & 0 & 1 \\ 0 & -(Z_q + mu) & 0 & -Z_w \end{bmatrix} \begin{bmatrix} \theta \\ q \\ z \\ w \end{bmatrix} + \begin{bmatrix} 0 \\ -Z_\delta \\ M_\delta \\ 0 \end{bmatrix} \delta. \tag{13}$$

3 Methods

3.1 Linear Quadratic Regulator (LQR)

The Linear Quadratic Regulator is used to find the control minimizing the performance index and results in the trajectory of state solution. Some steps of the LQR are computing the solution of the Riccati Equation $P(t)$ in equation (14) and the feedback gain $K(t)$ in equation (15). From the result of the feedback gain, we can compute the solution of state space and optimal control [11].

$$\dot{P} = -P(t)A - A^T P(t) - Q + P(t) B_u R^{-1} B_u^T P(t), P(t_f) = H, \tag{14}$$

$$K(t) = R^{-1} B_u^T P(t). \tag{15}$$

3.2 Linear Quadratic Gaussian (LQG)

3.2.1 State space model

The state space model of LQG is similar to the LQR model, however there is additional noise so that the state space model of LQG [12] is

$$\dot{x}(t) = Ax(t) + B_u u(t) + B_w w(t), \tag{16}$$

$$m(t) = C_m x(t) + v(t), \tag{17}$$

$$E[w(t) w^T(t+\tau)] = S_w \delta(\tau), \tag{18}$$

$$E[v(t) v^T(t+\tau)] = S_v \delta(\tau), \tag{19}$$

$$E[v(t) w^T(t+\tau)] = 0. \tag{20}$$

Here, equation (16) is the plant model, equation (17) is the measurement and equations (18) to (20) are the noise model of plant and measurement. The performance index which will be optimized is

$$J = \frac{1}{2} E\left[x^T(t_f) H x(t_f) + \int_0^{t_f} \left(x^T(t) Q x(t) + u^T(t) R u(t) \right) \right], \tag{21}$$

where H and Q are positive semidefinite and R is positive definite.

3.2.2 The development of the Riccati equation

The development of the Riccati equation can be computed in equation (22) with the initial condition $F_e(0) = 0$. After the solutions are obtained, the solution of the Riccati equation can be applied in computing the Kalman Gain in equation (23).

$$\dot{F}_e(t) = AF_e(t) + F_e(t)A^T + B_w S_w B_w^T - F_e(t)C_m^T S_v^{-1} C_m F_e(t), F_e(0) = 0, \tag{22}$$

$$G(t) = F_e(t)C_m^T S_v^{-1}. \tag{23}$$

3.2.3 Kalman filter equation

To estimate the state solution, we use the Kalman Filter equation in equations (24) and (25) using the feedback gain from the LQR and Kalman Gain in equation (26).

$$\dot{\hat{x}}(t) = A\hat{x}(t) + B_u u(t) + G(t)(m(t) - C_m \hat{x}(t)), \tag{24}$$

$$\dot{\hat{x}}(t) = [A - G(t)C_m - B_u K(t)]\hat{x}(t) + G(t)m(t), \tag{25}$$

$$u(t) = -K(t)\hat{x}(t). \tag{26}$$

4 Results and Discussion

4.1 Model of ROV

The model of ROV used is as in equations (27) and (28), where θ is the pitch angle, w is the heave velocity, q is the pitch rate, and z is the heave position. The simulations will be run with various rates of S_w and S_v. The solution of state space and its estimation, i.e., the pitch angle and heave position, and optimal control will be obtained as comparison.

$$\begin{bmatrix} \dot{\theta} \\ \dot{w} \\ \dot{q} \\ \dot{z} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0.0175 & -1.273 & -3.559 & 0 \\ -0.052 & 1.273 & -2.661 & 0 \\ -5 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \theta \\ w \\ q \\ z \end{bmatrix} + \begin{bmatrix} 0 \\ 0.085 \\ 21.79 \\ 0 \end{bmatrix} \delta, \tag{27}$$

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \theta \\ w \\ q \\ z \end{bmatrix}. \tag{28}$$

4.2 Simulation 1

In this simulation, the values of $S_w = 0.01$ and $S_v = 0.01$ will be used and the result of feedback gain can be seen in Figure 1, the Kalman gain can be seen in Figure 2. The state solution and its estimation, i.e., the pitch angle and heave position, can ben seen in Figure 3 and optimal control can be seen in Figure 4.

Figure 1 describes the solution of the feedback gain from the LQR with given final conditions being zero. The feedback gain is obtained from the Riccati solution which is computed backwardly. Figure 2 describes the solution of the Kalman gain with given initial condition being zero.

The Kalman gain is obtained from development of the Riccati solution which is computed forwardly. Figure 3 shows the state solution and its estimation of the pitch angle

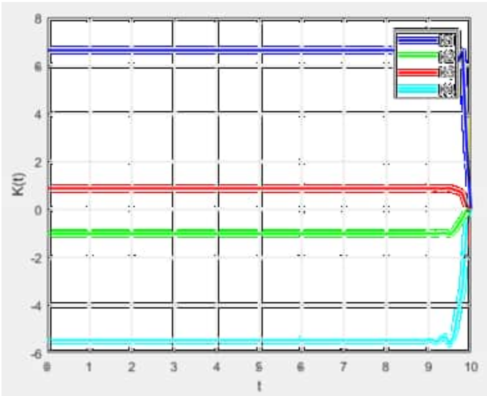


Figure 1: The Solution of the Feedback Gain.

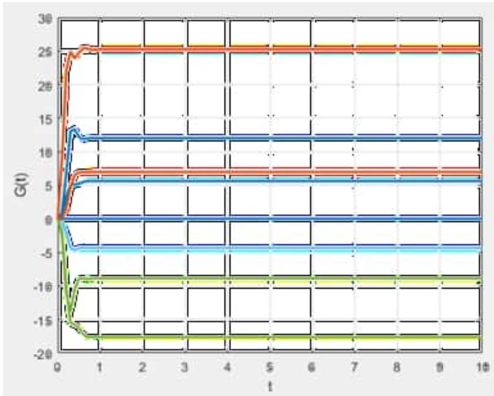


Figure 2: The Solution of the Kalman Gain.

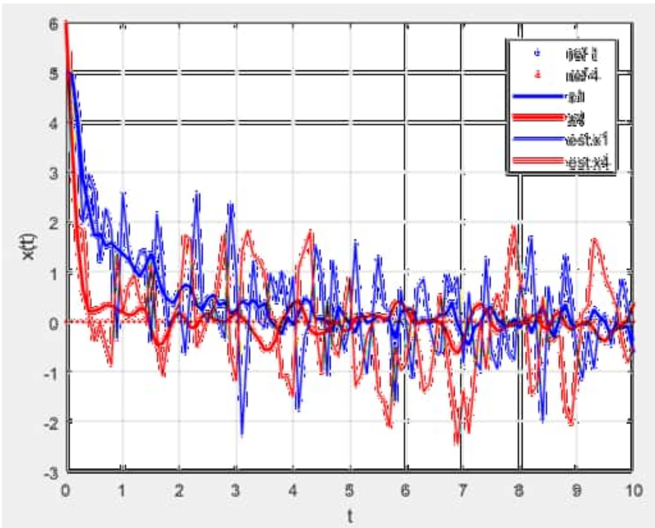


Figure 3: The State Solution and Estimation with Large Noise.

denoted by x_1 and the heave position denoted by x_4. The RMSE of the estimation is 0.9268 for the pitch angle estimation and 1.0646 for the heave position estimation. Figure 4 represents the optimal control and its solution with RMSE being 16.0780.

4.3 Simulation 2

In this simulation, the values of $S_w = 0.001$ and $S_v = 0.001$ will be used and the result of the feedback gain can be seen in Figure 5, the Kalman gain can be seen in Figure 6. The state solution and its estimation, i.e., the pitch angle and heave position, can be seen in Figure 7 and optimal control can be seen in Figure 8.

Figure 5 describes the solution of the feedback gain from the LQR with given final conditions being zero. The feedback gain is obtained from the Riccati solution which is

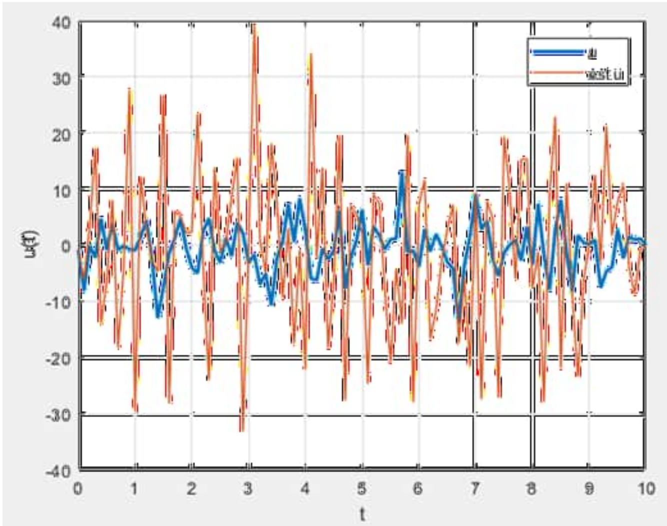


Figure 4: The Optimal Control and Estimation with Large Noise.

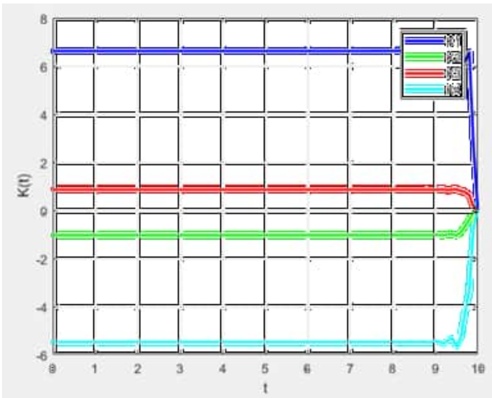


Figure 5: The Solution of the Feedback Gain.

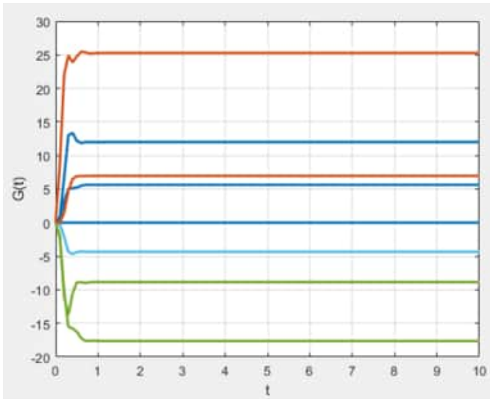


Figure 6: The Solution of the Kalman Gain.

computed backwardly. Figure 6 describes the solution of the Kalman gain with given initial condition being zero. The Kalman gain is obtained from development of the Riccati solution which is computed forwardly. Figure 7 shows state solution and its estimation of the pitch angle denoted by x_1 and the heave position denoted by x_4. The RMSE of the estimation is 0.2473 for the pitch angle estimation and 0.2652 for the heave position estimation. Figure 8 represents the optimal control and its solution with RMSE being 4.6209.

4.4 Simulation 3

In this simulation, the values of $S_w = 0.0001$ and $S_v = 0.0001$ will be used and the result of the feedback gain can be seen in Figure 9, the Kalman gain can be seen in Figure 10.

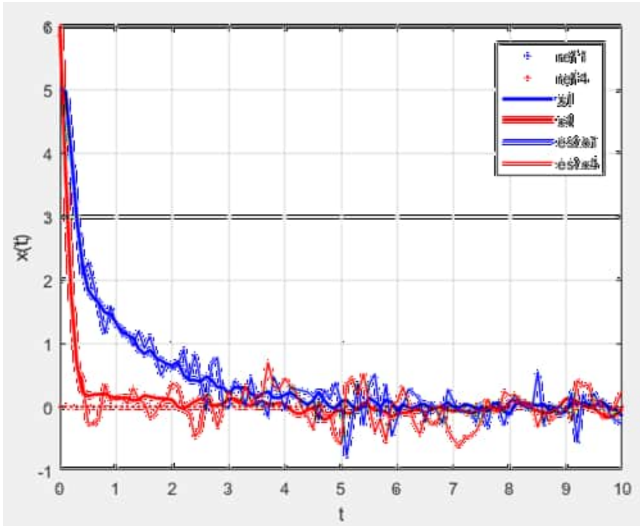


Figure 7: The State Solution and Estimation with Moderate Noise.

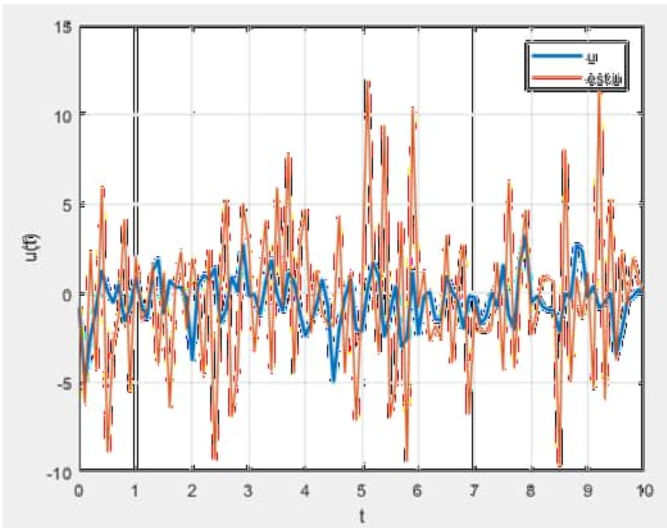
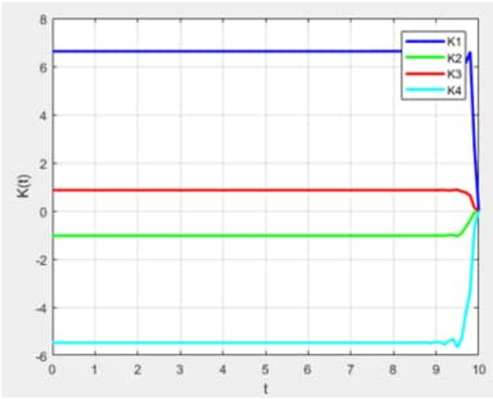


Figure 8: The Optimal Control and Estimation with Moderate Noise.

The state solution and its estimation, i.e., the pitch angle and heave position, can be seen in Figure 11 and optimal control can be seen in Figure 12.

Figure 9 describes the solution of the feedback gain from the LQR with given final conditions being zero. The feedback gain is obtained from the Riccati solution which is computed backwardly. Figure 10 describes the solution of the Kalman gain with given initial condition being zero. The Kalman gain is obtained from development of the Riccati solution which is computed forwardly. Figure 11 shows the state solution and



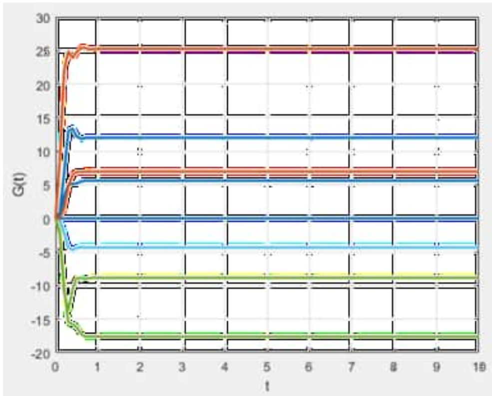


Figure 9: The Solution of the Feedback Gain. **Figure 10**: The Solution of the Kalman Gain.

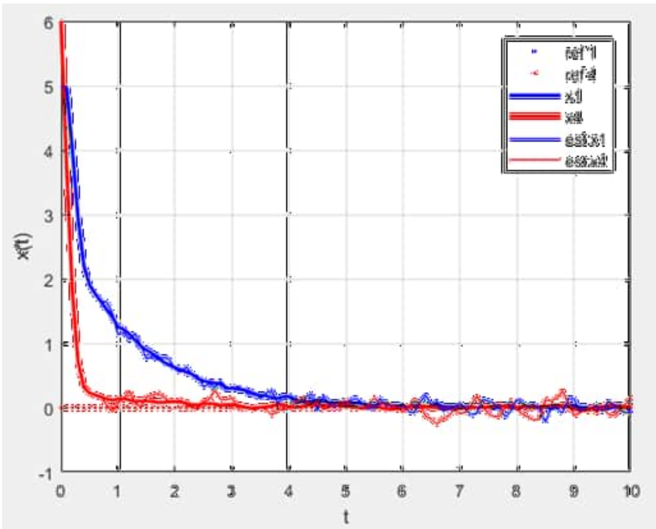


Figure 11: The State Solution and Estimation with Small Noise.

its estimation of the pitch angle denoted by x_1 and the heave position denoted by x_4. The RMSE of the estimation is 0.0762 for the pitch angle estimation and 0.0956 for the heave position estimation. Figure 12 represents the optimal control and its solution with RMSE being 1.3031.

5 Conclusion

The estimation of the pitch angle and heave position of a ROV is carried out by Linear Quadratic Gaussian (LQG). From the simulation, the estimation of the state solution and optimal control with various noise can be compared by LQG. From the results of three simulations, the small values of S_w and S_v can give better estimation with a small value of RMSE in both state solution and optimal control.

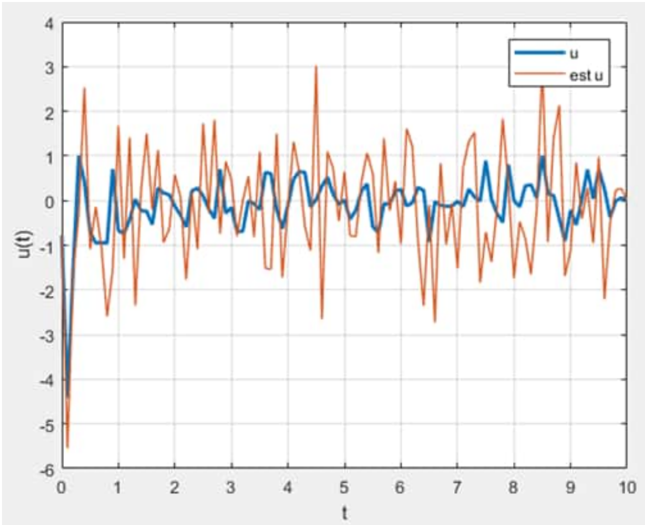


Figure 12: The Optimal Control and Estimation with Small Noise.

The developments of this research are the application of LQG to different objects besides the ROV. Moreover, the optimization of weight matrices in the performance index can be done by available methods.

6 Acknowledgement

High appreciation to the Kemdikbudristek for financial support for the completion of the research conducted in the year of 2024 with contract number 09/E5/PG.02.00.PL/2024, 083/SP2H/PT/LL7/2024, and 01/KP/LPPM/ITATS/2024.

References

[1] T. Herlambang, E. B. Djatmiko and H. Nurhadi. Ensemble Kalman filter with a square root scheme (EnKF-SR) for trajectory estimation of AUV SEGOROGENI ITS. *International Review of Mechanical Engineering (I. RE. ME)* **6** (9) (2015) 553–560.

[2] J. B. Burl. *Linear Optimal Control*. Addison Wesley, California, 1999.

[3] N. Syahroni and J. W. Choi. An Autonomous Underwater Vehicle Simulation Using Linear Quadratic Servo Based on Open Control Platform. *Modelling and Simulation in Engineering* (2012) Article ID 291318, 4 pages.

[4] T. Herlambang, D. Rahmalia, H. Nurhadi and D. Adzkiya. Optimization of Linear Quadratic Regulator with Tracking Applied to Autonomous Underwater Vehicle (AUV) Using Cuckoo Search. *Nonlinear Dynamics and Systems Theory* **20** (3) (2020) 282–298.

[5] T. Herlambang, D. Rahmalia and T. Yulianto. Particle Swarm Optimization (PSO) and Ant Colony Optimization (ACO) for Optimizing PID Paramaters on Autonomous Underwater Vehicle (AUV) Control System. In: *ICCGANT, Journal of Physics: Conference Series* **1211** (1) IOP Publishing (2019) 012039.

[6] T. Herlambang, D. Rahmalia, and A. Suryowinoto. H-Infinity for Autonomous Surface Vehicle Position Estimation. In: *ICCGANT, Journal of Physics: Conference Series* **2157** (1) IOP Publishing (2022) 012022.

[7] K. Ogata. *Modern Control Engineering.* Prentice Hall, New Jersey, 2022.

[8] T. Herlambang, D. Rahmalia and D. F. Karya. Optimal Control Model of Two Dimensional Missile Using Forward Backward Sweep Method (FBSM). In: *International Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE)*, IEEE (2021) 12–17.

[9] T. Herlambang, A. Muhith, D. Rahmalia, and H. Nurhadi. Motion Optimization Using Modified Kalman Filter for Invers-Kinematics based Multi DOF Arm Robot. *International Journal of Control and Automationt* **13** (2s) (2020) 64–71.

[10] T. Herlambang, D. Rahmalia, H. Nurhadi, and A. Suryowinoto. Estimation of shoulder and elbow joint angle from Linear Quadratic Tracking based on six muscle force vectors. In: *ICCGANT, Journal of Physics: Conference Series.* **2157** (1) IOP Publishing (2022) 012024.

[11] D. S. Naidu. *Optimal Control System.* CRC Press, Florida, 2003.

[12] F. L. Lewis, L. Xie and D. Popa. *Optimal and Robust Estimation.* CRC Press, London, 2008.