JOURNAL EDITORIAL BOARD

ISSN 2327-4352 (Print)  ISSN 2327-4379 (Online)
http://www.scirp.org/journal/jamp

Editor-in-Chief
Prof. Wen-Xiu Ma
University of South Florida, USA

Editorial Board (According to Alphabet)

Dr. Izhari Ahmad
King Fahd University of Petroleum and Minerals, Saudi Arabia

Dr. S. Joseph Antony
University of Leeds, UK

Prof. Roberto Oscar Aquilano
Instituto de Física Rosario, Argentina

Prof. Ping-Hei Chen
National Taiwan University, Chinese Taipei

Prof. Wanyang Dai
Nanjing University, China

Dr. Steven B. Damelin
The American Mathematical Society, USA

Prof. Beih El-Sayed El-Desouky
Mansoura University, Egypt

Prof. Chaudry Masood Khalique
North-West University, South Africa

Dr. Ki Young Kim
Samsung Advanced Institute of Technology, South Korea

Prof. Xiang Li
Beijing University of Chemical Technology, China

Prof. Xing Lü
Beijing Jiaotong University, China

Dr. Jafar Fawzi Mansi Al Omari
Al-Balqa’ Applied University, Jordan

Prof. Rosa Pardo
Complutense University of Madrid, Spain

Prof. Sanzheng Qiao
McMaster University, Canada

Dr. Daniele Ritelli
University of Bologna, Italy

Dr. Babak Daneshvar Rouyendeh
Atılım University, Turkey

Prof. Morteza Seddighin
Indiana University East, USA

Dr. Marco Spadini
University of Florence, Italy

Prof. Hari Mohan Srivastava
University of Victoria, Canada

Dr. Divine Tito Fongha Wanduku
Keiser University, USA

Prof. Ping Wang
Penn State University Schuylkill, USA

Prof. Xiaohui Yuan
Huazhong University of Science and Technology, China
# Table of Contents

**Volume 5  Number 6  June 2017**

**Oscillation and Asymptotic Behaviour of Solutions of Nonlinear Two-Dimensional Neutral Delay Difference Systems**  
K. Thangavelu, G. Saraswathi

**A Multi-MW Proton/Electron Facility at KEK**  
R. Belusevic

**Modeling for Collapsing Cavitation Bubble near Rough Solid Wall by Multit-Relaxation-Time Pseudopotential Lattice Boltzmann Model**  
M. L. Shan, Y. P. Zhu, C. Yao, Q. B. Han, C. P. Zhu

**Non-Native English Speaker Readability Metric: Reading Speed and Comprehension**  
A. Boudjella, M. Sharma, D. Sharma

**Optimal Distributed Control Problem for the $b$-Equation**  
C. Y. Shen

**A Second Order Characteristic Mixed Finite Element Method for Convection Diffusion Reaction Equations**  
T. J. Sun

**A-Equation and Its Connections to Nonlinear Integrable System**  
Y. L. Zhang

**Approximate and Invariant Solutions of a Mathematical Model Describing a Simple One-Dimensional Blood Flow of Variable Density**  
R. Ibragimov, N. Ibragimov, P. Mohazzabi

**Supply Chain Finance Decision Analysis with a Partial Credit Guarantee Contract**  
Y. L. Su, B. Y. Zhong

**N-Rotating Loop-Soliton Solution of the Coupled Integrable Dispersionless Equation**  
S. Abbagari, S. Youssoufa, H. T. Tchokouansi, V. K. Kuette, T. B. Bouetou, T. C. Kofane

**Influence of the Foundation on the Threshold of Stability for Rotating Machines with Roller Bearings—A Theoretical Analysis**  
U. Werner

**$PP$ and $P\bar{P}$ Multi-Particles Production Investigation Based on CCNN Black-Box Approach**  
E.-S. A. El-Dahshan
Journal of Applied Mathematics and Physics (JAMP)

Journal Information

SUBSCRIPTIONS

The *Journal of Applied Mathematics and Physics* (Online at Scientific Research Publishing, www.SciRP.org) is published monthly by Scientific Research Publishing, Inc., USA.

Subscription rates:
Print: $39 per issue.
To subscribe, please contact Journals Subscriptions Department, E-mail: sub@scirp.org

SERVICES

Advertisements
Advertisement Sales Department, E-mail: service@scirp.org

Reprints (minimum quantity 100 copies)
Reprints Co-ordinator, Scientific Research Publishing, Inc., USA.
E-mail: sub@scirp.org

COPYRIGHT

Copyright and reuse rights for the front matter of the journal:
Copyright © 2017 by Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY).
http://creativecommons.org/licenses/by/4.0/

Copyright for individual papers of the journal:
Copyright © 2017 by author(s) and Scientific Research Publishing Inc.

Reuse rights for individual papers:
Note: At SCIRP authors can choose between CC BY and CC BY-NC. Please consult each paper for its reuse rights.

Disclaimer of liability
Statements and opinions expressed in the articles and communications are those of the individual contributors and not the statements and opinion of Scientific Research Publishing, Inc. We assume no responsibility or liability for any damage or injury to persons or property arising out of the use of any materials, instructions, methods or ideas contained herein. We expressly disclaim any implied warranties of merchantability or fitness for a particular purpose. If expert assistance is required, the services of a competent professional person should be sought.

PRODUCTION INFORMATION

For manuscripts that have been accepted for publication, please contact:
E-mail: jamp@scirp.org
Abstract
This paper deals with the some oscillation criteria for the two-dimensional neutral delay difference system of the form
\[ \Delta(x_n + p_n x_{n-k}) = b_n y_n, \Delta(y_n) = -a_n x_{n-l+1}, \quad n \in \mathbb{N}(n_0) = 1,2,3,\ldots \]
Examples illustrating the results are inserted.

Keywords
Asymptotic, Two-Dimensional Neutral Delay Difference Systems

1. Introduction
Consider a nonlinear neutral type two-dimensional delay difference system of the form
\[ \Delta(x_n + p_n x_{n-k}) = b_n y_n, \Delta(y_n) = -a_n x_{n-l+1}, \quad n \in \mathbb{N}(n_0) = 1,2,3,\ldots \] (1.1)

Subject to the following conditions:

\( (c_1) \), \( \{a_n\} \) and \( \{b_n\} \) are nonnegative real sequences such that  
\[ \sum_{n=1}^{\infty} b_{n_0} = \infty. \]

\( (c_2) \), \( \{p_n\} \) is a positive real sequence.

\( (c_3) \), \( f, g : \mathbb{R} \rightarrow \mathbb{R} \) are continuous non-decreasing with \( u f(u) > 0 \), \( u g(u) > 0 \), for \( u \neq 0 \) and \( |f(u)| \geq k |u| \), where \( k \) is a constant.

\( (c_4) \), \( k \) and \( l \) are nonnegative integers.

Let \( \theta = \max \{k,1\} \). By a solution of the system (1.1), we mean a real sequence \( \{x_n, y_n\} \) which is defined for all \( n \geq n_0 - \theta \) and satisfies (1.1) for all \( n \in \mathbb{N}(n_0) \).
Let $W$ be the set of all solutions $X = \{x_n, y_n\}$ of the system (1.1) which exists for $n \in \mathbb{N} \left( n_0 \right)$ and satisfies

$$\sup \left\{ \left| x_n \right| + \left| y_n \right| : n \geq N \right\} > 0 \quad \text{for any integer } N \geq N_0.$$  

A real sequence defined on $\mathbb{N} \left( n_0 \right)$ is said to be oscillatory if it is neither eventually positive nor eventually negative and nonoscillatory otherwise.

A solution $X \in W$ is said to be oscillatory if both components are oscillatory and it will be called nonoscillatory otherwise.

Some oscillation results for difference system (1.1) when $p_n = 0$ for $n \in N \left( N_0 \right)$ and $n - l + 1 = n$ have been presented in [1]. In particular when $b_n > 0$ for all $n \in N \left( n_0 \right)$. The difference system (1.1) reduces to the second order nonlinear neutral difference equation

$$\Delta \left( \frac{1}{b_n} \Delta \left( x_n + p_n x_{n-k} \right) \right) = -a_n x_{n-l+1}.$$  

If $b_n = 1$, in Equation (1.2), we have a second order linear equation

$$\Delta^2 \left( x_n + p_n x_{n-k} \right) = -a_n x_{n-l+1}.$$  

For oscillation criteria regarding Equations (1.1)-(1.3), we refer to [2]-[12] and the references cited therein. In Section 2, we present some basic lemmas. In Section 3, we establish oscillation criteria for oscillation of all solutions of the system (1.1). Examples are given in Section 4 to illustrate our theorems.

### 2. Some Basic Lemmas

Denote $A_n = \sum_{i=0}^{n-1} a_i$, $n \in \mathbb{N} \left( n_0 \right)$. For any $x_n$, we define $z_n$ by

$$z_n = x_n + p_n x_{n-k}.$$  

We begin with the following lemma.

2.1. Let $(c_1) - (c_4)$ hold and let $\left( x_n, y_n \right) \in W$ be a solution of system (1.1) with $\{x_n\}$ either eventually positive or eventually negative for $n \in \mathbb{N} \left( n_0 \right)$. Then $\left( x_n, y_n \right)$ is nonoscillatory and $\{z_n\}$ and $\{y_n\}$ are monotone for $n \in N \left( N \right)$ for $N \in \mathbb{N} \left( n_0 \right)$.

Proof. Let $\left( x_n, y_n \right) \in W$ and let $\{x_n\}$ be nonoscillatory on $\mathbb{N} \left( n_0 \right)$. Then from the second equation of system (1.1), we have $\Delta y_n \leq 0$ for all $n \geq N_1 \in \mathbb{N} \left( n_0 \right)$ and $\Delta x_n$ and $y_n$ are not identically zero for infinitely many values of $n$. Thus $\{y_n\}$ is monotone for $n \geq N_1$. Hence $\{x_n\}$ is either eventually positive or eventually negative for $n \geq N_1$. Then, $\left( x_n, y_n \right)$ is nonoscillatory. Further from the first equation of the system (1.1), we have $\Delta x_n > 0$ or $\Delta x_n < 0$ eventually. Hence $\{z_n\}$ is monotone and nonoscillatory for all $n \geq N \geq N_1$. The proof is similar when $\{x_n\}$ is eventually negative.

Lemma 2.2. In addition to conditions $(c_1) - (c_2)$ assume that $0 < p_n \leq 1$ for all $n \in \mathbb{N} \left( n_0 \right)$. Let $\{x_n\}$ be a nonoscillatory solution of the inequality

$$x_n \left( x_n + p_n x_{n-k} \right) \geq 0$$  

for sufficiently large $n$. If for $n - k$ for all $n \in \mathbb{N} \left( n_0 \right)$, then $\{x_n\}$ is bounded.

Proof. Without loss of generality we may assume that $\{x_n\}$ be an eventually
positive solution of the inequality (2.1), the proof for the case \( \{x_n\} \) eventually negative is similar. From (2.1) we have

\[
(x_n + p_n x_{n-k}) \geq 0, \quad \text{for } n \geq N(n_0).
\]

and \( 0 < p_n \leq 1 \), we have from (2.2), \( x_{n-k} \leq p_n x_{n-k} \leq x_n \) for all \( n \geq N \). Hence \( \{x_n\} \) is bounded.

Next, we state a lemma whose proof can be found in [1].

Lemma 2.3. Assume that \( \{a_n\} \) is a non negative real sequence and not identically zero for infinitely many values of \( n \) and \( l \) is a positive integer. If

\[
\liminf_{n \to \infty} \sum_{r=n+l+1}^{n+1} a_r > \left(\frac{l}{l+1}\right)^{l+1}
\]

Then the difference inequality

\[
\Delta y_n + a_n x_{n-l-1} \leq 0, \quad n \in \mathbb{N}(n_0)
\]

cannot have an eventually positive solution and

\[
\Delta y_n + a_n x_{n-l-1} \geq 0, \quad n \in \mathbb{N}(n_0)
\]

cannot have an eventually negative solution.

### 3. Oscillation Theorems for the System (1.1)

Theorem 3.1. Assume that \( \{p_n\} \) is bounded and there exists an integer \( j \) such that \( l > j + k + 2 \). If

\[
\limsup_{n \to \infty} A_n \sum_{i=n-j+1}^{n} a_i > \frac{1}{k \beta}
\]

and

\[
\liminf_{n \to \infty} \sum_{i=n-l-j-k+1}^{n-l-j} k \beta b_i \left( \sum_{i=n-l-j+1}^{n-l-j} a_i \right)^{l-j-k+2} > \left(\frac{l-j-k}{l-j-k+2}\right)^{l-j-k+2}
\]

Then every solution \( \{(x_n, y_n)\} \in W \) is a nonoscillatory solution of system (1.1), with \( \{x_n\} \) bounded. Without loss of generality we may assume that \( \{x_n\} \) is eventually positive and bounded for all \( n \geq n_1 \in \mathbb{N}(n_0) \). From the second equation of (1.1), we obtain \( \Delta y_n \leq 0 \) for sufficiently large \( n \geq n_2 \in \mathbb{N}(n_1) \). In view of Lemma 2.1, we have two cases for sufficiently large \( n \in \mathbb{N}(n_2) \):

1) \( y_n < 0 \) for \( n \geq n_3 \);
2) \( y_n > 0 \) for \( n \geq n_3 \).

Case (1). Because \( \{y_n\} \) is negative and nonincreasing there is constant \( L > 0 \). Such that

\[
y_n \leq -L \quad \text{for all } n \geq n_3
\]

Since \( \{x_n\} \) and \( \{p_n\} \) are bounded, \( \{z_n\} \) defined by (2.1) is bounded. Summing the first equation of (1.1) from \( n_3 \) to \( n-1 \) and then using (3.3), we obtain

\[
z_n - z_{n_3} \leq -L \sum_{i=n_3}^{n-1} a_i, \quad n \geq n_3.
\]

From (3.3), we see that \( \lim_{n \to \infty} z_n = -\infty \) which contradicts the fact that \( \{z_n\} \)
is bounded. Case (1) cannot occur.

Case (2). Let $z_n > 0$ for $n \geq n_4$ where $n_4 \in N(n_3)$ is sufficiently large. Because $\{z_n\}$ is nondecreasing there is a positive constant $M$ such that

$$z_n \geq M, \quad \text{for all } n \geq n_4.$$  

(3.5)

From (2.1), we have $z_n > x_n$, and by hypothesis, we obtain

$$a_n z_{n-1} \geq a_n \frac{x_{n-1}}{k}, \quad n \geq n_5 \in N(n_4)$$  

(3.6)

summing the second equation of (1.1) from $n$ to $i$, using (3.5) and then letting $i \to \infty$, we obtain

$$y_n \geq k \sum_{j=n}^{\infty} a_j z_{j-1}, \quad n \geq n_5.$$  

(3.7)

From condition (3.1), we have

$$\frac{1}{k \beta} < \lim \sup \frac{A_n A_n}{a_n}$$  

(3.8)

we claim that the condition (3.1) implies

$$\sum_{n=N}^{\infty} A_n a_n = \infty, \quad N \in N(n_0).$$  

(3.9)

Otherwise, if $\sum_{n=N}^{\infty} A_n a_n < \infty$, we can choose an integer $N_1 \geq N$ So large that

$$\sum_{n=N_1}^{\infty} A_n a_n < \frac{1}{k \beta}$$  

which contradicts (3.6).

Using a summation by parts formula, we have

$$\sum_{n=N}^{\infty} A_n \Delta g(y_n) = A_N y_N - A_N y_N - z_N + z_N.$$  

(3.10)

From (3.3), (3.4) and (3.6) and the second equation of (1.1), we have

$$\sum_{n=1}^{N} A_n \Delta g(y_n) \leq \beta \sum_{n=N}^{\infty} A_n \Delta y_s$$

$$\leq -M \beta \sum_{n=1}^{N} A_n y_s$$

$$\leq -M \beta \sum_{n=1}^{N} A_n y_s$$

$$M \beta \sum_{n=1}^{N} A_n y_s = -A_N y_N + A_N y_N + z_N - z_N, n \geq N.$$  

combining (3.6) with (3.8), we obtain

$$\lim_{n \to \infty} (z_n - A_N y_N) = \infty.$$  

and

$$z_n \geq A_N g(y_n) \geq \beta A_N y_n, \quad n \geq n_6 \in N(n_6).$$  

The last inequality together with (3.4) and the monotonicity of $\{z_n\}$ implies

$$z_n \geq k \beta A_N \sum_{j=n}^{\infty} a_j z_{j-1} \geq k \beta A_N \sum_{j=n-1}^{\infty} a_j z_{j-1}$$

$$\geq k \beta A_N z_n \sum_{j=n-1}^{\infty} a_j$$

$$\sum_{j=n-1}^{\infty} a_j$$
and \( 1 \geq k\beta A_nz_n \sum_{s=1}^{n} a_s, \quad n \in \mathbb{N}(n_k) \) which contradicts (1.1). This case cannot occur. The proof is complete.

**Theorem 3.2.** Assume that \( 0 < p_n \leq 1 \), then there exists an integer \( j \) such that \( l > j+k \) and the conditions (3.1) and (3.2) are satisfied. Then all solutions of (1.1) are oscillatory.

**Proof.** Let \( \left\{ (x_n, y_n) \right\} \in \mathcal{W} \) be a nonoscillatory solution of (1.1). Without loss of generality we may assume that \( \{x_n\} \) is positive for \( n \in \mathbb{N}(n_1) \). As in the proof of above theorem we have two cases.

**Case (1).** Analogous to the proof of case (1) of above theorem, we can show that \( \lim_{n \to \infty} z_n = -\infty \). By Lemma 2.2, \( \{x_n\} \) is bounded and hence \( \{z_n\} \) is bounded which is a contradiction. Hence case (1) cannot occur.

**Case (2).** The proof of case (2) is similar to that of the above theorem and hence the details are omitted. The proof is now complete.

**Theorem 3.3.** Assume that \( 0 < p_n \leq 1 \) and

\[
\limsup_{n \to \infty} \sum_{s=1}^{n} \frac{k\beta (A_n - A_{s+1})}{p_{s-l-k+1}} A_n > 1.
\]

(3.14)

\[
\sum_{n=N}^{\infty} \left( \sum_{s=1}^{n} a_s \right) = \infty, \quad N \in \mathbb{N}(n_0)
\]

(3.15)

\[
\limsup_{n \to \infty} \left( k\beta A_n \sum_{s=1}^{n} a_s \right) > 1.
\]

(3.16)

Then all solutions of (1.1) are oscillatory.

**Proof.** Let \( \left\{ (x_n, y_n) \right\} \in \mathcal{W} \) be a nonoscillatory solution of (1.1). Without loss of generality we may assume that \( \{x_n\} \) is positive for \( n \in \mathbb{N}(n_1) \). As in the proof of above theorem we have two cases.

**Case 1.** From (2.1), we have

\[
z_n > p_n x_{n-k}, \quad n \geq n_1 \in \mathbb{N}(n_1)
\]

and

\[
f(x_{n-l+1}) \geq kx_{n-l+1} > k \frac{z_{n-l+k+1}}{p_{n-l-k+1}}, \quad n \geq n_4
\]

(3.17)

where \( n_4 \in \mathbb{N}(n_1) \) is sufficiently large. Then the following equality

\[
z_n = z_i + \left( A_n - A_i \right) y_i + \sum_{s=1}^{n-i} (A_n - A_{s+1}) \Delta y_s
\]

\[
z_n < \sum_{s=i}^{n-1} (A_n - A_{s+1}) \Delta y_s, \quad n > i \geq n_5.
\]

Combining the last inequality with the second equations of (1.1) and (3.17), we have

\[
z_n < \beta \sum_{s=1}^{n-1} (A_n - A_{s+1}) \left( -a_s f(x_{s-l+k+1}) \right)
\]

\[
< k\beta \sum_{s=1}^{n-1} \frac{(A_n - A_{s+1}) z_{s-l-k+1}}{p_{s-l-k+1}}, n \geq n_5.
\]

Let \( i = n-l+k+1 \) and using the monotonocity of \( \{z_n\} \), from the last inequality, we obtain
\[ z_n < z_{s} \sum_{s=n-t-k+1}^{n-1} k \beta \left( A_s - A_{s+1} \right) \frac{a_s}{p_{s+1+k-1}} \]

and

\[ 1 > z_{s} \sum_{s=n-t-k+1}^{n-1} k \beta \left( A_s - A_{s+1} \right) \frac{a_s}{p_{s+1+k-1}} \]

which contradicts the condition (3.14).

Case 2. The proof for this case is similar to that of Theorem (3.1). Here we use condition (3.16) instead of condition (2.1). The proof is complete.

4. Examples

Example 4.1. Consider the difference system

\[ \Delta \left( x_n + \frac{1}{2} x_{n-3} \right) = \frac{1}{n} y_n \]
\[ \Delta y_n = -nx_{n-2}, \quad n \geq 1. \]

The conditions (3.1) and (3.2) are

\[ \limsup_{n \to \infty} \frac{1}{n} \sum_{s=0}^{\infty} s = \infty. \]
\[ \liminf_{n \to \infty} \frac{1}{n} \sum_{s=0}^{n-1} \left( \sum_{t=1}^{s} 2t \right) = 4. \]

All conditions of Theorem 3.2 are satisfied and so all solutions of the system (4.1) are oscillatory.

Example 4.2. Consider the difference systems

\[ \Delta \left( x_n + \frac{1}{4} x_{n-2} \right) = (n+1) y_n \]
\[ \Delta y_n = -\frac{c}{n+1} x_{n-1}, \quad n \geq 1, \]

where \( c \) is a positive constant. The conditions (3.1) and (3.2) are

\[ \limsup_{n \to \infty} \left( n+1 \right) \sum_{s=1}^{n} \frac{c}{s+1} = \infty \]

and

\[ \liminf_{n \to \infty} \sum_{s=1}^{n-2} \left( s+1 \right) \left( \sum_{t=1}^{s+1} \frac{4c}{t+1} \right) = 12c. \]

For \( c > \frac{1}{12} \), all conditions of Theorem 3.2 are satisfied and so all solutions of the system (4.2) are oscillatory.

References

[1] Agarwal, R.P. (2000) Difference Equations and Inequalities. 2nd Edition, Marcel Dekkar, New York.

[2] Elizebeth, S., Graef, J.R., Sundaram, P. and Thandapani, E. (2005) Classifying Non-oscillatory Solutions and Oscillations of Neutral Difference Equations. Journal of Difference Equations and Applications, 11, 605-618.
[3] Sternal, A. and Szamanda, B. (1996) Asymptotic and Oscillatory Behaviour of Certain Difference Equations. *Le Matematiche*, 51, 77-86.

[4] Sternal, A., Szamanda, B. and Szafrański, Z. (1998) Oscillatory and Asymptotic Behaviour of Some Difference Equations. *Publications. DE L’INSTITUT MATHEMATIQUE*, 63, 66-74.

[5] Thandapani, E. (1992) Asymptotic and Oscillatory Behaviour of Solutions of a Second Order Neutral Difference Equations. *Riv. Mat. Univ. Parma*, 1, 105-113.

[6] Thandapani, E. and Mahalingam, K. (2003) Necessary and Sufficient Condition for Oscillation of Second Order Neutral Difference Equations. *Tamkang Journal of Mathematics*, 34, 137-145.

[7] Thandapani, E. and Mohankumar, P. (2007) Oscillation of Difference Systems of the Neutral type. *Computers and Mathematics with Applications*, 54, 556-566. https://doi.org/10.1016/j.camwa.2006.12.029

[8] Thandapani, E., Sundaram, P., Graef, J.R. and Spikes, P.W. (1995) Asymptotic Properties of Solutions of Nonlinear Second Order Neutral Delay Difference Equations. *Dynam. Systems Appl.*, 4, 125-136.

[9] Thandapani, E. and Sundaram, P. (2000) Oscillation and Nonoscillation Theorems for Second Order Quasilinear Functional Difference Equations. *Indian J. Pure Appl. Math*, 31, 37-47.

[10] Zhang, B.G. and Saker, S.H. (2003) Kamenev-Type Oscillation Criteria for Nonlinear Neutral Delay Difference Equations. *Indian J. Pure Appl. Math.*, 34, 1571-1584.

[11] Zhang, G. (2002) Oscillation for Nonlinear Neutral Difference Equations. *Applied Math. E-Notes*, 2, 22-24.

[12] Zhang, Z., Chen, J. and Zhang, C. (2001) Oscillation of Solutions of Second Order Nonlinear Difference Equations with Nonlinear Neutral Term. *Computers and Mathematics with Applications*, 41, 1571-1584. https://doi.org/10.1016/s0898-1221(01)00113-4

Submit or recommend next manuscript to SCIRP and we will provide best service for you:

Accepting pre-submission inquiries through Email, Facebook, LinkedIn, Twitter, etc.
A wide selection of journals (inclusive of 9 subjects, more than 200 journals)
Providing 24-hour high-quality service
User-friendly online submission system
Fair and swift peer-review system
Efficient typesetting and proofreading procedure
Display of the result of downloads and visits, as well as the number of cited articles
Maximum dissemination of your research work

Submit your manuscript at: [http://papersubmission.scirp.org/](http://papersubmission.scirp.org/)
Or contact jmp@scirp.org
A Multi-MW Proton/Electron Facility at KEK

R. Belusevic

IPNS, High Energy Accelerator Research Organization (KEK), Ibaraki, Japan
Email: belusev@post.kek.jp

Abstract

The main “bottleneck” limiting the beam power in circular machines is caused by space charge effects that produce beam instabilities. To increase maximally the beam power of a “proton driver”, it is proposed to build a facility consisting solely of a 2.5 GeV injector linac (PI) and a 20 GeV pulsed superconducting linac (SCL). Such a facility could be constructed using the existing KEK accelerator infrastructure. The PI, based on the European Spallation Source (ESS) linac, would serve both as an injector to the SCL and a source of proton beams that could be used to copiously produce, e.g., muons and “cold” neutrons. Protons accelerated by the SCL would be transferred through the KEK Tristan ring in order to create neutrino, kaon and muon beams for fixed-target experiments. At a later stage, a 70 GeV proton synchrotron could be installed inside the Tristan ring. The SCL, comprising 1.3 GHz ILC-type rf cavities, could also accelerate polarized or unpolarized electron beams. After acceleration, electrons could be used to produce polarized positrons, or may traverse an XFEL undulator.

Keywords
Proton Driver, Superconducting Linac, Neutrino Oscillations, Rare Kaon Decays, XFEL

1. Introduction

The Standard Model (SM) of particle physics is supported by two theoretical “pillars”: the gauge principle and the Higgs mechanism for particle mass generation. In this theoretical model, the mass of a particle depends on its interaction with the Higgs field, a medium that permeates the universe. The SM predicts the existence of a neutral spin-0 particle (the Higgs boson) associated with the Higgs field, but it does not predict its mass. Whereas the gauge principle has been firmly established through precision electroweak measurements, the Higgs mechanism is yet to be fully tested.
The Higgs-boson mass, $m_H$, affects the values of electroweak observables through radiative corrections. Many of the electroweak measurements obtained so far may be combined to provide a global test of consistency with the SM. The best constraint on $m_H$ is obtained by making a global fit to the electroweak data. Such a fit is consistent with the value of the Higgs mass measured at the Large Hadron Collider (LHC) [1] [2].

To discover a new particle (such as the Higgs boson), or to search for physics beyond the SM, usually requires the use of high-energy hadron or electron-positron colliders. However, many important discoveries in particle physics have been made using proton beams with relatively low energies but high intensities (flavor mixing in quarks and in neutrinos are noteworthy examples). Experiments with high-intensity neutrino beams, e.g., are designed primarily to explore the mass spectrum of the neutrinos and their properties under the CP symmetry.

Some of the most important discoveries emerged from high-precision studies of $K$ mesons ("kaons"), in particular neutral kaons. A deeper insight into CP violation is expected to be gained from measurements of ultra-rare kaon decays such as $K^0 \rightarrow \pi^0 \nu \bar{\nu}$ and $K^- \rightarrow \pi^- \nu \bar{\nu}$. These decays provide important information on higher-order effects in electroweak interactions, and therefore can serve as a probe of new phenomena not predicted by the Standard Model.

The physics programs envisaged at the proposed facility are, to a large extent, complementary to each other. For instance, neutrino oscillation experiments and searches for permanent electric dipole moments both look for new sources of CP violation, a phenomenon which reflects the fundamental difference between matter and antimatter.

A unique feature of the proposed facility is the use of superconducting ILC-type cavities to accelerate both protons and electrons, which considerably increase its physics potential. Polarized electrons and positrons can be used to study the structure of composite particles and the dynamics of strong interactions, as well as to search for new physics beyond the Standard Model.

2. The Proposed Proton/Electron Facility at KEK

The main “bottleneck” limiting the beam power in circular machines is caused by space charge effects that produce beam instabilities. Such a “bottleneck” exists at the J-PARC proton synchrotron complex, and is also intrinsic to the “proton drivers” envisaged at CERN and Fermilab. To increase maximally the beam power of a “proton driver”, it is proposed to build a facility consisting solely of a low-energy injector linac and a high-energy pulsed superconducting linac. Pulsed operation is preferred over the CW mode (continuous wave, 100% duty) mainly because the former allows the use of rf cavities with high accelerating gradients. This would considerably reduce the overall length of the machine, which is limited by the size of the KEK site.

The layout of the proposed proton/electron facility at KEK is shown in Figure 1. A 2.5 GeV proton linac (PI) serves both as an injector to a superconducting
linac (SCL) and a source of proton beams that can be used to copiously produce neutrons and muons. Protons accelerated by the SCL to 20 GeV are transferred through the KEK Tristan ring in order to create beams for various fixed-target experiments. At a later stage, a 70 GeV proton synchrotron could be installed inside the Tristan ring. The SCL, comprising 1.3 GHz superconducting ILC-type rf cavities, can also accelerate polarized or unpolarized electron bunches. After acceleration, electrons may be used to produce polarized positrons. An SCL-based XFEL and a synchrotron light source for applications in materials science and medicine are also envisaged.

The proposed facility could be constructed using the existing KEK accelerator infrastructure. As shown in Figure 2, the present KEK linac tunnel and klystron gallery could be extended to increase the length, and hence the maximum energy, of each linac in Figure 1. The cryomodules, RF sources and cryogenic plant units of the proposed linac complex would be installed inside these extended structures (see Figure 3). The Tristan ring (TR) would initially be equipped only with magnets capable of steering the SCL proton bunches to various fixed targets. The four 200 m-long straight sections of the TR, each with an experimental hall in the middle (see Figure 2), would house beam lines and detectors.

The beam power of a pulsed linear accelerator is given by the expression

\[ P_b = E_b [\text{MV}] \times I [\text{A}] \times \tau_p [\text{s}] \times R [\text{Hz}] \]  

where \( P_b \) is the beam power, \( E_b \) is the beam energy, \( I \) is the average current per pulse, \( \tau_p \) is the beam pulse length, and \( R \) is the repetition rate. The duty cycle of a pulsed linac is \( D = \frac{\tau_p}{R} \). Using the values from Table 1, and assuming \( E_b = 20 \text{ GeV} \), one obtains \( D = 0.024 \) and

\[ P_b = 20000 \text{ MV} \times 31 \text{ mA} \times 1.2 \text{ ms} \times 20 \text{ s}^{-1} \approx 15 \text{ MW} \]  

The beam parameters in Table 1 are mutually constrained by the following relations: The number of protons per second \( N = \frac{P_b}{E_b} \) and the number of protons per pulse \( N_p = N / R \); the average current per pulse is \( I = \left( N_p \times 1.6 \times 10^{-19} \text{ C} \right) / \tau_p \). The klystron pulse length is the sum of the rf cavity fill time (current dependent) and the beam pulse length: \( \tau = \tau_f + \tau_p \). For ILC-
Figure 2. A sketch of the KEK site showing the Tristan ring and the existing electron linac (in blue). The lines drawn in red indicate possible extensions of the present linac tunnels. Alternatively, a new underground linac tunnel could be excavated at a greater depth in case a larger SCL beam energy (up to 25 GeV) is required.

Figure 3. Front view of the linac tunnel and the klystron gallery housing the cryomodules, RF sources and cryogenic plant units of the proposed superconducting linac complex.
Table 1. Parameters of the SCL.

| Parameter                          | Value       |
|-----------------------------------|-------------|
| Beam energy                       | 20 GeV      |
| Beam power                        | 15 MW       |
| Repetition rate                    | 20 Hz       |
| Protons per pulse                 | $2.3 \times 10^{14}$ |
| Beam pulse length                 | 1.2 ms      |
| Average current per pulse         | 31 mA       |
| Duty cycle                        | 2.4%        |
| RF frequency                      | 1.3 GHz     |
| Klystron average power            | 150 kW      |
| Klystron peak power               | 5 MW        |
| Klystron pulse length             | 1.5 ms      |
| Effective accelerating gradient   | 20 MV/m     |
| Peak power per coupler            | 460 kW      |

2.1. Main Characteristics of an ILC-Type Linac

The main characteristics of a linear accelerator are determined by the properties of its rf source (klystrons) and accelerating cavities. For a pulsed ILC-type superconducting linac, one of the currently available rf sources is the Toshiba E3736 Multi-Beam Klystron [3]. This source has the following well-tested specifications: rf frequency—1.3 GHz; peak rf power—10 MW; average power—150 kW; efficiency—65%; pulse length—1.5 ms; repetition rate—10 Hz. If the repetition rate of the Toshiba klystron is increased by a factor of two, while its peak power is reduced by the same factor (thus keeping the average power constant) one obtains the klystron specifications presented in Table 1. For such a klystron, a suitable 20 Hz pulse modulator has to be developed.

A very important parameter that determines, to a large extent, the power conversion efficiency of a klystron is its perveance, defined by

$$K \equiv \frac{I_0}{U^{3/2}}$$  \hspace{1cm} (3)

In this expression, $I_0$ is the beam current and $U$ is the anode voltage. Since there is an upper limit to the applied voltage, low perveance can be only obtained by operating with low currents. For single-beam klystrons, this requirement is not compatible with the need for high output power. With this in mind, multi-beam klystrons (MBK) were originally developed in the 1960s [4]. An MBK is a parallel assembly of low-current (low-perveance) beamlets within a common rf structure, which efficiently generates high output power. Using Equation (3), the output rf power of an MBK can be expressed as

$$P_k = \eta I_0 U = \eta KU^{3/2}$$  \hspace{1cm} (4)

where $\eta$ is the klystron efficiency and $I_0 = N_b I_b$ is the total beam current;
$N_b$ is the number of beamlets and $I_b$ is the current carried by each beamlet. For Toshiba’s E3736 MBK, $\eta = 65\%$, $U = 116$ kV and $I_b=134$ A. Hence, $K = 3.4 \times 10^{-6} \text{ A}/\sqrt{\text{V}}^2$, klystron’s peak power $P_\text{K} = 10$ MW and its average power $\overline{P}_\text{K} = R \times \tau \overline{R} = 150$ kW for $\tau = 1.5$ ms and $R = 10$ Hz. Since the klystron has six beamlets, $I_b = 22.3$ A.

In order to increase the beam power of the SCL, $R$ could be increased to 20 Hz (see Equation (1)). As already mentioned, the peak power of the klystron would then have to be reduced to 5 MW, so that its average power is still 150 kW for the klystron pulse length $\tau = 1.5$ ms (see Table 1).

The basic properties of a 1.3 GHz superconducting ILC-type cavity are presented, e.g., in [5]. There are two important parameters that characterize rf cavities: the accelerating gradient $E_{\text{acc}}$ and the unloaded quality factor $Q_0$. The former is a measure of how much the energy of a particle is increased over a given length of the linac (typically expressed in units of MV/m), while the latter specifies how well the cavity can sustain the stored rf power. A higher value of $Q_0$ implies a lower rate of power loss relative to the stored energy$^1$. ILC-type cavities must have a nominal $Q_0$ greater than $1 \times 10^{10}$ (a dimensionless parameter) at $E_{\text{acc}} = 31.5$ MV/m.

Each ILC-type cryomodule for the proposed SCL would contain eight niobium 9-cell cavities and a quadrupole magnet at its centre. Other major components of such a cryomodule are the vacuum vessel, thermal and magnetic shields, cryogenic piping, interconnections, etc. The inactive spaces between cavities or cryomodules (the “packing fraction”) are responsible for a substantial reduction in the average accelerating gradient of the linac. The beam physics and the lattice design of the superconducting L-band linac described in [6] are applicable to the SCL.

The average usable accelerating gradient in ILC-type cavities is $E_{\text{acc}} = 29.3 \pm 5.1$ MV/m [7]. Taking into account an estimated linac “packing fraction” of about 70%, the effective accelerating gradient of the SCL is $E_{\text{eff}} \approx 20$ MV/m. Hence, the total length of a 20 GeV linac is ~1000 m.

Since the length of an ILC 9-cell cavity is 1 m, a linac with $E_b = 20$ GeV would require about $N_{\text{cav}} = (20000 \text{ MeV})/(29 \text{ MeV}) = 690$ cavities. The average input rf power per cavity is thus $\overline{P}_{\text{cav}} = P_c / N_{\text{cav}} \approx 22$ kW, and the corresponding peak power $P_{\text{cav}} = \overline{P}_{\text{cav}} / D = 916$ kW. Although this value is acceptable for a pulsed linac with $D \approx 2\%$, it would be prudent to use two rf couplers per cavity. In that case the peak rf power per coupler would be about 460 kW.

For $E_{\text{acc}} = 30$ MV/m, ohmic losses in an ILC 9-cell cavity amount to $P_c = 100$ W in the CW mode of operation, but only $P_c = (10\times D)$ W = 2.4 W (plus static loss) in the pulsed mode with a duty factor $D = 0.024$. Because of large ohmic losses, which scale with the square of the accelerating gradient, $E_{\text{acc}}$ is generally lower in the CW mode of operation than in the pulsed mode.

$^1$The $Q$ factor of an rf cavity is defined as $Q = 2\pi\times$ (energy stored/energy dissipated per cycle). For large values of $Q$, the $Q$ factor is approximately the number of oscillations required for the energy of a freely oscillating system to fall off to $e^{-2\pi}$, or 0.2%, of its original value.
2.2. Proton Injector (PI)

A typical ~1 GeV proton linear accelerator consists of three main sections:

- Front end, comprising a proton source and a radiofrequency quadrupole accelerator;
- Medium-velocity linac, which accelerates proton beams to ~100 MeV;
- High-velocity linac, which accelerates protons to energies exceeding 1 GeV.

The most complex part of a proton linac is the low-energy (low-$\beta$) section, situated between the proton (or ion) source and the first drift-tube-based accelerating section. The continuous beam of protons coming from an electron cyclotron resonance (ECR) source [8] has to be focused, bunched and accelerated in the first rf structure. These three essential functions are nowadays successfully performed by radio frequency quadrupoles (RFQ) [9]. However, the beam has to be shrunk before it can be fed into an RFQ. This is accomplished within a low-energy beam transport (LEBT) section by means of cylindrical magnets (solenoids).

As soon as the beam is bunched—which is essential for further acceleration—it enters a medium-energy beam transport (MEBT) section, where it is collimated and steered from the RFQ into the medium-velocity linac (MVL). The MEBT may also contain a number of buncher cavities. Inside the MVL, the beam is accelerated to about 100 MeV ($\beta \sim 0.1$ to 0.5). The MVL usually contains normal-conducting drift-tube linac (DTL) and cell-coupled drift tube linac (CCDTL) structures. A DTL incorporates accelerating components of increasing length in order to match precisely the increase in beam velocity, while quadrupole magnets provide strong focusing. The main advantage of using CCDTL structures is that they provide longitudinal field stability.

High-velocity linac (HVL) structures accelerate the beam to energies around 1 GeV. They consist either of normal-conducting side-coupled linac (SCL) structures\(^2\) or superconducting elliptical cavities. The latter offer some advantages over the former, such as higher accelerating gradients and lower operating costs. The superconducting HVL can also feature spoke resonators, characterized by their simplicity, high mechanical stability and compact size [10].

The European Spallation Source (ESS) is an example of a typical ~1 GeV proton linear accelerator [11] (see Figure 4). The transverse beam size along the linac varies in the range 1 - 4 mm, and the bunch length decreases from 1.2 cm to 3 mm towards the end of the linac.

One of the main concerns in the design of a high-power proton linac is to restrict beam losses. A careful beam dynamics study is therefore needed in order to avoid halo formation, a major source of beam loss. Another important issue is the preservation of beam emittance [6].

High-power proton linear accelerators have a wide range of applications including spallation neutron sources, nuclear waste transmutation, production of radioisotopes for medical use, etc. A number of laboratories worldwide have ex-

\(^2\)The main reason for using these π/2-mode structures is that long chains of coupled cavities are often required for an efficient use of high-power rf sources [10].
Figure 4. Block diagram of the ESS linac [11]. The RFQ and DTL structures are normal-conducting, while the spoke resonator and elliptical cavities are superconducting.

pressed interest in building “proton drivers” that would primarily deliver high-intensity neutrino, kaon and muon beams [12] [13].

3. Physics at the Proposed Facility

The physics potential of a multi-MW “proton driver” is extensively discussed, for instance, in [14]. This note is mainly concerned with the application of a high-intensity proton source to investigate the properties of long-baseline neutrino oscillations. A unique feature of the proposed facility is the use of a superconducting linac to accelerate both protons and electrons, which considerably increases its physics potential. As described in [15] [16], polarized electron and positron beams can be used to study the structure of composite particles and the dynamics of strong interactions, as well as to search for new physics beyond the Standard Model.

3.1. Neutrino Flavor Oscillations and Leptonic CP Violation

The observed transformation of one neutrino flavor into another (“neutrino oscillation”) indicates that these particles have finite masses. Cosmological data suggest that the combined mass of all three neutrino species is a million times smaller than that of the next-lightest particle, the electron [17]. The phenomenon of neutrino oscillations implies not only the existence of neutrino mass, but also of neutrino mixing. That is, the neutrinos of definite flavor are not particles of definite mass (mass eigenstates), but coherent quantum-mechanical superpositions of such states. Conversely, each neutrino of definite mass is a superposition of neutrinos of definite flavor. Neutrino mixing is large, in striking contrast to quark mixing. Whatever the origin of the observed neutrino masses and mixings, it implies a profound modification of the Standard Model.

Mathematically, the phenomenon of neutrino mixing can be expressed as a unitary transformation relating the flavor and mass eigenstates. The neutrino oscillation rate depends, in part, on (1) the difference between neutrino masses and (2) the three parameters in the transformation matrix known as mixing angles. The complex phase factors in the transformation matrix (also called mixing matrix) are associated with the violation of CP symmetry in the lepton sector. The size of the CP violation is determined both by the phases and the mixing angles.

Experiments with high-intensity neutrino beams are designed primarily to explore the mass spectrum of the neutrinos and their properties under the CP symmetry, and thus provide a deeper insight into the nature of these elusive par-
articles and their role in the universe. For instance, if there is experimental evidence for CP violation in neutrino oscillations, it could be used to explain the observed asymmetry between matter and antimatter [18].

To search for CP violation in neutrino oscillations, a 100 kiloton water Cherenkov detector could be built at Okinoshima, located at a distance 

\[ L_2 \approx 650 \text{ km} \] from KEK. Using the proposed KEK “proton driver”, the detector at Okinoshima and the existing 0.022 Mt Super-Kamiokande detector (situated at a distance \( L_1 \approx 300 \text{ km} \) from KEK), the neutrino mass hierarchy could be determined either by comparing the \( \nu_\mu \) appearance probabilities measured at the two vastly different baseline lengths \( L_1 \) and \( L_2 \), or by measuring at \( L_1 \) and \( L_2 \) the neutrino energy of the first oscillation maximum. Once the mass hierarchy is determined, the CP-violating phase in the mixing matrix can be measured with a precision of \( \pm 20^\circ \), assuming that \( 2.5 \times 10^{21} \) protons are delivered on target for both \( \nu_\mu \) and \( \nu_\tau \) beams [19].

**Proton Target and Magnetic Horn**

The main challenge in the design of a multi-MW neutrino beam facility is to build a *proton target* that could dissipate large amounts of deposited energy, withstand the strong pressure waves created by short beam pulses, and survive long-term effects of radiation damage. Simulation studies of the pion production and energy deposition in different targets (liquid mercury jet, tungsten powder jet, solid tungsten bars and gallium liquid jet) are presented in [20]. Those studies also provide estimates of the amount of concrete shielding needed to protect the environment from the high radiation generated by each target. A proof-of-principle demonstration of a 4 MW target station comprising a liquid mercury jet inside a 20 T solenoidal magnetic field is described in [21]. Alternatively, one could use a rotating, gas-cooled tungsten target that would require the least amount of development effort, and would also have good thermal and mechanical properties [11]. A 15 MW proton beam could be separated by a series of magnets into four beam lines. Each of the four beams would be focused by a series of quadrupoles and correctors to an assembly consisting of four targets and the same number of magnetic horns (see, e.g., [22]).

To maximize the discovery potential of a neutrino beam facility, it is important to properly design the *magnetic horn* that focuses the charged particles produced in the proton target. For proton beam pulses lasting 1 ms, a DC horn has been designed at KEK by Yukihide Kamiya [23]. The toroidal magnetic field of the horn, characterized by \( B(r) = \text{const.} \), is generated by hollow aluminium conductors containing water. The strength of the magnetic field \( B = 0.2 \text{ T} \), and its length \( \ell = 5 \text{ m} \); hence, \( B \cdot \ell = 1 \text{T} \cdot \text{m} \). The radius of the magnet, \( r \), is determined by \( r = L \tan(\theta) + \ell \tan(\theta/2) \), where \( \theta \approx 0.03 + 0.3/p \) is the initial angle a charged pion makes with respect to the proton beam direction, \( L \) is the distance from the target to the horn, and \( p \) is the pion momentum. For example, if \( L = 5 \text{ m} \) then \( r \approx 5 \text{ m} \). The total power generated in the conductors is about 10 MW.
3.2. Physics with Polarized Electrons and Positrons

Electron and positron beams, polarized and/or unpolarized, can be used to study the structure of composite particles and the dynamics of strong interactions, as well as to search for new physics beyond the Standard Model. A detailed description of the physics potential of a facility that can provide such beams (e.g., the upgraded CEBAF facility at Jefferson Lab or the proposed KEK superconducting linac) is presented in [15] [16].

Polarized positrons are created in a conversion target by circularly polarized photons, which themselves are produced when polarized laser light is Compton-backscattered on a high-energy electron beam [24]. Circularly polarized photons can also be produced by bremsstrahlung from polarized electrons [25]. Using polarized electrons and positrons, the nucleon electromagnetic form factors and generalized parton distributions can be determined in a model-independent way [16].

Among the physics topics discussed in [15], parity violation in electron-electron (Møller) scattering is of particular interest. Møller scattering is a purely leptonic process that allows high-precision tests of the Standard Model. At four-momentum transfers much smaller than the mass of the Z boson \( q^2 \ll m_Z^2 \), the parity-violating asymmetry, \( A \), is dominated by the interference between the electromagnetic and neutral weak amplitudes [26]. By definition,

\[
A = \frac{d\sigma_R - d\sigma_L}{d\sigma_R + d\sigma_L} \approx \frac{f_R^2 - f_L^2}{f_r^2}
\]  

In this expression, \( d\sigma_R \) ( \( d\sigma_L \) ) is the differential cross-section for right-handed (left-handed) electron scattering on an unpolarized target:

\[
d\sigma_R \propto \left| f_r + f_L^{\gamma Z} \right|^2 \approx \left| f_r \right|^2 + 2 f_r f_L^{\gamma Z}
\]

where \( f_r \) and \( f_L^{\gamma Z} \) are the scattering amplitudes with \( \gamma \) and \( Z \) exchange, respectively. From the four Feynman diagrams in Figure 6 of [27], one can readily obtain the Born amplitudes for Møller scattering mediated by photons and Z bosons. The weak neutral current amplitudes are functions of the weak mixing (or Weinberg) angle \( \theta_w \), which relates the weak coupling constants \( g_w \) and \( g_Z \) to the electromagnetic coupling constant. As shown in [27], the polarization asymmetry for polarized electron scattering on an unpolarized target is given by

\[
A^{\text{Born}} = m_e E \frac{G_F Q_w'}{\sqrt{2} \pi \alpha} F(\theta)
\]

where \( m_e \) is the mass of the electron, \( E \) is the incident beam energy, \( G_F \) is the Fermi coupling constant characterizing the strength of the weak interaction, \( \alpha \) is the fine structure constant, and \( F(\theta) \) is a function of the scattering angle in the center-of-mass frame. The weak charge of the electron, \( Q_w' = 1 - 4 \sin^2 \theta_w \), is proportional to the product of the electron’s vector and axial-vector couplings to the Z boson.

Since the value of \( \sin^2 \theta_w \) is close to 1/4, there is an enhanced sensitivity of \( A \)
to small changes in the Weinberg angle. The value of $\theta_w$ varies as a function of the four-momentum transfer, $q$, at which it is measured. This variation, or “running”, is a key prediction of the Standard Model. The one-loop electroweak radiative corrections to $A^{\text{Born}}$ (calculated once the renormalized parameters in (7) are properly defined) reduce its Born value by ~40%. This effect can be attributed to an increase of $\sin^2 \theta_w(q^2)$ by 3% as the four-momentum transfer “runs” from $q^2 = m_Z^2$ to $q^2 \approx 0$ [28]. The precision with which $\sin^2 \theta_w$ can be measured in the MOLLER experiment [29] (where a 11 GeV longitudinally polarized electron beam scatters on atomic electrons in a liquid hydrogen target) is shown in Figure 5.2 of [15].

Many of the electroweak measurements obtained so far may be combined to provide a global test of consistency with the SM. Since the Higgs-boson mass affects the values of electroweak observables through radiative corrections, it is of fundamental importance to test the agreement between the directly measured value of $m_H$ and that inferred from the measurements of electroweak parameters $m_W$, $m_{\text{top}}$, and $\sin^2 \theta_w$ (see Figure 5.2 in [15]). High-precision electroweak measurements, therefore, represent a natural complement to direct studies of the Higgs sector.

Apart from providing a comprehensive test of the SM, precision measurements of weak neutral current interactions at $q^2 \ll m_Z^2$ also allow indirect access to new physics phenomena beyond the TeV energy scale. For instance, such measurements can be used to look for hypothetical $Z'$ bosons, 4-fermion contact interactions, or very weakly coupled low-mass “dark bosons” [30].

### 3.3. Rare Kaon Decays

CP violation was introduced in the SM by increasing the number of quark and lepton families to at least three (M. Kobayashi and T. Maskawa, 1973). This idea became very attractive with the subsequent discovery (in 1977) of the bottom quark, which forms, together with the top quark (discovered in 1995), a third family of quarks. It is a remarkable property of the Kobayashi-Maskawa model that quark mixing and CP violation are intimately related [31].

![Figure 5](image.png)

**Figure 5.** Unitarity triangle from $K \to \pi \nu\bar{\nu}$ decays. The displacement of the bottom-right vertex is due to the charm-quark contribution to $K^+ \to \pi^+ \nu\bar{\nu}$. 

---

R. Belusevic
A deeper insight into CP violation is expected to be gained from precision measurements of rare kaon decays such as $K^0_L \rightarrow \pi^0 \nu \nu$ and $K^+ \rightarrow \pi^+ \nu \nu$. Both decays are theoretically 'clean' because hadronic transition amplitudes are matrix elements of quark currents between mesonic states, which can be extracted from the leading semileptonic decays using isospin symmetry. Since photons do not couple to neutrinos, $K \rightarrow \pi \nu \nu$ decays are entirely due to second-order weak processes determined by $Z$-penguin and $W$-box diagrams [31].

The process $K^0_L \rightarrow \pi^0 \nu \nu$ proceeds almost entirely through direct CP violation, and is completely determined by "short-distance" one-loop diagrams with top quark exchange. The Standard Model predicts its branching ratio to be [32]

$$B(K^0_L \rightarrow \pi^0 \nu \nu) = (2.43 \pm 0.39) \times 10^{-11} \text{ theoretical value} \quad (8)$$

This decay is an important source of information on higher-order effects in electroweak interactions, and thus can serve as a probe of new physics (see [33] and references therein).

The decay $K^+ \rightarrow \pi^+ \nu \nu$ receives both CP-conserving and CP-violating contributions. It has theoretical uncertainties that are somewhat larger than those in the process $K^0_L \rightarrow \pi^0 \nu \nu$. Since both decays involve one-loop Feynman diagrams with top quark exchange, they can yield valuable measurements of the CKM matrix elements $|V_{us}|$ and $|V_{cd}|$. The quantity $\text{Im} V_{us}^* V_{cd}$, which can be obtained from $K^0 \rightarrow \pi^0 \nu \nu$ alone, plays a central role in the phenomenology of CP violation in K decays; this quantity is related to the Jarlskog parameter, the invariant measure of CP violation in the Standard model [31] [34].

By measuring the branching ratios of both $K \rightarrow \pi \nu \nu$ decay modes, the unitarity triangle of the CKM matrix can be completely determined (see Figure 5), provided the matrix element $V_{ub}$ and the top quark mass are known [34]. Of particular interest is the unitarity triangle parameter $\sin 2 \beta$, which can also be determined from the decay $B_d \rightarrow \Psi K_s$. Both determinations of this parameter have to coincide if the Standard Model is valid [33].

The current branching ratio measurement of the charged decay mode,

$$B(K^+ \rightarrow \pi^+ \nu \nu) = \left(17.3^{+11.5}_{-10.5}\right) \times 10^{-11} \text{ measured value} \quad (9)$$

is based on the seven candidate events observed by the experiment E787/E949 at Brookhaven [35]. This result is consistent with $(7.8 \pm 0.80) \times 10^{-11}$, the value predicted by the SM [32].

As shown in Figure 11 of [33], the kaon yield rises rapidly as a function of the incident proton momentum. From the figure one infers that the minimum energy of the proton beam should be about 20 GeV, for otherwise the kaon yield would be severely reduced. At the proposed KEK facility, the total proton beam energy would be $(2.5+20)$ GeV. A 70 GeV proton synchrotron could be installed, at a later stage, inside the Tristan ring in order to increase the proton beam energy -albeit at the cost of a considerably lower beam power. For a given kaon yield, the required beam power would be lowest at $E_b = 30 - 100$ GeV [33].
3.4. A Novel g-2 Experiment with Ultra-Slow Muons

A charged elementary fermion has a magnetic dipole moment \( \mu = g_s (q/2m) s \) aligned with its spin \( s \). The proportionality constant \( g_s \) is the Landé g-factor, \( q \) is the charge of the particle and \( m \) is its mass. Dirac’s theory of the electron predicts that \( g_s = \frac{2}{3} \). For the electron \((e)\), muon \((\mu)\) and tau lepton \((\tau)\), this prediction differs from the observed value by a small fraction of a percent. The difference is the anomalous magnetic moment; the anomaly is defined by \( a \equiv (g_s - 2)/2 \sim 10^{-3} \). In the Standard Model, three distinct classes of Feynman diagrams contribute to the value of the anomaly for each lepton species: (1) the dominant QED terms that contain only leptons and photons; (2) terms that involve hadrons; and (3) electroweak terms containing the Higgs, W and Z bosons. The muon anomaly \( a_\mu \) is about \( (m_\mu/m_e)^2 \sim 43000 \) times more sensitive to the existence of yet unknown heavy particles than the electron anomaly \( a_e \). The value of \( a_\mu \) (\( a_e \)) is sensitive to new physics at the scale of a few hundred GeV (MeV) \([36]\).

The current experimental uncertainty on \( a_\mu \) is \( \pm 0.54 \) ppm. In a novel g-2 experiment \([37]\), the main aim of which is to reduce this uncertainty to \( \pm 0.1 \) ppm, 3 GeV protons impinge on a graphite target and produce pions that are stopped in the target. Some of the positive pions are brought to rest near the surface of the target, where they decay into positive muons with momenta \( p_\mu = 30 \text{ MeV}/c \) and 100% spin polarization. The muons are collected using a large-aperture solenoid and transported to a silica-aerogel target in which they form muonium (electron—\( \mu^+ \)) atoms. As the atoms slowly diffuse from the target, they are ionized by a pulsed laser to produce 50% polarized muons with very low momenta\(^1\). Those “ultra-slow” muons \((4 \times 10^4/\text{pulse})\) are then accelerated to \( p_\mu = 300 \text{ MeV}/c \) by two linacs, and injected into a magnetic storage ring that contains a 3 T solenoid with a diameter of 66 cm. After injection, the muons circulate orthogonal to the magnetic field \( B \). An orbiting muon decays within 6.6 \( \mu \) s into a positron, a neutrino and an antineutrino:

\[
\mu^+ \rightarrow e^+ + \nu_e + \bar{\nu}_\mu.
\]

The experiment is also designed to measure the electric dipole moment \( d_\mu = \eta (q\hbar/2m) s \) of the muon (see below).

The highest-energy positrons, preferentially emitted parallel to the muon spin direction in the \( \mu^+ \) rest frame, are Lorentz-boosted to become the highest-energy positrons in the lab frame. Hence, the angular distribution of those positrons has its maximum in the direction of the muon spin \([17]\). By measuring the energy and time distributions of positrons one can determine the average spin direction. The time spectrum will show the muon lifetime modulated by the spin precession frequency. The relative precession of the spin with respect to the direction of the particle velocity \( u \) is given by

\[
\omega_s + \omega_\eta \propto a_\mu B - (\eta/2)(\beta \times B)
\]

where \( \omega_s \) and \( \omega_\eta \) arise from \( a_\mu \) and \( d_\mu \), respectively, and \( \beta \equiv u/c \).

\(^1\)A much higher level of polarization can be obtained by using a magnetic field to align the particle spins \([37]\).
Since the rotation axes due to $a_\mu$ and $d_\mu$ are orthogonal, the corresponding signals can be separated [37]. In the case of $\mu_\mu$, the anomalous precession period is $2.2 \mu s$, about 300 times the cyclotron period. Assuming that muons are 100% polarized, $1.5 \times 10^7$ positrons have to be detected for a measurement precision of 0.1 ppm [37].

4. An XFEL Based on the Proposed Superconducting Linac

To record the dynamics of atoms requires a probe with Ångstrom ($10^{-10}$ m) wavelength and femtosecond temporal duration ($10^{-15}$ s). Such probes have recently become available with the advent of X-ray free-electron lasers (XFELs). The ultrashort pulse duration of an XFEL matches the timescale of non-equilibrium microscopic processes such as electron transfer in molecules, evolution of chemical reactions, vibration dynamics in solid state systems, etc. Optical lasers are also capable of producing pulses of femtosecond duration, but lack the required spatial resolution.

The spectral brightness (or brilliance), $B$, of a radiation field can be expressed, in practical units, as the number of photons per second passing through a given cross-sectional area and within a given solid angle and spectral bandwidth (BW). This quantity determines how much monochromatic radiation can be focused onto a tiny spot on the target. The peak spectral brightness—the brightness measured during the very short duration of an FEL pulse—of the two presently most powerful XFEL facilities (LCLS at SLAC in the United States and SACLA at SPring-8 in Japan) is billion times higher than that of any synchrotron radiation source.

Owing to the high intensity of XFEL radiation, laser-irradiated atoms, molecules and atomic clusters can be excited into previously unknown states. Although high-intensity pulses may also destroy molecular structures, they can still be used to produce high-resolution X-ray diffraction patterns, from which real-space images of the atomic positions in molecules can be reconstructed. In a typical “pump-probe” experiment, the evolution of a chemical (or biochemical) reaction, initiated by an optical or IR laser pulse, is observed by a time-delayed X-ray pulse. By varying the delay, such stroboscopic measurements result in femtosecond “movies” of the evolving system. Note that, due to their long pulse duration, X-rays from synchrotron light sources can be used to image atomic structures only in static measurements.

Nanometre-scale molecular imaging is made possible by the high degree of coherence of the XFEL radiation. The coherence quality of a light source is best described by the degeneracy parameter $D$, defined as the number of emitted photons per coherent phase-space volume per coherence time:

$$D = \frac{B\lambda^3}{4c} \approx 8.3 \times 10^{-25} B\lambda^3$$

Here $B$ is the brilliance and $\lambda$ [0.1 nm] the wavelength of the source.

Free-electron lasing is achieved by a single-pass, high-gain FEL amplifier op-
erating in the so-called self-amplified spontaneous emission (SASE) mode. An FEL consists of an electron linear accelerator and an undulator, a long periodic array of magnets with period \( \lambda_u \). The wavelength of the first harmonic of the observed FEL radiation is given by [38] [39]

\[
\lambda = \frac{\lambda_u}{2\gamma^2 \left( 1 + \frac{K^2}{2} + \gamma^2 \theta^2 \right)}
\]

(12)

where the dimensionless quantity

\[
K = \frac{eB_p \lambda}{2\pi m_e c} = 0.934 B_p \lambda_u [\text{cm}]
\]

(13)

is the undulator deflection parameter and \( \gamma \) is the Lorentz factor of the electron beam. Typically, \( \lambda_u \approx 3 \text{ cm} \) and \( \gamma \approx 10^4 \); hence, \( \lambda \approx 0.1 \text{ nm} \).

The emission of radiation in an undulator does not occur at one wavelength, but in a wavelength band of width \( \Delta \lambda \) around the central value given by Equation (12). Each electron propagating through the undulator emits a wave train consisting of a number of wavelengths equal to the number of undulator periods, \( N_u \). The time duration \( \Delta t \) of this pulse is the pulse length \( L_p = N_u \lambda \) divided by the speed of light: \( \Delta t = N_u \lambda / c \). A pulse of duration \( \Delta t \) has a frequency bandwidth \( \Delta \nu \sim 1/\Delta t \). Hence, \( \Delta \nu \sim c/N_u \lambda = \nu/N_u \), because \( \lambda = c/\nu \). Thus,

\[
\frac{\Delta \nu}{\nu} = \frac{\Delta \lambda}{\lambda} = \frac{1}{N_u} \approx 10^{-3}
\]

(14)

The wave train is not monochromatic due to its finite length. For typical values \( N_u \approx 10^4 \) and \( \lambda \approx 0.1 \text{ nm} \), one obtains \( \Delta t \approx 0.33 \text{ fs} \). Since the electrons are distributed throughout a bunch, the pulse duration is increased to \( \tau_p \sim (\sigma_z/L_p) \Delta t \approx 80 \text{ fs} \), where \( \sigma_z \approx 25 \mu\text{m} \) is the bunch length and \( L_p \approx 0.1 \mu\text{m} \). An 80 fs pulse, therefore, consists of many micropulses of 0.33 fs duration.

The 'shot noise' in an electron beam, the origin of which is the random emission of the electrons from a photocathode, causes random fluctuations of the beam density. The radiation produced by such a beam has amplitudes and phases that are random in both space and time. For this reason, SASE X-ray FELs lack longitudinal (or temporal) coherence, characterized by the coherence length \( L_{\text{coh}} = \lambda^2 / \Delta \lambda \approx 0.1 \mu\text{m} \). This quantity is defined as the distance of propagation over which radiation with spectral width \( \Delta \lambda \) becomes 180° out of phase. The coherence time, defined by \( t_{\text{coh}} = L_{\text{coh}}/c \sim 1/\Delta \nu \), is much shorter than the pulse duration: \( t_{\text{coh}} \approx 0.3 \text{ fs} \).

In order to increase the coherence length in the hard X-ray regime (photons with 0.1 nm wavelength), a “self-seeding” method was tested at LCLS [40]. FEL pulses, generated in the first modular section of the LCLS undulator, are spectrally “purified” by a crystal filter (a diamond monochromator). Since a typical monochromator delays X-rays, the electron bunches exiting the first modular section are appropriately delayed after being diverted around the crystal by a
compact magnetic chicane (see Figure 1 in [40]). The crystal selects a very narrow part of the spectrum, which is further amplified in the second undulator section where the FEL radiation reaches saturation. At LCLS, “self-seeding” generated X-ray pulses with $\Delta \nu = 0.4 - 0.5$ eV at $\nu = 8 - 9$ keV, which represents a factor of 40 - 50 bandwidth reduction with respect to SASE [40].

The European XFEL as a Prototype for the Proposed X-Ray FEL

The European XFEL, currently under construction at DESY (Germany), could serve as a prototype for the proposed X-Ray FEL at KEK. The XFEL at DESY is a free-electron laser based on self-amplified spontaneous emission (SASE) in the X-ray regime. The FEL consists of a 17.5 GeV superconducting electron linear accelerator and a set of undulators that can generate both SASE FEL X-rays and incoherent radiation. A schematic layout of the European XFEL is shown in [41]. Electron bunches, each with a charge of 1 nC, are extracted from a photocathode by short ultraviolet laser pulses and then focused and accelerated inside a radio-frequency cavity (“RF gun”) to an energy of 120 MeV. In order to produce 5 kA peak currents necessary for lasing, the bunches are further accelerated and longitudinally compressed down to 25 µm using two magnetic chicanes (at 0.5 and 2.0 GeV). After traversing the main linac, where their energy is increased to 17.5 GeV, the bunches are sent through a number of FEL undulators.

A superconducting linac may accelerate 10 “bunch trains” per second, each train consisting of up to 2700 electron bunches. This results in 27,000 ultrashort X-ray flashes per second. The higher the number of electron bunches, the more scientific instruments can be operated simultaneously. The European XFEL facility will generate ultra-short pulses ($\leq 100$ fs) of spatially and temporally coherent X-rays with wavelengths in the range ~0.1 - 5 nm; its peak brilliance is expected to be

$$B \approx 5 \times 10^{33} \text{ photons/second/mm}^2/\text{mrad}^2/0.1\% \text{ BW}$$

The coherent superposition of the radiation fields from all microbunches within an electron bunch in the linac is responsible for the nearly monochromatic spectrum and small divergence of the radiation emitted in the forward direction [38]. Recall also that “self-seeding” can substantially improve longitudinal (temporal) coherence of SASE XFEL radiation. Thus, the radiation from an X-ray FEL has a narrow bandwidth, is transversely and longitudinally coherent, and is fully polarized. The coherently emitted XFEL spectral lines appear in addition to the spontaneously emitted undulator spectrum that extends into the MeV energy region (see, e.g., [42]).

The micropulses that form an FEL pulse give rise to “spikes” in the spectrum. The amplitudes of the micropulses vary greatly as a consequence of the amplified stochastic variations in the electron density. Within a micropulse, the radiation is both transversely and longitudinally coherent. The duration of a micropulse is roughly $t_{\text{coh}}$, the coherence time. In the SASE 1 and SASE 2 undulators at the European XFEL, $t_{\text{coh}} = 0.2 - 0.38$ fs [41]. The number of “spikes” in a
pulse is given by the ratio of the bunch length to the coherence length: 
\[ \sigma_z / L_{coh} = (25 \mu m)/(0.1 \mu m) \approx 250. \]

The spectrum of undulator radiation is sharply peaked around odd harmonics\(^4\) (see, e.g., [42]). The photon energy that corresponds to the \(n\)th harmonic is given by

\[
E_n[\text{keV}] = 0.9496 \frac{nE_e^2[\text{GeV}]}{\lambda_u[\text{cm}](1 + K^2/2 + \gamma^2 \theta^2)} \tag{16}
\]

where \(E_e\) is the electron beam energy and \(\theta\) is the radiation detection angle (with respect to the forward direction). For \(\theta = 0\) and the SASE 2 undulator parameters \(\lambda_u = 4.8 \text{ cm}\) and \(K = 6.1\), for example, Equation (16) yields \(E_i = 12.2 \text{ keV}\) [41].

At the exit of the SASE 1 undulator, the photon beam divergence is \(\sigma_\theta \sim 1/\gamma \sqrt{N_u} \approx 1 \mu\text{rad}\) [38] and the beam size is \(70 \mu\text{m} \times 70 \mu\text{m}\), the diameter of a fine needle. This beam can be focused to an area of \(0.1 \mu\text{m} \times 0.1 \mu\text{m}\) (the size of a virus) at an experimental station located a couple of hundred meters from the undulator exit [43]. Through variable focusing, the flux density of an XFEL beam can therefore be tuned by a factor of about one million. The SASE 1 undulator will deliver \(10^{12}\) photons in an ultra-short pulse of 100 fs duration (the timescale of molecular vibrations), yielding a peak power of about 20 GW at a photon energy \(E \approx 12 \text{ keV}\) (the photon wavelength is \(\lambda = h\gamma/E \approx 0.1 \text{ nm}\)).

Recall that all photons in a single micropulse are completely coherent. Since each pulse contains \(\sim 10^{12}\) photons and a few hundred micropulses, there are \(10^9\) indistinguishable (“degenerate”) photons in the coherence volume. In comparison, the degeneracy parameter \(D \approx 0.03\) at a synchrotron source with \(\lambda = 0.1 \text{ nm}\). Because of the large transverse coherence area of \(70 \times 70 \mu\text{m}^2\) and the large number of coherent photons per pulse, an interference (“speckle”) pattern can be recorded with a single XFEL pulse [44]. Therefore, X-ray FELs can be used not only to probe the structure of matter down to the size of an atom, but also to take ‘snapshots’ of the motion of atoms and molecules [45].

5. Summary

The main “bottleneck” limiting the beam power in circular machines is caused by space charge effects that produce beam instabilities. Such a “bottleneck” exists at the J-PARC proton synchrotron complex, and is also intrinsic to the “proton drivers” envisaged at CERN and Fermilab. In order to maximally increase the beam power of a “proton driver”, it is proposed to build a facility consisting of a low-energy injector linac (PI) and a high-energy pulsed superconducting L-band linac (SCL). The 2.5 GeV PI, based on the European Spallation Source (ESS) linac, would serve both as an injector to the SCL and a source of proton beams that could be used to copiously produce muons and “cold” neutrons. Protons

\(^4\)The occurrence of higher harmonics is explained in [38]. In the forward region (\(\theta = 0\)) of a planar undulator, only the odd higher harmonics are observed, while the off-axis radiation contains also the even harmonics. For the European XFEL, simulations predict that the relative contribution to the total radiation power of the 3rd and the 5th harmonic is about 1% and 0.03%, respectively [41].
accelerated by the SCL to 20 GeV would be transferred through the KEK Tristan ring in order to create neutrino, kaon and muon beams for fixed-target experiments. At a later stage, a 70 GeV proton synchrotron could be installed inside the Tristan ring. The proposed facility would be constructed using the existing KEK accelerator infrastructure and the most advanced linac technologies currently available.

Experiments with high-intensity neutrino beams are designed primarily to explore the mass spectrum of the neutrinos and their properties under the CP symmetry. To search for CP violation in neutrino oscillations, a 100 kt water Cherenkov detector could be built at Okinoshima, located at a distance \( L_2 \approx 650 \text{ km} \) from KEK. Using the proposed KEK “proton driver”, the detector at Okinoshima and the existing 0.022 Mt Super-Kamiokande detector (placed \( L_1 \approx 300 \text{ km} \) away from KEK), the neutrino mass hierarchy could be determined either by comparing the \( \nu_e \) appearance probabilities measured at the two vastly different baseline lengths \( L_1 \) and \( L_2 \), or by measuring at \( L_1 \) and \( L_2 \) the neutrino energy of the first oscillation maximum. Once the mass hierarchy is determined, the CP-violating phase in the mixing matrix can be measured with a precision of \( \pm 20^\circ \), assuming that \( 2.5 \times 10^{21} \) protons are delivered on target for both \( \nu_e \) and \( \bar{\nu}_e \) beams.

Some of the most important discoveries in particle physics emerged from high-precision studies of K mesons (“kaons”), in particular neutral kaons. A deeper insight into CP violation is expected to be gained from measurements of ultra-rare kaon decays such as \( K^0 \to \pi^0 \nu \bar{\nu} \) and \( K^+ \to \pi^+ \nu \bar{\nu} \). These decays provide important information on higher-order effects in electroweak interactions, and therefore can serve as a probe of new phenomena not predicted by the Standard Model.

A unique feature of the proposed facility is the use of superconducting ILC-type cavities to accelerate both protons and electrons, which considerably increases its physics potential. Polarized electrons and positrons can be employed to study the structure of composite particles and the dynamics of strong interactions, as well as to search for new physics beyond the Standard Model.

An SCL-based X-ray free-electron laser (XFEL) and a synchrotron light source for applications in materials science and medicine are also envisaged. The ultra-short pulse duration of an XFEL matches the timescale of non-equilibrium microscopic processes, allowing the dynamics of atoms and molecules to be recorded in the form of femtosecond “movies”.

Acknowledgements

For useful comments and suggestions concerning various aspects of this work, I am grateful to K. Fujii, K. Hagiwara, T. Higo, E. Kako, Y. Kamiya, K. Oide, K. Takayama and K. Tokoshuku. I wish to express my special gratitude to Kaoru Yokoya for his invaluable help and encouragement.

This paper is an abridged version of KEK Preprint 2014-35 (2014), which was subsequently published as an arXiv e-print (arXiv:1411.4874).
References

[1] Aad, G., et al. (2012) Observation of a New Particle in the Search for the Standard Model Higgs Boson with the ATLAS Detector at the LHC. *Physics Letters B*, **716**, 1-29. https://doi.org/10.1016/j.physletb.2012.08.020

[2] Chatrchyan, S., et al. (CMS Collab.) (2012) Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC. *Physics Letters B*, **716**, 30-61. https://doi.org/10.1016/j.physletb.2012.08.021

[3] Yano, A., et al. (2004) The Toshiba E3736 Multi-Beam Klystron. *Proceedings of LINAC* 2004, Lubeck, Germany, 706-708.

[4] Gelvich, E., et al. (1993) The New Generation of High-Power Multiple-Beam Klystrons. *IEEE Transactions on Microwave Theory and Techniques*, **41**, 15-19. https://doi.org/10.1109/22.210224

[5] Yamamoto, A. (2009) Superconducting RF Cavity Development for the International Linear Collider. *IEEE Trans. Appl. Supercond.*, **19**, 1387-1393.

[6] Ostroumov, P.N. (2006) Physics Design of the 8 GeV H-Minus Linac. *New Journal of Physics*, **8**, 281. https://doi.org/10.1088/1367-2630/8/11/281

[7] Reschke, D. (2013) Infrastructure, Methods and Test Results for the Testing of 800 Series Cavities for the European XFEL. *Proceedings of SRF* 2013, Paris, France, 812-815.

[8] Faircloth, D. (2013) Ion Sources for High-Power Hadron Accelerators. arXiv:1302.3745

[9] Kapchinsky, I.M. and Tepliakov, V.A. (1970) Linear Ion Accelerator with Spatially Homogenous Focusing. *Pribory i Tekhnika Eksperimenta*, **2**, 19-22.

[10] Vretenar, M. (2012) Low-Beta Structures. CERN Yellow Report 2011-007, 319-340. arXiv:1201.2593

[11] Danared, H., Lindroos, M. and Theroine, C. (2014) ESS: Neutron Beams at the High-Intensity Frontier. CERN Courier, June 2014, 21-24.

[12] Garoby, R., et al. (2013) Proton Drivers for Neutrino Beams and Other High Intensity Applications. *Journal Physics: Conference Series*, **408**, 012016. https://doi.org/10.1088/1742-6596/408/1/012016

[13] Baussan, E., et al. (2014) A very Intense Neutrino Super Beam Experiment for Leptonic CP Violation Discovery Based on the European Spallation Source Linac. *Nuclear Physics B*, **885**, 127-149. https://doi.org/10.1016/j.nuclphysb.2014.05.016

[14] Al-Binni, U., et al. (2013) Project-X, Physics Opportunities. arXiv:1306.5009v2

[15] Dudek, J., et al. (2012) Physics Opportunities with the 12 GeV Upgrade at Jefferson Lab. arXiv:1208.1244v2

[16] Voutier, E. (2014) Physics Perspectives at JLab with a Polarized Positron Beam. arXiv:1402.2814v1

[17] Belusevic, R. (2008) Relativity, Astrophysics and Cosmology. Wiley-VCH Verlag., Weinheim.

[18] Shaposhnikov, M. (2009) Baryon Asymmetry of the Universe and Neutrinos. *Progress of Theoretical Physics*, **122**, 185-203. https://doi.org/10.1143/PTP.122.185

[19] Hagiwara, K., et al. (2012) Physics Potential of Neutrino Oscillation Experiment with a Far Detector in Oki Island along the T2K Baseline, KEK-TH-1568. arXiv:1209.2763v1

[20] Back, J., et al. (2013) Particle Production and Energy Deposition Studies for the Neutrino Factory Target Station. *Physical Review Accelerators and Beams*, **16**, 021001. https://doi.org/10.1103/PhysRevSTAB.16.021001
[21] McDonald, K.T., et al. (2010) The MERIT High-Power Target Experiment at the CERN PS. *Proceedings of First International Particle Accelerator Conference*, Kyoto, Japan, 3527-3529.

[22] Baussan, E., et al. (2013) Target, Magnetic Horn and Safety Studies for the CERN to Frejus Super Beam. *Journal of Physics: Conference Series*, 408, 012061. https://doi.org/10.1088/1742-6596/408/1/012061

[23] Kamiya, Y. DC Neutrino Horn. The manuscript (in Japanese) has not been published.

[24] Omori, T., et al. (2006) Efficient Propagation of Polarization from Laser Photons to Positrons through Compton Scattering and Electron-Positron Pair Creation. *Physical Review Letters*, 96, 114801. https://doi.org/10.1103/PhysRevLett.96.114801

[25] Grames, J., et al (2013) PEPPo: Using a Polarized Electron Beam to Produce Polarized Positrons. 15th *International Workshop on Polarized Sources, Targets and Polarimetry*, Charlottesville, Virginia, USA.

[26] Zel’dovich, Y.B. (1959) Parity Nonconservation in the First Order in the Weak-Interaction Constant in Electron Scattering and Other Effects. *Journal of Experimental and Theoretical Physics (JETP)*, 36, 964-966.

[27] Derman, E. and Marciano, W.J. (1979) Parity Violating Asymmetries in Polarized Electron Scattering. *Annals of Physics*, 121, 147-180.

[28] Czarnecki, A. and Marciano, W.J. (1996) Electroweak Radiative Corrections to Polarized Moller Scattering Asymmetries. *Physical Review D*, 53, 1066-1072. https://doi.org/10.1103/PhysRevD.53.1066

[29] Mammei, J. (2012) The MOLLER Experiment. *Nuovo Cimento*, 35, 203-208.

[30] Kumar, K.S., et al. (2013) Low-Energy Measurements of the Weak Mixing Angle. *Annual Review of Nuclear and Particle Science*, 63, 237-267. https://doi.org/10.1146/annurev-nucl-102212-170556

[31] Belusevic, R. (1999) Neutral Kaons. Springer-Verlag, Berlin.

[32] Brod, J., Gorbahn, M. and Stamou, E. (2011) Two-Loop Electroweak Corrections for the $K \to \pi \nu \bar{\nu}$ Decays. *Physical Review D*, 83, 034030. https://doi.org/10.1103/PhysRevD.83.034030

[33] Belyaev, A., et al. (2001) Kaon Physics with a High Intensity Proton Driver, CERN-TH/2001-175. arXiv:hep-ph/0107046v2

[34] Buras, A. (1999) CP Violation and Rare Decays of K and B Mesons. arXiv:hep-ph/9905437

[35] Artamonov, A., et al. (2009) Study of the Decay $K^{+} \to \pi^{+} \nu \bar{\nu}$ in the Momentum Region $140 < p_{\nu} < 199$ MeV/c. *Physical Review D*, 79, 092004. https://doi.org/10.1103/PhysRevD.79.092004

[36] Miller, J., de Rafael, E. and Roberts, B.L. (2007) Muon ($g$-2): Experiment and Theory. *Reports on Progress in Physics*, 70, 795-881. https://doi.org/10.1088/0034-4885/70/5/R03

[37] Mibe, T. (2010) New g-2 Experiment at J-PARC. *Chinese Physics C*, 34, 745-748.

[38] Schmuser, P., Dohlus, M. and Rossbach, J. (2008) Ultraviolet and Soft X-Ray Free-Electron Lasers. Springer-Verlag, Berlin.

[39] Huang, Zh. and Kim, K. (2007) Review of X-Ray Free-Electron Laser Theory. *Physical Review Accelerators and Beams*, 10, 034801. https://doi.org/10.1103/PhysRevSTAB.10.034801

[40] Amann, J., et al. (2012) Demonstration of Self-Seeding in a Hard-X-Ray Free-Electron Laser. *Nature Photonics*, 6, 693-698. https://doi.org/10.1038/nphoton.2012.180
[41] Abela, R., et al. (2007) The European X-Ray Free-Electron Laser. Technical Design Report, DESY 2006-097.

[42] Arthur, J. (1999) Prospects for an X-Ray FEL Light Source and Some Possible Scientific Applications, SLAC-PUB-8276.

[43] Agapov, I., et al. (2014) SASE Characteristics from Baseline European XFEL Undulators in the Tapering Regime. *Proceedings of FEL* 2014, Basel, Switzerland, 159-163.

[44] Pellegrini, C. and Stohr, J. (2003) X-Ray Free-Electron Lasers—Principles, Properties and Applications. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 500, 33-40. [https://doi.org/10.1016/S0168-9002(03)00739-3](https://doi.org/10.1016/S0168-9002(03)00739-3)

[45] Barty, A. (2010) Time-Resolved Imaging Using X-Ray Free Electron Lasers. *Journal of Physics B: Atomic, Molecular and Optical Physics*, 43, 194014. [https://doi.org/10.1088/0953-4075/43/19/194014](https://doi.org/10.1088/0953-4075/43/19/194014)
Modeling for Collapsing Cavitation Bubble near Rough Solid Wall by Multi-Relaxation-Time Pseudopotential Lattice Boltzmann Model

Minglei Shan1,2, Yipeng Zhu1, Cheng Yao1,2, Qingbang Han1,2, Changping Zhu1,2

1College of Internet of Things Engineering, Hohai University, Changzhou, China
2Jiangsu Key Laboratory of Power Transmission and Distribution Equipment Technology, Hohai University, Changzhou, China

Email: shanming2003@126.com

Abstract

Cavitation bubble collapse near rough solid wall is modeled by the multi-relaxation-time (MRT) pseudopotential lattice Boltzmann (LB) model. The modified forcing scheme, which can achieve LB model's thermodynamic consistency by tuning a parameter related with the particle interaction range, is adopted to achieve desired stability and density ratio. The bubble collapse near rough solid wall was simulated by the improved MRT pseudopotential LB model. The mechanism of bubble collapse is studied by investigating the bubble profiles, pressure field and velocity field evolution. The eroding effects of collapsing bubble are analyzed in details. It is found that the process and the effect of the interaction between bubble collapse and rough solid wall are affected seriously by the geometry of solid boundary. At the same time, it demonstrates that the MRT pseudopotential LB model is a potential tool for the investigation of the interaction mechanism between the collapsing bubble and complex geometry boundary.

Keywords
Cavitation Bubble, Bubble Collapse, Lattice Boltzmann Method, Pseudopotential Model, Rough Solid Wall

1. Introduction

Cavitation is ubiquitous in liquid, and happens when the local pressure is below the saturated vapor pressure. As cavitation bubble collapse near a solid wall, the associated phenomena include instant high pressure, high velocity jets and high temperature, which closely relate with the cavitation erosion of the solid material surface. On the other side, the collapse of the cavitation bubble has been applied
in environmental protection, ultrasonic therapy, lab on chip and material surface cleaning [1] [2].

The mechanism of the cavitation bubble collapse near solid wall is a fundamental issue for the above applications. However, as too many phenomena are involved, theoretical model of cavitation bubble collapse is difficult to be established, and for complex geometry boundary, the analytical solution is even impossible. Many publications are devoted to the investigation of the bubble collapse near the planar wall starting from the experimental work [3]. In recent years, many works have been developed to investigate the interaction between collapsing bubble and non-planar solid wall [1] [4], and numerical methods are becoming more and more important tools to study the bubble collapse near solid wall [5] [6]. Commonly used numerical methods include the finite volume method (FVM) [4], the finite element method (FEM) [7], and the boundary element method (BEM) [8]. These macroscopic numerical modeling methods based on solving partial differential equations are limited in processing the multiphase flows and complex geometry boundary conditions. As describing the multiphase flow, macroscopic methods need the assistance of the schemes of the interface tracking or capturing, which will reduce the computational efficiency. For the complex geometry boundary, it is difficult and inefficient to implement by macroscopic methods.

Owing to the flexibility for complex geometry boundary and the simplicity of the algorithm, the lattice Boltzmann method (LBM) has been developed into a powerful tool for the flow, heat and mass transfer simulations relating with complex geometry boundary [9] [10]. In LBM community, many multiphase models have been presented, which can be generally classified as the color-gradient model [11], the pseudopotential model (or Shan-Chen model) [12] [13], the free-energy model [14] and the phase-field model. The pseudopotential model is widely and successfully used in the LBM multiphase community due to its conceptual simplicity and computation efficiency [15] [16]. In the pseudopotential model, the fluid interactions are mimicked by an interparticle potential, from which a non-monotomic equation of state (EOS) can be obtained. As a result, the separation of fluid phases or components can be achieved automatically in this method, and the methods to track or capture the interfaces are not required.

In recent years, the pseudopotential model, as the top choice of the multiphase LB model, was introduced into the issue of cavitation by Sukop and Or [17]. Subsequently, several research efforts emerged to investigate the mechanism of cavitation. Chen [18] simulated the cavitation bubble growth using the modified pseudopotential LB model with the exact difference method (EDM) force scheme. The results in quiescent flows agree fairly well with the solution of Rayleigh- Plesset equation. Mishra [19] introduced a model of cavitation based on the pseudopotential LB model that allows for coupling between the hydrodynamics of a collapsing cavity and supported solute chemical species. Unfortunately, the above researches do not involve the interaction between bubble and solid wall. Until, the pseudopotential model was introduced into the researches
on the mechanism investigations of cavitation bubble collapse near solid wall by Shan et al. [20] [21]. However, these previous works involve just planar wall. In engineering cases, the rough walls are ubiquitous. The mechanism investigation of cavitation bubble collapse near rough wall is the basic issue for the efficient applications of cavitation bubble collapse.

One of the key challenges of the simulation of cavitation bubble collapse near rough wall is the stability of the LB model. For pseudopotential multiphase model, the stability is closely related to the thermodynamic consistency [22], and great efforts have been made for this issue [23] [24] [25]. Recently, Li et al. [24] [25] [26] founded that there exists a suitable forcing scheme, which can meet the thermodynamic consistency requirement in an efficient way. In order to investigate the interaction mechanism between the collapsing bubble and complex geometry boundary, in the present work, the improved forcing scheme for the pseudopotential MRT LB model developed by Li et al. [25] is adopted to achieve the sufficient density ratio and stability of the numerical model.

2. Numerical Method

The pseudopotential LB model, also well known as Shan-Chen model, was developed by Shan and Chen in 1993 [12]. In pseudopotential model, the fluid interactions are mimicked by an interparticle potential, which is now widely called pseudopotential. In original pseudopotential LB model, the Single-Relaxation-Time (SRT) collision operator was employed. In recent years, the MRT collision operator has been verified that it is superior to the SRT operator in terms of numerical stability. The MRT LB evolution equation can be given as follows:

\[
f'_a(x + e_a \delta_t, t + \delta_t) = f_a(x, t) - (M^{-1} \Lambda^{\alpha\beta}(f_\beta - f_\beta^{eq}) + 2 \delta_t F'_a),
\]

where \( f_a \) is the density distribution function, \( f_\beta^{eq} \) is its equilibrium distribution, \( t \) is the time, \( x \) is the spatial position, \( e \) is the discrete velocity along the \( \alpha \)th direction, \( \delta_t \) is the time step, \( F'_a \) is the forcing term in the velocity space, and \( M \) is an orthogonal transformation matrix. \( \Lambda \) in Equation (1) is a diagonal matrix, and for D2Q9 lattice, \( \Lambda \) is given by

\[
\Lambda = \text{diag}(\tau_{\rho}, \tau_{\rho}, \tau_{\rho}, \tau_{\rho}, \tau_{\rho}, \tau_{\rho}, \tau_{\rho}, \tau_{\rho}, \tau_{\rho}).
\]

Through the transformation matrix \( M \), \( f \) and \( f_\alpha^{eq} \) can be projected onto the moment space via \( \mathbf{m} = Mf \) and \( \mathbf{m}^{eq} = Mf^{eq} \), and the collision step of MRT LB equation (Equation (1)) can be rewritten as [25]

\[
\mathbf{m}' = \mathbf{m} - \Lambda(\mathbf{m} - \mathbf{m}^{eq}) + \delta_t \left( \mathbf{I} - \frac{\Lambda}{2} \right) \mathbf{S}.
\]

where \( \mathbf{I} \) is the unit tensor, and \( \mathbf{S} \) is the forcing term in the moment space with \((\mathbf{I} - 0.5\Lambda)\mathbf{S} = \mathbf{M} \mathbf{F}'\). For D2Q9 lattice, \( \mathbf{m}^{eq} \) can be given by

\[
\mathbf{m}^{eq} = \rho \{ 1 - 3v_\rho^2, 1 - 3v_\rho^2, v_x, v_y, -v_y, v_x, v_y, -v_x, v_x, v_y \}^T.
\]

Here \( \rho = \sum_a f_a \) is the macroscopic density, \( v \) is the macroscopic velocity which satisfies \( v^2 = v_x^2 + v_y^2 \) and is calculated by
\[ \mathbf{v} = \sum_{a} \mathbf{e}_a f_a + 0.5 \delta^{\mathbf{F}} \rho, \]

where \( \mathbf{F} = (F_x, F_y) \) for two dimensional space is the force action on the fluid system. Then the streaming step of the MRT-LB equation can be formulated as

\[ f_a(x + e_a \delta^x, t + \delta^t) = f_a(x, t), \]

where \( f^* = M^{-1} m^* \).

For the pseudopotential LB model in D2Q9 lattice case, the \( \mathbf{F} \) in Equation (5) is given by

\[ \mathbf{F} = -G \psi(x) \sum_{a=4}^8 w_a \psi(x + e_a) \mathbf{e}_a, \]

where \( \psi(x) \) is the interparticle potential, \( G \) is the interaction strength, and \( w_{1,2,3,4} = 1/3 \) and \( w_{5,6,7,8} = 1/12 \) are the weights. In the present work, the form of \( \psi \) proposed by Yuan et al. [27] is adopted which can be formulated as

\[ \psi = \sqrt{\frac{2(p_{\text{EOS}} - p c_s^2)}{G c^2}}, \]

where \( p_{\text{EOS}} \) is the pressure calculated by equation of state (EOS). The \( G \) in Equation (8) loses the meaning of the interaction strength and is used to ensure that the whole term inside the square root is positive [27]. The Carnahan-Starling (CS) EOS is adopted in the present work, which can be given by [27]

\[ p_{\text{EOS}} = \rho RT \left(1 + \frac{b \rho / 4 + (b \rho / 4)^2 - (b \rho / 4)^3}{(1 - b \rho / 4)}ight) - a \rho^2, \]

where \( a = 0.4963 R^2 T_c^2 / p_c \) and \( b = 0.1873 R T_c / p_c \). Here \( T_c \) and \( p_c \) are the critical temperature and pressure, respectively.

\( \mathbf{F} \) in Equation (7) can be incorporated in evolution equation via \( \mathbf{S} \) with specific forcing scheme. Li et al. [25] proposed a MRT version forcing scheme to achieve thermodynamic consistency. For the D2Q9 lattice, Li’s forcing scheme can be given by

\[ \begin{bmatrix} 0 \\ 6(\mathbf{v}_x F_x + \mathbf{v}_y F_y) \frac{0.75 \varepsilon |\mathbf{F}|^2}{\psi \delta^x (\tau_x - 0.5)} \\ -6(\mathbf{v}_x F_x + \mathbf{v}_y F_y) \frac{0.75 \varepsilon |\mathbf{F}|^2}{\psi \delta^y (\tau_y - 0.5)} \\ F_x \\ -F_x \\ F_y \\ -F_y \\ 2(\mathbf{v}_x F_x - \mathbf{v}_y F_y) \\ \mathbf{v}_x F_y + \mathbf{v}_y F_x \end{bmatrix}, \]

where \( \varepsilon \) is a parameter related with the particle interaction range [28] and is proved by Li et al. to have the function of adjusting the thermodynamic consis-
Li’s forcing scheme has the following advantages: a) maintaining a uniform layout with the a general form of the LB forcing scheme; b) achieving thermodynamic consistency only by tuning one constant parameter; and c) fully retaining the LBM’s advantages of simple and efficient. In [21] it is found that the thermodynamic consistency is independent of kinematic viscosity for Li’s improved forcing scheme, and the surface tension is independent of the relaxation time \( \tau \). These features make it more convenient to investigate the physical mechanism of the multiphase flows.

Unless otherwise specified, the unit adopted in this paper is the lattice unit of LBM, i.e. the units of the length, time, mass, velocity, density and pressure are \( lu \) (lattice unit), \( ts \) (time step), \( mu \) (mass unit), \( mu \, lu^{-2} \), \( lu \, ts^{-1} \) and \( mu \, ts^{-2} \).

### 3. Cavitation Bubble Collapse near Rough Solid Wall

It is confirmed that the geometry is a crucial role on the process of the cavitation bubble collapse [29]. The interaction between collapse bubble and rough solid wall is a common issue abstracted from the cavitation applications involving the interactions between bubble and complex geometry boundary. In the present work, the rough solid wall is described by the wall with the geometries of periodic grooves with equal widths as shown in Figure 1. In this diagram, \( D_g \) and \( W_g \) are the depth and the width of groove, respectively. \( W_b \) is the width of bulge, and \( P_{gb} = W_g + W_b \) is the period width of the periodically arranged geometry.

The computational domain is shown in Figure 2. \( R_0 \), \( P_r \) and \( P_\infty \) are the initial radius, the pressure inside bubble and the ambient pressure, respectively. \( d_g \) and \( d_b \) are the distance from bubble center to groove bottom and bulge top, respectively. Based on the concept of the stand-off parameter [30], two non-dimensional parameters are introduced as follows to describe the positional relations between bubble and rough solid wall:

\[
\lambda_b = \frac{d_b}{R_0}, \tag{11}
\]

\[
\lambda_g = \frac{d_g}{R_0}. \tag{12}
\]
Here, $\lambda_b$, $\lambda_g$ express the stand-off distances between bubble and bulge top, groove bottom, respectively.

The pressure boundary conditions, implemented by the nonequilibrium extrapolation scheme [31], are adopted in the directions of left and right. And, a constant pressure boundary condition is adopted in the top of computational domain by Zou-He scheme [32].

A $401 \times 401$ lattice is adopted in the simulations of this section. A vapor bubble with radius of $R = 80$ is initially placed at the center of the domain. The density field is initialized as [33]

$$\rho(x, y) = \frac{\rho_l + \rho_v}{2} + \frac{\rho_l - \rho_v}{2} \tanh \left( \frac{2\left(\sqrt{(x-x_0)^2 + (y-y_0)^2} - R_0\right)}{W} \right),$$

(13)

where $(x_0, y_0)$ is the center of the bubble, $W$ is the prescribed width of the phase interface and is set as 5 in present work, \tanh is the hyperbolic tangent function. The parameters of solid wall geometry are set as $D_s = 144$, $W_b = W_g = 10$, $W_p = 20$, $d_s = 160$ and $d_b = 15$, respectively.

The parameters for the present MRT pseudopotential LB model are chosen as follows: $r_\rho = r_f = 1.0$, $r_v = r_g = 1.0$, $r_\eta = 1.1$, $\tau_v = 0.6$ and $\varepsilon = 1.86$. For CS EOS, $a = 0.5$, $b = 4$ and $R = 1$. The temperature $T = 0.7T_c$. In order to simulate the bubble collapse process, a positive pressure difference $\Delta P = P_\infty - P_v$ is achieved by artificially tuning the initial liquid density based on the equilibrium state. In this section, the pressure difference is $\Delta P = 0.0116$.

4. Results

4.1. Evolution of Density Field

The evolution of density field is shown in Figure 3. The deformation of bubble profile can be investigated from density field. The initial spherical bubble begins
to collapse motivated by the pressure difference from \( t = 0 \). The velocity of collapse is slow at the starting stage. Until \( t = 202 \), the bubble profile appears obvious deformation, \textit{i.e.}, the radius of curvature reduce towards the rough solid wall. Then, the bubble becomes an ellipsoid and the density just over the bubble increases. Along with the diffusion of the higher density area, the top of the bubble is flatten, and then concaved directing the solid wall. At the same time, the more dense and concentrated density area appears at the concave and accelerate the shift of the bubble profile. As a result, the jet is formed. The first collapse happens when the upper bubble wall clashes the bottom wall. The jet perforates the bottom wall and produces a crucial bubble. The shock wave generated by the first collapse propagates rapidly towards the rough solid wall and bounce back. The bubble experiences the complex and volatile deformation at the last stage under the jet which includes shock wave and bounce wave.

\textbf{Figure 3} describes the whole process of bubble collapse with time \( t \). Comparing with the flat solid wall case, the bubble collapse near rough solid wall appears the similar dynamics process \cite{20}. The differences are mainly reflected in two
aspects. The first one is the bounce of the shock wave in rough solid wall case is weaker than the one in the planar wall case. The strength and the reflection path of the shock wave are affected by the geometry of the rough solid wall. The second one is the geometry of the rough solid wall affects the bubble deformation, especially at the last stage. In order to be more intuitive, pressure field evolution of the collapsing bubble is investigated in the next section.

4.2. Evolution of Pressure and Velocity Field

Several representative moments of the pressure and velocity evolution are shown in Figure 4. At $t = 528$, one high pressure area emerges at the concave of bubble top. It is consistent with the analysis of the density distribution at the same moment. At the same time, the fluid velocity is increased towards the solid wall which formats the inceptive jet. The jet velocity at the concave is significantly higher than that of other areas in the moment. Along with the sagging of the upper bubble wall, the jet velocity is higher and higher. At $t = 717$, the first collapse occurs. The clash between the upper and the bottom walls induces shock wave as shown in $t = 741$. And at the same moment, the bubble is broken by the high velocity jet. Affected by the vortexes, which are induced by the jet and the bouncing back shock wave, the crucial bubble deforms and then collapses close to the rough solid wall. From Figure 4, we can find that the shock waves bounce intensively at the solid surfaces of the bulges top, and propagate unceasingly into the grooves with decreasing strength. The fluid jet follows the similar mode. It is inevitable that the shock wave and the jet will erode and impact the rough solid wall.

Figure 4. Pressure and velocity field evolution of the collapsing bubble near rough solid wall.
Figure 5 gives the finer detail display of the interaction between the pressure, velocity and the rough solid wall at \( t = 998 \). It is found that in the groove and near the bulge, which are off-center the collapse position, the vortexes are formed. Conversely, the laminar flows are formed in the groove just under the collapse position. The vortexes and the laminar flows induce the remarkable eroding effect on the solid surface of the bulge and the groove, respectively.

4.3. Analysis of Eroding Effect

In order to investigate the eroding process and effect induced by bubble collapse close to the rough solid wall, four test sections are set at the bulges top and the grooves side walls as shown in Figure 5. The coordinates of these four test sections are \( a \mid x = 191, y \in [257, 270] \), \( b \mid x = 211, y \in [257, 270] \), \( c \mid y = 256, x \in [200, 210] \) and \( d \mid y = 256, x \in [220, 230] \), respectively. The pressure evolution and the velocity evolution of these four test sections, from 730 ts to 1000 ts, are displayed in Figure 6 and Figure 7.

Form Figure 6, we can find that the test section \( a \) experiences three pressure peaks during the time interval of observation. The closer to the groove entrance, the greater the pressure amplitude is. The first pressure peak with the greatest amplitude is the result of shock wave. The second is the water hammer effect of jet. However, as the test section is parallel with jet direction, the pressure amplitude is weaker than the first pressure peak. The third pressure peak is induced by the second collapse. The pressure evolution at the test section \( b \) displays similar pattern. Due to the impeding effect of the bulge, the lower pressure area is induced near the test section \( b \). The test section \( c \) is just under the collapse position, so the three pressure peaks are greater than the test section \( a \) and \( b \). The pressure pattern of the test section \( d \) is similar as that of \( b \) on account of the larger deviation.
Figure 6. Pressure evolution at the test sections.

Figure 7. Velocity evolution at the test sections.
from the collapse station. But the pressure amplitudes are higher than the test section b because of the closer distance from the solid wall. The pressure evolution at the four test sections demonstrates that no matter the bulge solid wall or the groove entrance experience the oscillation of pressure.

The velocity evolution of the four test sections are shown in Figure 7. For the test section a, every test point experiences the similar process of the fluid velocity, which increases sharply and then gradually rises to the peak. The pattern of the velocity is same as the pressure, i.e., the closer to the groove entrance, the greater the velocity is. It demonstrated that the groove geometry impedes both of the pressure and the velocity at the groove side walls. Although the same near the entrance, the test section b shows a different pattern. Under the combined action of shock wave and jet, the vortexes are induced at b. The velocity evolution pattern of the test section c is symmetric in x direction. Due to the non-slid effect of solid wall, the fluid velocity at center of the bulge is lower than the sides, which are closer to the groove entrances. For the test section d, the velocity pattern is more diversified with lower fluid velocity than other test sections. The oscillation of velocity is more violent for farther distance from collapse position. The last velocity peaks are related with the collapse of the crucial bubble.

In summary, the eroding effect at the solid wall of the bulge part is weaker than that at the side walls, and is more obvious at the side walls of the grooves than other sections. From the pressure and velocity evolution patterns, it demonstrates that the fluid velocity accounts for the eroding effect at the side walls of the grooves, and the pressure accounts for the eroding effect at the bulge part of the rough solid wall.

5. Conclusions

For the modeling of the collapsing bubble near rough solid wall, an improved MRT pseudopotential LB model was adopted with the modified forcing scheme by Li et al. Then the bubble collapse near rough solid wall was simulated. The bubble collapse mechanism was investigated from the dynamic process including bubble profiles evolution, pressure and velocity field evolution. The eroding effects of shock wave and jet were analyzed in details. It is found that the process and the effect of bubble collapse are affect seriously by the geometry of solid boundary. In the present work, the interaction between the collapsing bubble and the geometries boundary structured with periodic grooves was investigated. We found that the fluid velocity is the major cause for the eroding effect at the side walls of the grooves, and the pressure is the major cause for the eroding effect at the bulge part of the rough solid wall.

The improved forcing scheme proposed by Li et al. provides a convenient and efficient approach to achieve thermodynamic consistency. By the modelling of the bubble collapse near the rough solid wall, it is demonstrated that the MRT pseudopotential LB model improved by the modified forcing scheme has enough stability to describe the whole process of bubble collapse and the interaction between the collapsing bubble and complex geometry boundary.
Acknowledgements

The present work is supported by the National Key Research and Development Program of China (Grant No. 2016YFC0401600), the Primary Research & Development Plan of Jiangsu Province, China (Grant No. BE2016056) and the Natural Science Foundation of Jiangsu Province, China (Grant No. SBK2014043338).

References

[1] Abboud, J.E. and Oweis, G.F. (2013) Themicrojetting Behavior from Single Laser-Induced Bubbles Generated above a Solid Boundary with a Through Hole. *Experiments in Fluids*, **54**, 1438. https://doi.org/10.1007/s00348-012-1438-6

[2] Hashmi, A., Yu, G., Reilly-Collette, M., Heiman, G. and Xu, J. (2012) Oscillating Bubbles: A Versatile Tool for Lab on a Chip Applications. *Lab on a Chip*, **12**, 4216-4227. https://doi.org/10.1039/c2lc40424a

[3] Naudé, C.F. and Ellis, A.T. (1961) On the Mechanism of Cavitation Damage by Nonhemispherical Cavities Collapsing in Contact with a Solid Boundary. *Journal of Basic Engineering*, **83**, 648-656. https://doi.org/10.1115/1.3662286

[4] Li, B.B., Jia, W., Zhang, H.C. and Lu, J. (2014) Investigation on the Collapse Behavior of a Cavitation Bubble near a Conical Rigid Boundary. *Shock Waves*, **24**, 317-324. https://doi.org/10.1007/s00193-013-0482-3

[5] Aganin, A.A., Ilgamov, M.A., Kosolapova, L.A. and Malakhov, V.G. (2016) Dynamics of a Cavitation Bubble near a Solid Wall. *Thermophysics and Aeromechanics*, **23**, 211-220. https://doi.org/10.1134/S0869864316020074

[6] Magaletti, F., Gallo, M., Marino, L. and Casciola, C.M. (2016) Shock-Induced Collapse of a Vapor Nanobubble near Solid Boundaries. *International Journal of Multiphase Flow*, **84**, 34-45. https://doi.org/10.1016/j.ijmultiphaseflow.2016.02.012

[7] Hsu, C.Y., Liang, C.C., Nguyen, A.T. and Teng, T.L. (2014) A Numerical Study on the Underwater Explosion Bubble Pulsation and the Collapse Process. *Ocean Engineering*, **81**, 29-38. https://doi.org/10.1016/j.oceaneng.2014.02.018

[8] Zhang, A.M. and Liu, Y.L. (2015) Improved Three-Dimensional Bubble Dynamics Model Based on Boundary Element Method. *Journal of Computational Physics*, **294**, 208-223. https://doi.org/10.1016/j.jcp.2015.03.049

[9] Chen, S. and Doolen, G.D. (1998) Lattice Boltzmann Method for Fluid Flows. *Annual Review of Fluid Mechanics*, **30**, 329-364. https://doi.org/10.1146/annurev.fluid.30.1.329

[10] Succi, S. (2001) The Lattice Boltzmann Equation: For Fluid Dynamics and Beyond. Oxford University Press, Oxford.

[11] Gunstensen, A.K., Rothman, D.H., Zaleski, S. and Zanetti, G. (1991) Lattice Boltzmann Model of Immiscible Fluids. *Physical Review A*, **43**, 4320. https://doi.org/10.1103/physreva.43.4320

[12] Shan, X. and Chen, H. (1993) Lattice Boltzmann Model for Simulating Flows with Multiple Phases and Components. *Physical Review E*, **47**, 1815. https://doi.org/10.1103/physreve.47.1815

[13] Shan, X. and Chen, H. (1994) Simulation of Non Ideal Gases and Liquid-Gas Phase Transitions by the Lattice Boltzmann Equation. *Physical Review E*, **49**, 2941. https://doi.org/10.1103/physreve.49.2941

[14] Swift, M.R., Osborn, W.R. and Yeomans, J.M. (1995) Lattice Boltzmann Simulation of Non-Ideal Fluids. *Physical Review Letters*, **75**, 830. https://doi.org/10.1103/physrevlett.75.830
[15] Chen, L., Kang, Q., Mu, Y., He, Y.L. and Tao, W.Q. (2014) A Critical Review of the Pseudo Potential Multiphase Lattice Boltzmann Model: Methods and Applications. *International Journal of Heat and Mass Transfer*, 76, 210-236.

[16] Huang, H., Sukop, M. and Lu, X. (2015) Multiphase Lattice Boltzmann Methods: Theory and Application. John Wiley & Sons Ltd., Chichester. [https://doi.org/10.1002/9781118971451](https://doi.org/10.1002/9781118971451)

[17] Sukop, M.C. and Or, D. (2005) Lattice Boltzmann Method for Homogeneous and Heterogeneous Cavitation. *Physical Review E*, 71, Article ID: 046703. [https://doi.org/10.1103/physreve.71.046703](https://doi.org/10.1103/physreve.71.046703)

[18] Chen, X.P., Zhong, C.W. and Yuan, X.L. (2011) Lattice Boltzmann Simulation of Cavitating Bubble Growth with Large Density Ratio. *Computers & Mathematics with Applications*, 61, 3577-3584.

[19] Mishra, S.K., Deymier, P.A., Muralidharan, K., Frantziskonis, G., Pannala, S. and Simunovic, S. (2010) Modeling the Coupling of Reaction Kinetics and Hydrodynamics in a Collapsing Cavity. *Ultrasonics Sonochemistry*, 17, 258-265.

[20] Shan, M.L., Zhu, C.P., Xi, Z.H.O.U., Cheng, Y.I.N. and Han, Q.B. (2016) Investigation of Cavitation Bubble Collapse near Rigid Boundary by Lattice Boltzmann Method. *Journal of Hydrodynamics*, 28, 442-450.

[21] Shan, M.L., Zhu, C.P., Yao, C., Yin, C. and Jiang, X.Y. (2016) Pseudopotential Multi-Relaxation-Time Lattice Boltzmann Model for Cavitation Bubble Collapse with High Density Ratio. *Chinese Physics B*, 25, Article ID: 104701. [https://doi.org/10.1088/1674-1056/25/10/104701](https://doi.org/10.1088/1674-1056/25/10/104701)

[22] Li, Q. and Luo, K.H. (2014) Thermodynamic Consistency of the Pseudopotential Lattice Boltzmann Model for Simulating Liquid-Vapor Flows. *Applied Thermal Engineering*, 72, 56-61.

[23] Kupershtokh, A.L., Medvedev, D.A. and Karpov, D.I. (2009) On Equations of State in a Lattice Boltzmann Method. *Computers & Mathematics with Applications*, 58, 965-974.

[24] Li, Q., Luo, K.H. and Li, X.J. (2012) Forcing Scheme in Pseudopotential Lattice Boltzmann Model for Multiphase Flows. *Physical Review E*, 86, Article ID: 016709. [https://doi.org/10.1103/physreve.86.016709](https://doi.org/10.1103/physreve.86.016709)

[25] Li, Q., Luo, K.H. and Li, X.J. (2013) Lattice Boltzmann Modeling of Multiphase Flows at Large Density Ratio with an Improved Pseudopotential Model. *Physical Review E*, 87, Article ID: 053301. [https://doi.org/10.1103/physreve.87.053301](https://doi.org/10.1103/physreve.87.053301)

[26] Li, Q., Luo, K.H., Kang, Q.J., He, Y.I.L., Chen, Q. and Liu, Q. (2016) Lattice Boltzmann Methods for Multiphase Flow and Phase-Change Heat Transfer. *Progress in Energy and Combustion Science*, 52, 62-105.

[27] Yuan, P. and Schaefer, L. (2006) Equations of State in a Lattice Boltzmann Model. *Physics of Fluids*, 18, Article ID: 042101. [https://doi.org/10.1063/1.2187070](https://doi.org/10.1063/1.2187070)

[28] Shan, X. (2008) Pressure Tensor Calculation in a Class of Nonideal Gas Lattice Boltzmann Models. *Physical Review E*, 77, Article ID: 066702. [https://doi.org/10.1103/physreve.77.066702](https://doi.org/10.1103/physreve.77.066702)

[29] Shi, D.Y., Wang, Z.K. and Zhang, A.M. (2014) Study on Coupling Characteristics between Bubble and Complex Walls at the Same Scale. *Acta Physica Sinica*, 63, 533-538.

[30] Plesset, M.S. and Chapman, R.B. (1971) Collapse of an Initially Spherical Vapour Cavity in the Neighbourhood of a Solid Boundary. *Journal of Fluid Mechanics*, 47, 283-290. [https://doi.org/10.1017/S0022112071001058](https://doi.org/10.1017/S0022112071001058)

[31] Guo, Z.-L., Zheng, C.-G. and Shi, B.-C. (2002) Non-Equilibrium Extrapolation Me-
Method for Velocity and Pressure Boundary Conditions in the Lattice Boltzmann Method. *Chinese Physics*, **11**, 366. [https://doi.org/10.1088/1009-1963/11/4/310](https://doi.org/10.1088/1009-1963/11/4/310)

[32] Zou, Q. and He, X. (1997) On Pressure and Velocity Boundary Conditions for the Lattice Boltzmann BGK Model. *Physics of Fluids*, **9**, 1591-1598. [https://doi.org/10.1063/1.869307](https://doi.org/10.1063/1.869307)

[33] Huang, H., Krafczyk, M. and Lu, X. (2011) Forcing Term in Single-Phase and Shan-Chen-Type Multiphase Lattice Boltzmann Models. *Physical Review E*, **84**, Article ID: 046710. [https://doi.org/10.1103/physreve.84.046710](https://doi.org/10.1103/physreve.84.046710)

Submit or recommend next manuscript to SCIRP and we will provide best service for you:

Accepting pre-submission inquiries through Email, Facebook, LinkedIn, Twitter, etc.
A wide selection of journals (inclusive of 9 subjects, more than 200 journals)
Providing 24-hour high-quality service
User-friendly online submission system
Fair and swift peer-review system
Efficient typesetting and proofreading procedure
Display of the result of downloads and visits, as well as the number of cited articles
Maximum dissemination of your research work

Submit your manuscript at: [http://papersubmission.scirp.org/](http://papersubmission.scirp.org/)
Or contact jamp@scirp.org
Non-Native English Speaker Readability Metric: Reading Speed and Comprehension

Aissa Boudjella1,2, Mukti Sharma3, Deepti Sharma4

1Department of Information Technology and Computer Science, American University of Afghanistan, Kabul, Afghanistan
2Department of Electrical Engineering, Bircham International University, Madrid, Spain
3General Education KIC, Abu Dhabi, United Arab Emirates
4BML Munjal University, Gurgaon, India

Abstract

This paper presents an investigation to evaluate the reading speed and reading comprehension of non-native English speaking students by presenting a simple analytical model. For this purpose, various readability softwares were used to estimate the average grade level of the given texts. The relationship between the score obtained by the students and their reading speed under average grade level 9 and 14 using font size 12 and 14 is presented. The experimental results show that the reading speed and the score versus the students may be explained by a linear regression. Reading speed decreases as the score decreases. The students with a higher magnitude of reading speed scored better marks. More importantly, we find that the reading speed of our students is lower than the native English speakers. This approach of modeling the readability in linear form significantly simplifies the readability analysis.

Keywords

Readability, Reading Speed, Software Readability, Flesch-Kincaid Grade Level, Gunning-Fog Score, Coleman-Liau Index, Smog Index, Automatic Readability Index, Linear Regression

1. Introduction

Language is an essential tool to convert our ideas into words. Language is a form of communication which has two important parts: 1) The expressions or experiences which we want to express or share and 2) the suitable words that we use to convey these ideas or experiences. English as a Lingua Franca is an existing common language used for communication between speakers of different languages. When we learn English language we aim to develop all four major skills
of the language: Listening, Speaking, Reading and Writing.

All four skills are either receptive or transmittal skills.

Reading is an active and diligent process. To establish a successful and productive communication between a reader and a writer, the prerequisite is the clarity of thoughts and correct usage of the words in the text. The relevance of writing is unattainable unless the reader understands and comprehends the text in its actual context. Reading involves the understanding of words, and their meanings to have an accurate interpretation of the text. When reader reads a text he or she should be able to connect the words with the given situation for better understanding.

Readability is defined as the ease with which a written text can be understood by a reader. The readability of a particular text depends on both: The complexity of its vocabulary and syntax.

To measure readability, various computer-based formulas have been proposed that include only two factors [1]-[13]: 1) the number of syllables or (letters) in a word and 2) the number of words in a sentence. The results generated by these various formulas are not accurate and not always in a good agreement.

In the present paper, we will examine the reading speed \( V_{rs} \) and reading comprehension text shown as a black text color on a plain background for non-native English speaking undergraduate students of our institution. Different presentations of text with different font sizes were assigned to the students. The obtained results are compared with those of similar investigations. Correlation between performance goals such as success rate, time to complete tasks versus students will be presented.

2. Methodology, Experimental Result & Discussion

In order to check the readability a planned survey was conducted in two phases. In the first phase, a drafted text was chosen with a few direct questions to evaluate the students’ reading speed. The difficulty level of the text was checked thoroughly using readability software [1]-[11] before it was given to the students. The students were asked to read the text carefully for a few minutes and write the answer in the space provided after each paragraph. The time they took to read the text was noted by the instructor. The purpose of this investigation was just to see whether the students had understood the text by considering the time it took to complete the tasks. In the second phase, after monitoring the reading speed, a reading comprehension was given to the students. An elaborated and structured text having variety of questions as multiple choice questions was given to the students. The purpose of this survey was to check the understanding of the students. The total time taken by the students to complete the text was also noted by the instructor.

The sample subjects of this study are 70 students both male and female, 60% male and 40% female. In term of language proficiency, we used a placement test with average level 9 and 14, calculated using the readability software indicated in Table 1 [1]-[11]. The selected subjects were employed and non-employed students.
Table 1. Grade level and readability score [1]-[11].

| Readability formula | Text for comprehension | Text for reading speed |
|---------------------|------------------------|------------------------|
| Flesch-Kincaid grade level | 12.6 | 8.7 |
| Gunning-Fog Score | 16.2 | 11 |
| Fry-Graph | 14.1 | 8.9 |
| Coleman-Liau index | 16.8 | 8.3 |
| Smog index | 11.7 | 8.7 |
| Automatic readability index | 13.2 | 9.1 |

Average grade level 14.1 8.7

| Character count | 1.621 | 2.168 |
| syllable | 534 | 711 |
| Word count | 293 | 517 |
| Sentence count | 17 | 25 |
| Characters per word | 5.5 | 4.2 |
| Syllabus per word | 1.8 | 1.4 |
| Words per sentence | 17.2 | 20.7 |

Readability Formula

| Flesch-Kincaid grade level | 35.2 | 68.4 |

with different age ranges between 20 and 35. Two different reading tests were used to evaluate the students’ performance. Speed of reading test \( (V_r) \) and comprehension test with different presentation of font sizes text assigned to the subjects divided into two groups.

For the speed of reading test, the average grade level and score of the text are 8.7 and 68.4, respectively. The details about the grade level texts are indicated in Table 1 and can be calculated using Fry graph software Figure 1. Scores calculated by readability software are different. The scores obtained by the students usually range between 0 and 100. A higher score indicates easier readability. Score readability in the range between [90 - 100] is too easily understood by an average 11 old year student, while score in the range between [60 - 70] can be easily understood by 13 - 15 old year student. Score in the range between [0 - 30] is best understood by university student. Based on the USA education system, a grade level is equivalent to the number of years of education a person has had. Scores over 22 should generally be taken to mean graduate level text.

The reading speed text contained 516 words. The subjects were allowed to do the test after they confirmed that they had understood about how to answer the questions. They were also told to be a little fast because of the time constraint.

Presenting text in darker colors and larger size with more pixel aids helped students in performing better. Black text on a plain background has been found to yield faster reading than black text on a medium textured background [14], up to 32 percent faster than reading light text on a dark background [14].
seems that larger text with more pixel draws more attention than smaller ones. In this investigation, for rapid reading and understanding we present black text on white page with high contrast backgrounds. The font size is either 12 or 14 points.

Long sentences using unnecessary words are often used to express more than one idea in a sentence. Research indicates that brief and simple sentences are easily and readily understood than long sentences. Sentences over 20 words in length cause a loss in reading comprehension [15]. In addition to adequate contrast between text and its background, it is also recommended that the number of sentences in a paragraph should not exceed six as indicated in Table 1 which is provided in this investigation.

The comprehension reading text contained 295 words. The test was supposed to be consistent with the information items, to be tested with intermediate learners. The average grade level and score of reading comprehension text are 14.1 and 35.2, respectively. For the reading speed text, the average grade level and score are 9 and 68.4, respectively. The grade level and score are indicated in Table 1 and can be calculated using Raygor graph software Figure 2. The score is the number of items the subject completes accurately within the time limit indicated in Figures 3-8.

Data are represented using a plot called a scatter plot or $x - y$ scatter diagram plot. During analysis we try to find the equation of a line that fits the data. This is called the regression line. Points are $(x = \text{student}, y = \text{reading speed or score})$ pairs can be plotted on the Cartesian coordinate system. From the study of correlation when the slope of the regression line is positive the value of $y$ increases.
Figure 2. Reading speed text. \( \mathbf{X} = \) Average number words with 6+ characters per 100 words: 22. \( \mathbf{Y} = \) Average number of sentences per 100 words: 5.1. \( \mathbf{G} = 7.9 = \) Grade level (the number between the pink parallel lines is the grade level).

Figure 3. Speed and score versus student (Group 1): Font size 12.

as the value of \( x \) increases. This is called a positive correlation. When the slope of the regression line is negative the value of \( y \) decreases as \( x \) increases. The strength of these relationships is given by the correlation coefficient \( R \) which can be calculated. Regression analysis is used to predict the value of the variable based on the value of a second variable which is controlled by the experimenter. Results may be plotted on a scatter plot. Correlation is used to give information about the relationship between \( x \) and \( y \). When the regression equation is calculated, the correlation results indicate the nature and the strength of the relationship. The correlation coefficient, \( R \), indicates this relationship between \( x \) and \( y \).
Figure 4. Speed and score versus student Group 2: Font size 12.

Figure 5. Score versus speed (Group 1). Font size 12.

Figure 6. Score versus speed (Group 2). Font size 12.
Values of $R$ range from $-1$ to $+1$. A correlation coefficient of $0$ means that there is no relationship. A value of $-1$ is a perfect negative coefficient and a correlation value of $+1$ indicates a perfect positive correlation. Another value of use in correlation analysis is the coefficient of determination which is represented as $R^2$, and varies between 0 and 1.

In Figure 3 and Figure 4, after collecting the data, the sample for $V_{rs}$ and students’ score are sorted in ascending order. Then, the sorted data of $V_{rs}$ and score (y-axis) versus the corresponding students (x-axis) are displayed as a scatter plot. The x-axis coordinates $x_i$ corresponds to the first student’s name point where $x_n$ corresponds to the $n^{th}$ student’s name point.

The variation in $V_{rs}$ using font size 12 associated with the student scores versus student names are illustrated in Figure 3 and Figure 4, for two student groups, group1 and group 2, respectively. The experimental results show that the
students with a higher magnitude of $V_{rs}$ scored better marks. On the other hand, lower mark is observed when the $V_{rs}$ decreases by approximately from 120 to 70 [words/min]. More importantly, each group shows a linear relationship between $V_{rs}$ or score versus the reader. The average slopes obtained for the group 1 and 2 are given in Equation (1.1), (1.2) and (2.1), (2.2), respectively with a slight difference ($a_{rs1} = -0.015$, $a_{score1} = -3.6759$) and ($a_{rs2} = -0.088$, $a_{score2} = -1.87$). The test for the reading speed under the average level 9, shows approximately the same extracted parameters for group 1 ($a_{rs1} = -0.015$, $b_{rs1} = 1.0034$) and group 2 ($a_{rs2} = -0.0088$, $b_{rs2} = 0.8427$), respectively. Based on these extracted results, it seems that these groups perform quite similarly. On other hand, with higher average grade level 14, for the text comprehension, the extracted parameters for group 1 ($a_{score1} = -3.6759$, $b_{score1} = 104.82$) and for group 2 ($a_{score2} = -1.87$, $b_{score2} = 78.911$) show that the readability performance is quite different. Note that the x-axis is the name of the student and not the number of the student, and the graph was obtained after sorting the data from the smallest to the highest score.

Linear regression for reading speed and score obtained by the students for group 1 (Figure 3) are given in Equations (1.1) and (1.2), respectively.

\[
y = -0.015x + 1.0034 \quad \text{(readings peed)} \quad \text{(1.1)}
\]
\[
y = -3.6759x + 104.82 \quad \text{(Score)} \quad \text{(1.2)}
\]

Linear regression for reading speed and score versus the students for group 2 (Figure 4), are given in Equations (2.1) and (2.2), respectively.

\[
y = -1.87x + 0.78911 \quad \text{(readings peed)} \quad \text{(2.1)}
\]
\[
y = -0.088x + 0.8427 \quad \text{(Score)} \quad \text{(2.2)}
\]

$R$-squared is a statistical measure of how close the data are to the fitted regression line. The reading speed data indicates a very strong positive correlation ($R^2 > 0.77$) between $V_{rs}$ and student. While, the score graph versus the student shows very weak correlation in the range of $R^2 = [0.1055 - 0.1958]$.

Figure 5 and Figure 6 show the score obtained by the student versus $V_{rs}$ using font size text 12. When $V_{rs}$ changes approximately from 30 to 130 [words/min], the score increases linearly by approximately from 70% to 120%. The relationship between $V_{rs}$ and its corresponding score may be described by linear regression given by the Equations (3.1) and (3.2), respectively. With an average slope and intercept ($a_{ss1} = 0.0038$, $b_{ss1} = 0.5931$) and ($a_{ss2} = 0.0039$, $b_{ss2} = 0.59097$) for the group 1 and 2, respectively. The two samples appear radically the same.

Linear regression score versus reading speed for the group 1 (Figure 5) and group 2 (Figure 6) are given in Equations (3.1) and (3.2), respectively.

The value of the coefficient of determination $R^2$ is less than 0.2. This indicates very weak correlation.

\[
y = 0.0380x + 0.5931 \quad \text{(3.1)}
\]
\[
y = 0.039x + 0.5097 \quad \text{(3.2)}
\]
For reading comprehension under different size font 12 and 14, the relationship between the students and their scores is illustrated in Figure 7 and Figure 8 for the group 1 and 2, respectively. This correlation can be described by linear relationship indicated by the equations 4.1 and 4.2 for group 1 and 2, respectively. Note that the x-axis is the name of the student and not the number of the student, and the graph was obtained after sorting the data from the lowest to the highest score. In this investigation, the subjects were told that there was no time limitation and they could do this part at their own pace, but they were not supposed to take a lot of time. The value of the coefficient of determination under this condition is larger than 0.84, which indicates very strong correlation.

\[ y = 0.0544x + 1.1492 \]  
\[ y = -0.01x + 1.0023 \]

For the same group, the score obtained under font 12 and 14 can also be described by linear relationship. The average slopes for group 1 \( (a = -0.0544, \text{ Figure 7}) \) and for group 2 \( (a = -0.01, \text{ Figure 8}) \) are different. The minimum value scored for group 1, is about 30%, while for the group 2 the minimum value is 87%.

The survey indicated that more than 90% of the students do not read. In comparison with the native speaker, it seems that non English speaking students do not spend more time in reading. This could be explained by many factors such as, difficulty in reading, lack of motivation, environment, etc.

The coefficient of determination gives an indication of the contribution of the factor being studied in the regression analysis to the relationship between reading speed, score and student. In the case of reading speed data (Figure 3 and Figure 4) and score data (Figure 7, Figure 8), the value of the coefficient of determination, \( R^2 \) is larger than 0.77. This indicates a very strong correlation between the student and their performance. On other hand, the coefficient of determination, the score obtained by the student under controlled time shows very weak correlation (Figure 3 and Figure 4) with \( R^2 \) magnitude less than 0.2. This indicates that that the time is an important factor changing the magnitude of the coefficient of determination \( R^2 \) from 0.2 (Figure 3, Figure 4 score data) to more than 0.84 (Figure 7, Figure 8).

It is extremely important for a reader to get enough motivation to read and then using his skills to comprehend and understand the text. Motivation can be instilled through consistent encouragement by the educator and parents. But interest can only be developed through regular reading. The complexity of the text affects the reader’s motivation. In order to develop interest, it is essential that text should not be very complex to understand. If the text has difficult vocabulary and complex syntax then the reader loses interest in the very first phase of reading.

Reading skill, once developed, can be most easily maintained at a high level by the students. If the students have poor receptive skills then they will find reading...
monotonous. As a result they will never interpret the text in the right context. They will face multiple problems to comprehend and understand English especially in their academic courses where English Language is most commonly used. Due to lack of student’s proficiency in the language, to bridge the language gap, it requires a lot of collaborative efforts and coordination from the student and the teacher to overcome the difficulties in the process of learning. Higher complexity requires more learning and results in less efficient human performance.

In addition to the effect of the environment, the difference of reading speed and performance may be explained by individual differences and basic skills. We all differ in learning abilities and in task completion [16].

Knowledge, experience, environment, and familiarity will help to remember and increase the readability speed. For example, it seems that most native English-speaking people remember English words easily than non-native speaker. Reading in one’s native language is easier than reading something in a foreign language. A non-native speaker who uses English or learns English as a second language faces many difficulties in the process of learning English. Reading doesn’t come easily. To enjoy reading and achieve proficiency, interest should be created in reading independently without facing any difficulties. Reading should be developed as a habit so that the reader enjoys what he reads.

Learning is the process of encoding in long-term memory information that is contained in short-term memory. It is a complex process that requires consistent efforts. Learning process is improved through repetition and deep analysis, only if the information being transferred from short-term memory, has structure and is meaningful and familiar. Based on above learning processes, it can be ascertained that high readability requires high skill. In case of sample students due to lack of adequate reading, their process of encoding is likely to be reduced which results in less efficient performance indicated by the slow reading speed and low score due to the low degree of familiarity and deep analysis.

The essence of skill is in the performance of actions characterized by consistency and economy of effort. Given enough time, people can improve their performance in almost any task. Usability goals versus performance in the form of measurable objectives must also be established. Performance goals such as the time it takes to complete tasks versus success rates must be defined. In performance, research indicates that a greater working memory is positively related to increase reading comprehension, reasoning skill, and learning technical information [17]. In addition, information stored within working memory is variously thought to last from 5 to 30 seconds, with estimates of working memory storage capacity size is about of 3 to 4 items [14]. Based on the above theories, it seems that low reading speed can be explained by the minimum information stored time with low capacity storage.

Lind et al., reported that there are two levels of information processing [18]. Both levels function simultaneously. The highest level used for reading and understanding, which consists of consciousness and working memory, performing
reasoning and problem solving. This level is limited, slow, and sequential. In contrast to the lowest level, perceiving the physical form of information sensed, it processes familiar information rapidly. It seems that the process of both levels of our students is likely to be reduced.

3. Conclusion

Through our study we made an attempt to investigate the degree of readability of non-native English speakers. The study demonstrates that speed of reading, which is an indicator of assessing a non-native English speaker’s readability, is lower than that of a native English speaker. More importantly, the results show that the relationship between the reading speed and score versus the reader can be described by linear regression with very strong correlation. The magnitude of extracted parameters namely the slope and y-intercept maybe used as guideline to assess and evaluate the readability. In our future work, to clarify the weak correlation obtained in the present investigation, experiments will be carried under separated gender, controlled time and different grade level.

Authors’ Contribution

The co-authors (Dr. Muktisharma and Dr. Deepti Sharma) prepared the drafted text that was given to the students to carry out the survey with sharing contribution writing introduction, methodology and conclusion by all the authors. While the experimental results analysis, modelling, discussion, and full citations of references in the press manuscript are the contributions of the main author (Dr. Aissa Boudjella).

Acknowledgements

The first author would like to acknowledge the Wikipedia foundation to make available the free and open source software used in this investigation.

References

[1] Flesch, R. (1948) A New Readability Yardstick. *Journal of Applied Psychology, 32*, 221-233. [https://doi.org/10.1037/h0057532](https://doi.org/10.1037/h0057532)

[2] Klare, G.R. (1963) The Measurement of Readability. University of Iowa Press, Ames, IA.

[3] Chall, J.S. (1958) Readability: An Appraisal of Research and Application. Bureau of Educational Research, Ohio State University, Columbus, OH.

[4] Farr, J.N., Jenkins, J.J. and Paterson, D.G. (1951) Simplification of the Flesch Reading Ease Formula. *Journal of Applied Psychology, 35*, 333-357. [https://doi.org/10.1037/h0062427](https://doi.org/10.1037/h0062427)

[5] Dale, E. and Chall, J.S. (1948) A Formula for Predicting Readability. *Educational Research Bulletin, 27*, 1-20, 37-54.

[6] Gunning, R. (1952) The Technique of Clear Writing. McGraw-Hill, New York.

[7] Coleman, M. and Liau, T.L. (1975) A Computer Readability Formula Designed for Machine Scoring. *Journal of Applied Psychology, 60*, 283-284. [https://doi.org/10.1037/h0076540](https://doi.org/10.1037/h0076540)
[8] Fry, E.B. (1963) Teaching Faster Reading. Cambridge University Press, London.

[9] Fry, E.B. (1968) A Readability Formula That Saves Time. Journal of Reading, 11, 513-516.

[10] McLaughlin, G.H. (1969) SMOG Grading: A New Readability Formula. Journal of Reading, 22, 639-664.

[11] Senter, R.J. and Smith, E.A. (1967) Automated Readability Index. Wright-Patterson Air Force Base. p. 3. AMRL-TR-6620.

[12] DuBay, W.H. (2000) Smart Language: Readers, Readability, and the Grading of Text. Impact Information, Costa Mesa.

[13] Caylor, J.S., Stitch, T.G., Fox, L.C. and Ford, J.P. (1973) Methodologies for Determining Reading Requirements of Military Occupational Specialties: Technical Report No. 73-5. Human Resources Research Organization, Alexander.

[14] Koyani, S.J., Bailey, R.W. and Nall, J.R. (2004) Research-Based Web Design and Usability Guidelines. National Cancer Institute.

[15] Galitz, W.O. (2002) The Essential Guide to User Interface Design an Introduction to GUI Design Principles and Techniques. John Wiley & Sons, Inc., Hoboken.

[16] Nielsen, J. (2006) Ten Best Intranets of 2006. www.useit.com/Alertbox

[17] Baddeley, A. (1992) Working Memory. Science, 255, 556-559.

[18] Lind, N.E., Johnson, M. and Sandblad, B. (1992) The Art of the Obvious. CHI 95. Monterey, 3-7 May 1992.

Submit or recommend next manuscript to SCIRP and we will provide best service for you:

Accepting pre-submission inquiries through Email, Facebook, LinkedIn, Twitter, etc.
A wide selection of journals (inclusive of 9 subjects, more than 200 journals)
Providing 24-hour high-quality service
User-friendly online submission system
Fair and swift peer-review system
Efficient typesetting and proofreading procedure
Display of the result of downloads and visits, as well as the number of cited articles
Maximum dissemination of your research work

Submit your manuscript at: http://papersubmission.scirp.org/
Or contact jamp@scirp.org
Optimal Distributed Control Problem for the $b$-Equation

Chunyu Shen

Nonlinear Scientific Research Center, Faculty of Science, Jiangsu University, Zhenjiang, China
Email: shenchunyus8228@sina.com

Abstract

This paper is concerned with the optimal distributed control problem governed by $b$-equation. We firstly investigate the existence and uniqueness of weak solution for the controlled system with appropriate initial value and boundary condition. By contrasting with our previous result, the proof without considering viscous coefficient is a big improvement. Secondly, based on the well-posedness result, we find a unique optimal control for the controlled system with the quadratic cost functional. Moreover, by means of the optimal control theory, we obtain the sufficient and necessary optimality condition of an optimal control, which is another major novelty of this paper. Finally, we also present the optimality conditions corresponding to two physical meaningful distributive observation cases.

Keywords

Weak Solution, Existence and Uniqueness, Optimal Control, Sufficient and Necessary Optimality Condition, $b$-Equation

1. Introduction

Recently, Escher and Yin [1] studied the following nonlinear dispersive equation ($b$-equation):

$$
\begin{cases}
    u_t - \alpha^2 u_{xx} + c_0 u_x + (b+1) uu_x + \Gamma u_{xxx} = \alpha^2 \left( bu_x u_x + uu_{xx} \right), & t > 0, x \in \mathbb{R}, \\
    u(0,x) = u_0(x), & x \in \mathbb{R},
\end{cases}
$$

(1.1)

where $c_0$, $b$, $\Gamma$ and $\alpha$ are arbitrary real constants. Denoting $y = u - \alpha^2 u_x$, we can rewrite $b$-equation in the following form:

$$
\begin{cases}
    y_t + c_0 y_x + bu_x y_x + bu_x y + \Gamma u_{xxx} = 0, & t > 0, x \in \mathbb{R}, \\
    u(0,x) = u_0(x), & x \in \mathbb{R}.
\end{cases}
$$

(1.2)

Equation (1.2) can be derived as a family of asymptotically equivalent shallow
water wave equations that emerge at quadratic order accuracy for \( \forall b \neq -1 \) by an appropriate Kodama transformation [2] [3]. For the case \( b = -1 \), the corresponding Kodama transformation is singular and the asymptotic ordering is violated [2] [3]. The solutions of the \( b \)-Equation (1.2) with \( c_0 = \Gamma = 0 \) were studied numerically for various values of \( b \) in [4] [5], where \( b \) was taken as a bifurcation parameter. The symmetry condition necessary for integrability of the \( b \)-Equation (1.2) was investigated in [6]. The Korteweg-de Vries (KdV) equation, the Camassa-Holm (CH) equation and the Degasperis-Procesi (DP) equation are the only three integrable equations in the \( b \)-Equation (1.2), which was shown in [7] [8] by using Painleve analysis. The \( b \)-equation with \( c_0 = \Gamma = 0 \) admits peaked solutions for \( \forall b \in R \) [4] [5] [7]. The peaked solutions replicate a feature that is characteristic for the waves of great height: waves of the largest amplitude that are exact solutions of the governing equations for water waves [9] [10] [11].

If \( \alpha = 0 \) and \( b = 2 \), then \( b \)-Equation (1.1) becomes the well-known KdV equation

\[
u_t + c_0 \nu_x + 3 \nu \nu_x + \Gamma \nu_{xxx} = 0, \ t > 0, \ x \in R,
\]

which describes the unidirectional propagation of waves at the free surface of shallow water under the influence of gravity [12]. In this model, \( u(t, x) \) represents the wave’s height above a flat bottom; \( x \) is proportional to distance in the direction of propagation and \( t \) is proportional to the elapsed time. The KdV equation is completely integrable, and its solitary waves are solitons [13]. The Cauchy problem of the KdV equation has been studied by many authors [14] [15] [16] and a satisfactory local or global (in time) existence theory is now available (for example, in [15] [16]). The solution of the KdV equation is global for \( u_0 \in L^2(S) \) [15] [16]. It is also observed that the KdV equation does not accommodate wave breaking (by wave breaking we mean the phenomenon that a wave remains bounded but its slope becomes unbounded in finite time) [17].

For \( \Gamma = 0 \) and \( b = 2 \), \( b \)-Equation (1.1) becomes the CH equation

\[
u_t - \alpha^2 \nu_{xxx} + c_0 \nu_x + 3 \nu \nu_x = 2\alpha^2 \nu u_x + \alpha^2 \nu u_{xx}, \ t > 0, \ x \in R,
\]

modelling the unidirectional propagation of shallow water waves over a flat bottom. Again \( u(t, x) \) stands for the fluid velocity at time \( t \) in the spatial \( x \) direction and \( c_0 \) is a nonnegative parameter related to the critical shallow water speed [18]. The CH equation is derived physically by approximating directly the Hamiltonian for Euler’s equations in the shallow water regime (it also appears in the context of hereditary symmetries studied by Fuchssteiner and Fokas [19]). Recently, the alternative derivations of the CH equation as a model for water waves, respectively, as the equation for geodesic flow on the diffeomorphism group of the circle were presented in [20] and in [21]. For the physical derivation, we refer to the work in [22]. The geometric interpretation is important because it can be used to prove that the least action principle holds for the CH equation [23]. It is worth pointing out that the fundamental aspect of the CH equation, the fact that it is a completely integrable system, was shown in [24] [25].
for the periodic case and in [26] [27] for the non-periodic case. Its solitary waves are smooth if \( c_0 > 0 \) and peaked in the limiting case \( c_0 = 0 \) [28]. They are orbitally stable and interact like solitons [29] [30] and the explicit interaction of the peaked solitons is given in [14].

Since the CH equation is structurally very rich, many physicists and mathematicians pay great attention to it. Local well-posedness for the initial datum \( u_0 \in H^s (I) \) with \( s > 3/2 \) was proved by several authors, as in [31] [32] [33] [34]. For the initial data with lower regularity, we refer to Molinet’s paper [35] and also the paper [36]. Moreover, wave breaking for a large class of initial data has been established in [31] [33] [37] [38]. However, in [39], global existence of weak solutions was proved but uniqueness was obtained only under a prior assumption that is known to hold only for the initial data \( u_0(x) \in H^1 \) such that \( u_0 - u_{0,x,x} \) is a sign-definite Radon measure (under this condition, global existence and uniqueness was shown in [40]). Also it is worth noting that CH equation has global conservative solutions in \( H^1(R) \) [36] [41] [42] and global dissipative solutions (with energy being lost when wave breaking occurs) in \( H^1(R) \) [43] [44]. In [45], the authors showed the infinite propagation speed for the CH equation in the sense that a strong solution of the Cauchy problem with compact initial profile cannot be compactly supported at any later time unless it is the zero solution, which is an improvement of the previous results in this direction obtained in [46].

For \( c_0 = \Gamma = 0 \) and \( b = 3 \) in b-Equation (1.1), then we find the DP equation of the form [8]

\[
u_t - \alpha^2 u_{xx} + 4uu_x = 3\alpha^2 u_x u_{xx} + \alpha^2 u_{xxx}, \quad t > 0, x \in R.
\]

Degasperis, Holm and Hone [47] proved the formal integrability of the DP Equation (1.5) by constructing a Lax pair. They also showed that DP equation has a bi-Hamiltonian structure and an infinite sequence of conserved quantities, and that it admits exact peakon solutions which are analogous to the CH peakons. Peakons for either \( b = 2 \) or \( b = 3 \) are true solitons that interact via elastic collisions under CH dynamics, or DP dynamics, respectively. Recently, Lundmark [48] showed that the DP equation has not only peaked solitons, but also shock peakons of the form

\[ u(t,x) = -\frac{1}{t+k} \text{sgn}(x) e^{1/t}, \quad k > 0. \]

The DP equation can be regarded as a model for nonlinear shallow water dynamics and its asymptotic accuracy is the same as for the CH shallow water equation [2] [3] [22]. An inverse scattering approach for computing n-peakon solutions to the DP equation was presented in [49]. Its traveling wave solutions were investigated in [50].

The Cauchy problem for the DP equation has been studied widely. Local well-posedness of this equation is established in [51] [52] for the initial data \( u_0 \in H^s (S) \) with \( s > 3/2 \). Similar to the CH equation, the DP equation has also global strong solutions [51] [53] [54] [55] as well as finite time blow-up solu-
tions [51] [53] [54] [56] [57]. On the other hand, it has global weak solutions in $H^1(S)$ [53] [56] [58]. Analogous to the case of the CH equation, Henry [59] and Mustafa [60] showed that smooth solutions to the DP equation have infinite speed of propagation. Coclite and Karlsen [61] also obtained global existence results for entropy weak solutions belonging to the class of $L^1(R) \cap BV(R)$ and the class of $L^1(R) \cap L^\infty(R)$.

Although the DP equation is similar to the CH equation in several aspects, these two equations are truly different. One of the novel features of the DP equation different from the CH equation is that it has not only peakon solutions [47] and periodic peakon solutions [58], but also shock peakons [48] [61] and the periodic shock waves [56].

Despite the abundant literature on the above three special cases of the $b$-equation, there are few results on the $b$-equation. Recently, some authors devoted to studying the Cauchy problem of the $b$-equation. Since the conservation laws of the $b$-equation are much weaker, there are only a few kinds of global or blow-up results.

In [1], Escher and Yin studied $b$-equation on the line for $\alpha > 0$ and $c_b, b, \Gamma \in \mathbb{R}$. They established the local well-posedness, described the precise blow-up scenario, and proved that the equation has strong solutions which exist globally in time and blow up in finite time. Moreover, the authors showed the uniqueness and existence of global weak solutions to $b$-equation, provided the initial data satisfy certain sign conditions. The similar discussions for $b$-equation on the circle can be found in [62]. The author expanded the result of corresponding solutions blow-up in finite time where conditions on the initial data and the bifurcation parameter $b \geq 3$ in [2] to the case $b \geq 2$ [63]. In [64], the authors established the local well-posedness for the nonuniform weakly dissipative $b$-equation which includes both the weakly dissipative CH equation and the weakly dissipative DP equation as its special cases. They studied the blow-up phenomena and the long time behavior of the solutions.

Recently Gui, Liu, and Tian [65] considered $b$-Equation (1.1) with $c_0 = \Gamma = 0$ on the real line. They proved that the equation is locally well-posed in the Sobolev space $H^s(R)$ for $s > 3/2$. Moreover, they give the precise blow-up scenario of strong solution of the equation with certain initial data. In [66], Zhou established blow-up results for $b$-equation with $c_0 = \Gamma = 0, \alpha = 1$ under various classes of initial data. He also proved that the solutions with compact support initial data do not have compact support. In the periodic case of $b$-equation with $c_0 = \Gamma = 0$, sufficient conditions on the initial data were obtained in [67] to guarantee the finite time blow-up and global existence. The local well-posedness of $b$-equation with $c_0 = \Gamma = 0, \alpha = 1$ in the critical Besov space $B^{3/2}_{2,1}$ was studied in [68]. They showed that if a weaker $B^{q}_{r,r}$-topology is used, the solution map becomes Hölder continuous. Moreover, they showed that the dependence on initial data is optimal in $B^{3/2}_{2,1}$ in the sense that the solution map is continuous but not uniformly continuous. They also obtained the periodic peaked solutions and applied them to obtain the ill-posedness in $B^{3/2}_{2,1}\sigma$. There are some
other papers concerned with \( b \)-equation of \( c_s = \Gamma = 0 \) and we will not attempt to mention all here.

In the past decades, the optimal control of distributed parameter systems has become much more active in academic field. Especially, the optimal control of nonlinear solitary wave equation lies in the front of the intersection of mathematics, engineering and computer science and so on. Recently, people have taken a considerable interest in realizing the operation mechanism of prototype tsunami in the laboratory and in looking for a really efficient control mechanism to generate exact long water waves in the man-made pool. The CH equation attracted much more attention also in the context of the relevance of integrable equations to the modelling of tsunami waves [69] [70] [71]. Naturally, an optimization problem needs to be considered in this shallow water wave equation. It seems to the author that the study of nonlinear shallow water equation from the point of view of control theory was an open field. There are only some research results reported. For instance, Zhang studied the control problems for two nonlinear dispersive wave equations--the KdV equation and the Benjamin-Bona-Mahony (BBM) equation. Moreover, for the BBM equation, he showed that the wave-maker, by choosing a proper boundary value, can make a wave to approach a given state as closely as desired as long as the given state is small in some sense [72]. Glass investigated the problem of exact controllability and asymptotic stabilization of the CH equation on the circle, by means of a distributed control. The results are global, and in particular the control prevents the solution from blowing up [73]. The distributed optimal control problems for the viscous CH equation, the viscous DP equation, the viscous Dullin-Gottwald-Holm (DGH) equation were considered by our research team respectively. We proved the existence and uniqueness of weak solution in short interval. Further, we employed the quadratic cost objective functional to be minimized within an admissible control set with the distributive observation and discussed the existence of optimal control which minimizes the quadratic cost functional [74] [75] [76]. Subsequently, by the Dubovitskii-Milyutin functional analytical approach, Sun considered the optimal distributed control problem of the viscous generalized CH equation and viscous DGH equation respectively and obtained the Pontryagin maximum principle of the systems studied. The necessary optimality condition is established for an optimal control problem in fixed final horizon case [77] [78]. In [79] [80], recently, our research team studied optimal distributed control of the Fornberg-Whitham equation and the \( \theta \)-equation which involve complex nonlinear items respectively. We clarified the well-posedness of weak solution without relying on viscous coefficient, which is major improvement in comparison with our previous results. Utilizing the Dubovitskii-Milyutin functional analytical approach, we also proved the necessary optimality condition for the control systems in fixed final horizon case. Hwang studied the quadratic cost optimal control problems for the viscous DGH equation. He derived the necessary optimality conditions of optimal controls, corresponding to physically meaningful distributive observations. By making use of the second
order Gateaux differentiability of solution mapping on control variables, he also proved the local uniqueness of optimal control [81].

Inspired by the papers mentioned above, in present work, we investigate the $b$-equation from the point of view of distributed control. More precisely, we consider the following governing equation

$$
\begin{align*}
\frac{\partial u}{\partial t} - u_{xx} + c \frac{\partial u}{\partial x} + (b + 1) \mu u_x + \Gamma u_{xxx} - bu_x + uu_{xxx} &= Bv, \\
u(t, x + L) &= u(t, x), \forall x \in \mathbb{R}, \forall t \in [0, T], \\
y(0, x) &= y_0(x) = u(0, x) - u(0, x) \in V,
\end{align*}
$$

(1.6)

where $Bv$ is the external control term which is $L$-periodic in spatial $x$, $v \in \mathcal{U}_{ad}$ is a control and $B$ be an operator called a controller. The explicit formulation of the control problem will be provided after the investigation of well-posedness of the state equation.

We mainly consider the two following problems:

- for the nonlinear control system governed by the $b$-equation with quadratic cost functional $I(v) = \| Cu(v, t, x) - z \|_m^2 + (Nv, v)_t$, can one find $v^* \in \mathcal{U}_{ad}$ such that $I(v^*) = \inf_{v \in \mathcal{U}_{ad}} I(v)$ and whether this $v^*$ is unique?

- if one finds the unique optimal control $v^* \in \mathcal{U}_{ad}$ for the above control problem, how can we characterize this optimal control?

The plan of the remaining sections can be summarized as follows. In Section 2, we study the initial-boundary problem of the $b$-equation with forcing function in a special space $S(0, T)$. Adopting the Faedo-Galerkin method and utilizing a uniformly prior estimate of the approximate solution, we prove the existence and uniqueness of weak solution under the definition introduced in the paper. For general $b \in \mathbb{R}$, the proof without relying on viscous coefficient is a major improvement in comparison with our results in [74] [75] [76] and other discussions in [77] [78] [81]. In Section 3, based on the well-posedness result, we give the formulation of the quadratic cost optimal control problem for the $b$-equation and investigate the existence and uniqueness of the optimal solution. In Section 4, by the method of control theory (for more detailed discussion, we refer readers to book [82]), we establish the sufficient and necessary optimality condition of an optimal control in fixed final horizon case. In order to obtain this result, we also prove the Gateaux differentiability of the state variable $u(y; t, x)$ which is used to define the associate adjoint systems. Comparing with the research in our previous works [74] [75] [76] and the related works [77] [78] [79] [80] [81], the sufficient and necessary optimality condition of an optimal control which is not limited to the necessary condition is another novelty in this paper. At last, in Section 5, we give the specific sufficient and necessary optimality condition of optimal control $v^*$ for two physical meaningful distributed observation cases employing the associate adjoint systems.

2. The Existence and Uniqueness of Weak Solution

Without loss of generality, we assume $\Omega = [0, L]$. Denote the usual Hilbert
space $H = L^2(\Omega)$ equipped with the norm $\|u\|_H = \left(\int_\Omega |u|^2 \, dx\right)^{\frac{1}{2}}$, and the inner product in $H$ is denoted by $(u, v)_H = \|u\|_H^2$. Let $H^s(\Omega) = W^{s,2}(\Omega)$, $s \in \mathbb{N}$ be the integral exponent Sobolev spaces. By using the Poincare's inequality in $H^s(\Omega)$, we can define norm $\|\xi\|_{H^s(\Omega)} = \left(\sum_{0 \leq j \leq s} \|\partial_x^j \xi\|_H^2\right)^{\frac{1}{2}}$, where $\partial_x^j \xi(0) = \partial_x^j \xi(L)$ and $s = 0, 1, 2, 3, \ldots$. Especially, taking $m = 1$, we get the Hilbert space $V = H^1(\Omega)$ supplied with the inner product $(\varphi, \psi)_V = (\varphi, \psi)_H$, where $\forall \varphi, \psi \in V$. Let us denote that $V' = H^{-1}(\Omega)$ and $H^s = L^2(\Omega)$ are the dual spaces of $V$ and $H$ respectively. Then we can find that $V$ embeds into $H$ and $H'$ embeds into $V'$, where each embedding is dense and corresponding injections are continuous.

For convenience, we shall consider the following initial-boundary value problem for Equation (1.1)

\begin{equation}
\begin{aligned}
u_t - \nu_{xx} + c \nu_x + \left(\frac{b+1}{2}\right) \nu u_x + \Gamma \nu_{xx} - b \nu u_x - u u_{xx} &= f(t, x), \\
u(x, x + L) &= u(t, x), \forall x \in R, \forall t \in [0, T], \\
u(0, x) &= y_0(x), \forall x \in R,
\end{aligned}
\end{equation}

where $f(t, x)$ is forcing item which is $L$-periodic in spatial $x$.

With $y(t, x) = u(t, x) - u_{xx}(t, x)$ and $y_0(x) = u(0, x) - u_{xx}(0, x)$, Equation (2.1) takes the form:

\begin{equation}
\begin{aligned}
u_t - \nu_{xx} + c \nu_x + \left(\frac{b+1}{2}\right) \nu u_x + \Gamma \nu_{xx} (t, x) &= f(t, x), \\
u(t, x + L) &= u(t, x), \forall x \in R, \forall t \in [0, T], \\
u(0, x) &= y_0(x), \forall x \in R.
\end{aligned}
\end{equation}

In order to study the weak solution of Equation (2.2), we introduce the following two special spaces firstly.

$\mathcal{W}(0, T)$ is defined by \( \mathcal{W}(0, T) = \left\{ \xi \big| \xi \in L^2(0, T; V), \xi_t \in L^2(0, T; V') \right\} \), which is equipped with the norm \( \|\xi\|_{\mathcal{W}(0, T)} = \left(\|\xi\|^2_{L^2(0, T; V)} + \|\xi_t\|^2_{L^2(0, T; V')}\right)^{\frac{1}{2}} \).

$S(0, T)$ is defined by \( S(0, T) = \left\{ \xi \big| \xi \in L^2(0, T; H^1(\Omega)), \xi_t \in L^2(0, T; V) \right\} \), endowed with the norm \( \|\xi\|_{S(0, T)} = \left(\|\xi\|^2_{L^2(0, T; H^1(\Omega))} + \|\xi_t\|^2_{L^2(0, T; V)}\right)^{\frac{1}{2}} \).

It is easy to verify that the spaces $\mathcal{W}(0, T)$ and $S(0, T)$ are both Hilbert spaces.

**Definition 2.1.** A function $u(t, x) \in S(0, T)$ is said to be a weak solution of Equation (2.2), if $y(t, x) = u(t, x) - u_{xx}(t, x) \in \mathcal{W}(0, T)$ satisfies

\begin{equation}
\begin{aligned}
\left( y_t(t, \cdot), \varphi(\cdot) \right)_H + \left( c \nu_x(t, \cdot), \varphi(\cdot) \right)_H = \left( u(t, \cdot) y_x(t, \cdot), \varphi(\cdot) \right)_H + \left( b \nu u_x(t, \cdot), \varphi(\cdot) \right)_H + \left( \Gamma \nu_{xx}(t, \cdot), \varphi(\cdot) \right)_H = \left( f(t, \cdot), \varphi(\cdot) \right)_H, \\
u(t, x + L) &= u(t, x), \forall x \in R, \forall t \in [0, T], \\
u(0, x) &= y_0(0, x) \in V,
\end{aligned}
\end{equation}
for \( \forall \varphi(\cdot) \in H \) in the sense of \( D'(0,T) \).

From now on, when we speak of a solution of Equation (2.2), we shall always mean the weak solution in the sense of Definition 2.1 unless noted otherwise.

We set an unbounded linear self-adjoint operator \( Au = -u'_x \), where \( \forall u \in D(A) = H \cap \{u | u(t,x+L) = u(t,x)\} \). Then the set of all linearly independent eigenvectors \( \{\omega_j\}_{j \in N^+} \) of \( A \) with the eigenvalues \( \{\lambda_j^\prime\}_{j \in N^+} \), i.e., \( A\omega_j = \lambda_j^\prime \omega_j \),

\[
0 < \lambda_1^\prime \leq \lambda_2^\prime \leq \cdots \leq \lambda_j^\prime \to \infty \quad \text{as} \quad j \to \infty,
\]
is an orthonormal basis of \( H \).

Furthermore, we can define the powers \( A^s \) of \( A \) for \( s \in N^+ \), where the space \( D(A^s) \) is a Hilbert space which is endowed with the norm \( \|A^s\|_H \). It can be found that the following expression holds

\[
A^s \omega_j = (-1)^s \hat{c}_{\omega_j} \omega_j = \lambda_j^s \omega_j,
\]

where \( \{\omega_j\}_{j \in N^+} \) are eigenvectors of \( A^s \) and \( \{\lambda_j^s\}_{j \in N^+} \) are eigenvalues.

**Definition 2.2.** A function \( u_m(t,x) = \sum_{j=1}^m a_{jm}(t) \omega_j(x) \in C^1([0,T];S_m) \) is called an approximate solution to Equation (2.2), if it satisfies

\[
\begin{align*}
&\left(y_m(t,x),\omega_j\right)_H + \left(c_{0} y_m,x, u_m\right)_H + \left(u_m(t,x) y_m,x, \omega_j\right)_H \\
&+ \left(b u_m,x \omega_j\right)_H + \left(\Gamma u_m,x, \omega_j\right)_H = \left(f(t,x),\omega_j\right)_H \\
u_m(t,x+L) = u_m(t,x), \forall x \in R, \forall t \in [0,T],
\end{align*}
\]

where \( y_m(t,x) = u_m(t,x) - u_{m,\xi}(t,x) \), \( S_m = \text{span}\{\omega_1(x), \omega_2(x), \cdots, \omega_m(x)\} \)

and

\[
a_{jm}(t) \in C^1([0,T];R).
\]

**Lemma 2.1.** Let \( y(t,x) = u(t,x) - u_{\xi}(t,x) \in \mathcal{W}(0,T) \) and \( u(t,x) \) satisfies the boundary conditions of Equation (2.1). Then, we get

\[
\|u(t,x)\|_{\mathcal{W}(0,T)} \leq C \|y(t,x)\|_{\mathcal{W}(0,T)},
\]

where \( C > 0 \) is a constant.

The proof of Lemma 2.1 can be referred to our article [79] [80].

**Theorem 2.1.** Assume that \( f(t,x) \in L^2(0,T;V) \) and \( y_0(x) \in V \). Then, Equation (2.2) exhibits a unique weak solution \( u(t,x) \in \mathcal{S}(0,T) \).

**Proof:** Multiplying both sides of the first equation in Equation (2.4) by \( a_{jm}(t) \) and summing up over \( j \) from 1 to \( m \), we have

\[
\begin{align*}
\left(y_m,t, y_m\right)_H + \left(c_{0} y_m,x, u_m\right)_H + \left(u_m(t,x) y_m,x, u_m\right)_H + \left(b u_m,x, y_m\right)_H + \left(\Gamma u_m,x, u_m\right)_H = \left(f, u_m\right)_H
\end{align*}
\]

This gives

\[
\frac{d}{dt}\left(\|u_m\|_V^2 + \|u_m\|_H^2\right) = (2-b) \int_{\Omega} u'_m dx + 2 \left(f, u_m\right)_H.
\]

(2.5)
Because \( f(t,x) \in L^2(0,T;V) \) is a forcing function, we can assume that \( \|f\|_V \leq M_I \), where \( M_I > 0 \) is constant.

It then derives from Equation (2.5) that
\[
\frac{d}{dt} \left( \|u_m\|^2 + \|u_m\|^2_{H^1(\Omega)} \right) \leq [2 - b_1] \lambda_2 \|u_m\|_{L^2(\Omega)}^2 + \lambda_2^2 M_I^2 + \|u_m\|^2_{H^1(\Omega)},
\] (2.6)
where \( \lambda_i > 0, i = 1, 2 \) are embedding constants. In order to estimate the term \( \|u_m\|^2_{H^1(\Omega)} + \|u_m\|^2_{L^2(\Omega)} \), we should estimate the term \( \{u_m\}_{n \in N^+} \) in \( H^2(\Omega) \).

Multiplying both sides of the first equation in Equation (2.4) by \( \lambda^*_i a_{m,t}(t) \) and summing up over \( j \) from 1 to \( m \), we get
\[
\left( y_{m,t} - u_{m,xx} \right)_H + \left( c_{ij} u_{m,x} - u_{m,xx} \right)_H + \left( u_{m,xx} - u_{m,xx} \right)_H + \left( \Gamma u_{m,xxx} - u_{m,xxx} \right)_H = \left( f - u_{m,xx} \right)_H.
\]

The above equation implies that
\[
\frac{d}{dt} \left( \|u_m\|^2 + \|u_m\|^2_{H^1(\Omega)} \right) + (b + 1) \int_{\Omega} u_{m,xx}^2 \, dx + (2b - 1) \int_{\Omega} u_{m,xx}^2 u_{m,xx} \, dx
\]
\[= 2 \left( f - u_{m,xx} \right)_H, \]
(2.7)

By the use of the Sobolev embedding theorem, we can estimate the following items as
\[-(b + 1) \int_{\Omega} u_{m,xx}^2 \, dx \leq \|f\|_{L^1(\Omega)} \|u_m\|_{H^1(\Omega)}^2 \leq \|f\|_{L^1(\Omega)} \|u_m\|_{H^1(\Omega)}^2;\]
\[-(2b - 1) \int_{\Omega} u_{m,xx}^2 u_{m,xx} \, dx \leq \|f\|_{L^1(\Omega)} \|u_m\|_{H^1(\Omega)}^2 \leq \|f\|_{L^1(\Omega)} \|u_m\|_{H^1(\Omega)}^2 ;\]
and
\[2 \left( f - u_{m,xx} \right)_H \leq 2 \|f\|_{L^1(\Omega)} \|u_m\|_{H^1(\Omega)} \leq 2 \lambda_2 M_I \|u_m\|_{H^1(\Omega)} ,\]

where \( \lambda_i > 0, i = 1, 2 \) are embedding constants.

Therefore, we can deduce from Equation (2.7) that
\[
\frac{d}{dt} \left( \|u_m\|^2 + \|u_m\|^2_{H^1(\Omega)} \right)
\leq \|f\|_{L^1(\Omega)} \|u_m\|_{H^1(\Omega)}^2 + \|u_m\|_{H^1(\Omega)}^2 + \|u_m\|_{H^1(\Omega)}^2 + 2 \lambda_2 M_I \|u_m\|_{H^1(\Omega)}^2
\leq \beta_i \lambda_2 \left( \|u_m\|^2 + \|u_m\|^2_{H^1(\Omega)} + \frac{2 \lambda_2 M_I}{\beta_i \lambda_2} \right),
\]
where
\[
\beta_i = \max \{b + 1,|2b - 1|\} . \tag{2.8}
\]

From inequality (2.8), we can obtain that
\[
\|u_m\|^2 + \|u_m\|^2_{H^1(\Omega)} \leq \left[ 1 - \frac{\beta_i \lambda_2}{2} \left( \|u_m\|^2 + \|u_m\|^2_{H^1(\Omega)} + \frac{2 \lambda_2 M_I}{\beta_i \lambda_2} \right)^2 \right]^{1/2} + \frac{2 \lambda_2 M_I}{\beta_i \lambda_2} \]
\[= \frac{2 \lambda_2 M_I}{\beta_i \lambda_2}, \tag{2.9}
\]
where $\forall t \in [0, T]$, 
\[
T < \frac{2}{\beta_z^2 \lambda_z^2 \left( \left\| u_m (0, x) \right\|_{H^2}^2 + \left\| u_m (0, x) \right\|_{H^2}^2 \right) + 2 \beta_z \lambda_z M_1}
\]
and $M_2 > 0$ is a constant.

Therefore, combining the boundedness of the sequence $\{u_m\}_{m \in \mathbb{N}^+}$ in $H^2(\Omega)$ with the inequality (2.6), we can derive that
\[
\left\| u_m \right\|_{H^2}^2 + \left\| u_m \right\|_{L^2}^2 \leq \left( \left\| u_m (0, x) \right\|_{H^2}^2 + \left\| u_m (0, x) \right\|_{H^2}^2 + \lambda_z^2 M_2^2 \right) \exp (\beta_z t)
\]
\[
- \lambda_z^2 M_2^2 \triangleq M_3^2,
\]
where $\forall t \in [0, T]$, $\beta_z = \max \{2 - b | \lambda_z M_2, 1\}$ and $M_3$ is some positive constant.

Similarly, multiplying both sides of the first equation in Equation (2.4) by $(\lambda_j^2) a_{j, m}(t)$ and summing up over $j$ from 1 to $m$, we can get
\[
\left( y_{m,t}, u_{m,xxxx} \right)_H + \left( (c_g u_m, u_{m,xxxx})_H + \left( u_m, y_{m,xxxx} \right)_H \right) + \left( bu_{m,t}, y_{m,xxxx} \right)_H + (\Gamma u_{m,xxxx}, u_{m,xxxx})_H = \left( f, u_{m,xxxx} \right)_H.
\]

By integration by parts in the above equation, we can deduce that
\[
\frac{d}{dt} \left( \left\| u_m \right\|_{H^2(\Omega)}^2 + \left\| u_m \right\|_{L^2(\Omega)}^2 \right) + 5(b + 1) \int_\Omega u_{m,t} u_{m,xx}^2 \, dx + (2b + 1) \int_\Omega u_{m,t} u_{m,xx}^2 \, dx = 2 \left( f, u_{m,xxxx} \right)_H.
\]

Using the Sobolev embedding theorem, inequality (2.9) and boundary conditions of Equation (2.4), we can estimate the following each item
\[
-5(b + 1) \int_\Omega u_{m,t} u_{m,xx}^2 \, dx \leq 5 \left\{ b + 1 \right\} \left\| u_m \right\|_{L^\infty} \left\| u_m \right\|_{H^2(\Omega)}^2
\]
\[
\leq 5 \left\{ b + 1 \right\} \lambda_z \left\| u_m \right\|_{H^2(\Omega)}^2
\]
\[
\leq 5 \left\{ b + 1 \right\} \lambda_z M_3^2,
\]
\[
- (2b + 1) \int_\Omega u_{m,t} u_{m,xxxx}^2 \, dx \leq 2(b + 1) \left\| u_m \right\|_{L^\infty} \left\| u_m \right\|_{H^2(\Omega)}
\]
\[
\leq 2(b + 1) \lambda_z M_2 \left\| u_m \right\|_{H^2(\Omega)}
\]
and
\[
2 \left( f, u_{m,xxxx} \right)_H \leq 2 \left( f_m, u_{m,xxxx} \right)_H \leq 2 \left\| f_m \right\|_{L^2(\Omega)} \left\| u_m \right\|_{H^2(\Omega)} \leq M_3^2 + \left\| u_m \right\|_{H^2(\Omega)}.
\]

Combining above estimates, Equation (2.11) can be deduced into the following inequality
\[
\frac{d}{dt} \left( \left\| u_m \right\|_{H^2(\Omega)}^2 + \left\| u_m \right\|_{L^2(\Omega)}^2 \right) \leq \left( [2b + 1] \lambda_z M_2 + 1 \right) \left\| u_m \right\|_{H^2(\Omega)} + \left( 5 \left[ b + 1 \right] \lambda_z M_3 + M_3^2 \right)
\]
\[
\leq \left( [2b + 1] \lambda_z M_2 + 1 \right) \left( \left\| u_m \right\|_{L^2(\Omega)} + \left\| u_m \right\|_{H^2(\Omega)} \right) + \left( 5 \left[ b + 1 \right] \lambda_z M_3 + M_3^2 \right).
\]

From inequality (2.12), we can obtain that
where $\forall t \in [0, T]$, $T < -\frac{2}{\sqrt{\beta^2 \lambda_2^2 \left( \|u_m(0,x)\|_{h_1}^2 + \|u_m(0,x)\|_{H^{(1)}}^2 \right) + 2\beta \lambda_2 \lambda_i M_i}}$ and $M_i > 0$ is a constant.

Hence, combining estimate inequality (2.9) and (2.13), we can find that
\[
\|y_m\|_{L^2(0,T;H)}^2 = \|y_m - u_{n,x}\|_{L^2}^2 = \|u_m\|_{L^2}^2 + 2\|u_m\|_{L^2}^2 + \|u_m\|_{H^{(1)}}^2 \leq M_2^2 + M_4^2,
\]
which indicate $y_m \in V$. We also can have $y_m \in H$ from the fact of $V$ embeds into $H$.

Combining estimate inequality (2.9) and (2.10), we also can know that
\[
\|y_m\|_{L^2(0,T;H)}^2 = \|y_m - u_{n,x}\|_{L^2}^2 = \|u_m\|_{L^2}^2 + 2\|u_m\|_{L^2}^2 + \|u_m\|_{H^{(1)}}^2 \leq M_2^2 + M_3^2.
\]

Therefore, we deduce from inequality (2.14) that
\[
\|y_m\|_{L^2(0,T;H)}^2 \leq \left( M_2^2 + M_4^2 \right) T,
\]
which indicates $\{y_m\}_{m \in \mathbb{N}^*}$ is uniformly bounded in $L^2(0,T;H^*)$.

Afterward, we will prove uniform boundedness of sequence $\{y_{m,t}\}_{m \in \mathbb{N}^*}$ in $L^2(0,T;V^*)$. Indeed, from the first equation of Equation (2.2) and the Sobolev embedding theorem, we have
\[
\|y_{m,t}\|_{L^2(0,T;V^*)} \leq \|y_m\|_{L^2} + \|c_0\| \|u_m\|_{L^2} + \lambda_2 \|u_m\|_{H^{(1)}} + \|f\| \|u_m\|_{H^{(1)}} \leq \lambda_2 M_1 + \|c_0\| M_3 + \lambda_2 M_3 \sqrt{M_2^2 + M_4^2}
\]
\[
+ \|f\| \lambda_i M_2 \sqrt{M_2^2 + M_4^2} + \|f\| M_2,
\]
where $\lambda_i > 0, i = 2, 3$ are embedding constants as before.

It derives from inequality (2.17) that
\[
\|y_{m,t}\|_{L^2(0,T;V^*)} \leq \left[ \lambda_2 M_1 + \|c_0\| M_3 + \lambda_2 M_3 \sqrt{M_2^2 + M_4^2} + \|f\| \lambda_2 M_3 \sqrt{M_2^2 + M_4^2} + \|f\| M_2 \right]^2 T.
\]

Collecting the analysis above, one has:

1. For $\forall t \in [0, T]$, where $T < \frac{2}{\sqrt{\beta^2 \lambda_2^2 \left( \|u_m(0,x)\|_{h_1}^2 + \|u_m(0,x)\|_{H^{(1)}}^2 \right) + 2\beta \lambda_2 \lambda_i M_i}}$, the sequence $\{y_{m,t}\}_{m \in \mathbb{N}^*}$ is bounded in $L^2(0,T;H)$ as well as in $L^2(0,T;V)$,
which is independent of the dimension of ansatz space \( S_m \).

(II) For \( \forall t \in [0, T] \), where

\[
T < \frac{2}{\sqrt{\beta \lambda^2 + \left( \| u_m(0, x) \|_2^2 + \| u_m(0, x) \|_2^2 \right) + 2\beta \lambda_2 M}},
\]

the sequence \( \{ y_{m,t} \}_{m \in N^+} \) is bounded in \( L^2(0, T; V') \), which is also independent of the dimension of ansatz space \( S_m \).

So, we obtain the boundedness of \( \{ y_{m,t} \}_{m \in N^+} \) in \( W(0, T) \) from (I) and (II) mentioned above. By the extraction theorem of Rellich’s, there may extract a subsequence \( \{ y_{m,t} \}_{m \in N^+} \) and find a \( y \in W(0, T) \) such that

\[
y_{m,t} \xrightarrow{weakly} y \quad \text{in} \quad W(0, T), \quad \text{as} \quad k \to \infty. \tag{2.18}
\]

Utilizing the fact that \( V \) embeds \( H \) compactly and (2.18), we can refer to the conclusion of Aubin-Lions-Teman’s compact embedding theorem to verify that \( \{ y_{m,t} \} \) is pre-compact in \( L^2(0, T; H) \). Hence we can choose a subsequence (denoted again by \( \{ y_{m,t} \} \)) of \( \{ y_{m,t} \} \) such that

\[
y_{m,t} \xrightarrow{strongly} y \quad \text{in} \quad L^2(0, T; H), \quad \text{as} \quad k \to \infty. \tag{2.19}
\]

Because \( W(0, T) \) embeds into \( C(0, T; H) \), we can obtain that \( u_m \in C(0, T; H^2(0,1)) \). Then, by virtue of (2.19), we can find a subsequence (denoted again by \( \{ u_{m,t} \} \)) of \( \{ u_{m,t} \} \) such that

\[
u_{m,t} \xrightarrow{strongly} u \quad \text{in} \quad H^2(\Omega), \quad \text{as} \quad k \to \infty, \quad \text{for} \quad \forall t \in [0, T] \quad \text{a.e.}. \tag{2.20}
\]

Combining (2.18)-(2.20) and the Lebesgue dominated convergence theorem, we have

\[
u_{m,t} y_{m,t} \xrightarrow{weakly} u y \quad \text{in} \quad L^2(0, T; H), \quad \text{as} \quad k \to \infty; \tag{2.21}
\]

\[
u_{m,t} y_{m,t} \xrightarrow{strongly} u y \quad \text{in} \quad L^2(0, T; H), \quad \text{as} \quad k \to \infty; \tag{2.22}
\]

\[
u_{m,t} \xrightarrow{weakly} u \quad \text{in} \quad L^2(0, T; H), \quad \text{as} \quad k \to \infty. \tag{2.23}
\]

We replace \( y_{m,t} \) and \( u_{m,t} \) by \( y_{m,t} \) and \( u_{m,t} \) respectively in the first equation of Equation (2.4), which yields

\[
\left( y_{m,t}, \omega \right)_H + \left( c_0 u_{m,t}, \omega \right)_H + \left( u_{m,t} y_{m,t}, \omega \right)_H
+ \left( b u_{m,t} y_{m,t}, \omega \right)_H + \left( \Gamma u_{m,t}, \omega \right)_H = \left( f, \omega \right)_H. \tag{2.24}
\]

Multiplying both sides of Equation (2.24) by \( \alpha(t) \), where \( \alpha(t) \in C^1 [0, T] \), \( \alpha(T) = 0 \) and integrating the result equation over \( [0, T] \), we have

\[
\int_0^T \left[ -\left( y_{m,t}, \alpha \omega \right)_H + \left( c_0 u_{m,t}, \alpha \omega \right)_H + \left( u_{m,t} y_{m,t}, \alpha \omega \right)_H
+ \left( b u_{m,t} y_{m,t}, \alpha \omega \right)_H + \left( \Gamma u_{m,t}, \alpha \omega \right)_H \right] dt
= \int_0^T \left( f, \alpha \omega \right)_H dt + \left( y_{m,t}(0, x), \alpha(0) \omega \right)_H. \tag{2.25}
\]
Utilizing (2.19), (2.21)-(2.23), we may pass to the limit in Equation (2.25). Then, we get
\[
\int_0^T \left[ -\left(y, \alpha, \omega_j \right)_H + \left(c_0 u_x, \alpha \omega_j \right)_H + \left(u y_x, \alpha \omega_j \right)_H \\
+ \left(b u_x, \alpha \omega_j \right)_H + \left(\Gamma u_{xx}, \alpha \omega_j \right)_H \right] dt \\
= \int_0^T \left( f, \alpha \omega \right)_H dt + \left( y_0, \alpha(0) \omega_j \right)_H.
\] (2.26)

We can find Equation (2.26) is true for any \( \alpha(t) \). Therefore, we may take \( \alpha(t) \in \mathcal{D}(0, T) \), then Equation (2.26) gives
\[
\frac{d}{dt} \left(y(t, x), \omega \right)_H + \left(c_0 u_x(t, x), \omega \right)_H + \left(u(t, x), y, \omega \right)_H \\
+ \left(b u_x(t, x) y(t, x), \omega \right)_H + \left(\Gamma u_{xx}(t, x), \omega \right)_H = \left(f(t, x), \omega \right)_H
\]
in the sense of \( \mathcal{D}'(0, T) \).

Since \( j \) is arbitrary and finite linear combinations of \( \omega_j \) is dense in \( H \), we can find that \( y(t, x) \in \mathcal{W}(0, T) \) satisfies Definition 2.1. Hence, from complex analysis above and Lemma 2.1, we obtain the existence of weak solution \( u(t, x) \in \mathcal{S}(0, T) \) to Equation (2.2).

Next we will discuss the uniqueness of this weak solution.

Let \( u_1 \) and \( u_2 \) be any two weak solutions of Equation (2.1) and set \( \eta(t, x) = u_1(t, x) - u_2(t, x) \). Then \( \eta \) satisfies
\[
\begin{cases}
\eta_t - \eta_{xx} + c_0 \eta_x + (b + 1) u_1 \eta_x + (b + 1) u_2 \eta + \Gamma \eta_{xx} - b u_x \eta_{xx} - u_2 \eta_{xx} = 0, \\
\eta(t, x + L) = \eta(t, x), \forall x \in R, \forall t \in [0, T], \\
\eta(0, x) = \eta_0(0, x) = \eta_{\alpha}(0, x) = 0, \forall x \in R.
\end{cases}
\] (2.27)

Taking the inner product of both sides of the first equation in Equation (2.27) with \( \eta \), we obtain
\[
\frac{1}{2} \frac{d}{dt} \left( \|\eta\|_H^2 + \|\eta\|_V^2 \right) = - ((b + 1) u_1 \eta_x, \eta) - ((b + 1) u_2 \eta_x, \eta) + \left( b u_x \eta_{xx}, \eta \right) \\
+ \left( b u_x \eta_x, \eta \right) + \left( u_1 \eta_{xx}, \eta \right) + \left( u_2 \eta_{xx}, \eta \right). 
\] (2.28)

The right hand side of Equation (2.28) can be estimated as follows:
\[
- ((b + 1) u_1 \eta_x, \eta) \leq \|b + 1\| \|u_1\|_{L^\infty} \int_{\Omega} \|\eta_x\| \|\eta\| dx \leq \frac{|b + 1| \lambda_2 C_1}{2} \left( \|\eta\|_H^2 + \|\eta\|_V^2 \right)
\]
\[
- ((b + 1) u_2 \eta_x, \eta) \leq \|b + 1\| \|u_2\|_{L^\infty} \int_{\Omega} \|\eta_x\| \|\eta\| dx \leq \frac{|b + 1| \lambda_2 C_2}{2} \left( \|\eta\|_H^2 + \|\eta\|_V^2 \right)
\]
\[
(b u_x \eta_{xx}, \eta) = - b \int_{\Omega} \eta_{xx} \eta \|dx - b \int_{\Omega} \eta \| \eta \| dx
\]
\[
\leq \|b\| \|\eta_{xx}\|_L^\infty \int_{\Omega} \|\eta\| \|\eta\| dx + \|b\| \|\eta\|_{L^2(\Omega)} \int_{\Omega} \|\eta_{xx}\| dx
\]
\[
\leq \frac{|b| \lambda_2}{2} \|\eta\|_{L^2(\Omega)} \left( \|\eta\|_H^2 + \|\eta\|_V^2 \right) + \|b\| \|\eta\|_{L^2(\Omega)} \|\eta\|_V^2
\]
\[
\leq \frac{|b| \lambda_2}{2} C_3 \left( \|\eta\|_H^2 + \|\eta\|_V^2 \right) + \|b\| \lambda_2 C_4 \|\eta\|_V^2; 
\]
(bu₂,η,η) ≤ \frac{|b|}{2} \|u₂\|_{H^2} + \frac{3}{2} \int_Ω u₂ dx ≤ C₃ \left( \|y\|_{H^2} + \|\eta\|_{H^2} \right);

(u_i,\eta,\eta) = \int_Ω u_i \eta dx + \frac{3}{2} \int_Ω u_i \eta^2 dx \leq \frac{|λ_i|}{2} \|u_i\|_{H^2} + \frac{3}{2} \int_Ω u_i dx \leq \frac{λ_i}{2} \left( \|y\|_{H^2} + \|\eta\|_{H^2} \right) + \frac{3λ_i}{2} C₅ \|\eta\|_{H^2};

(u₂,η,η) = -2 \int_Ω u₂ dx ≤ \frac{λ_2}{2} \left( \|y\|_{H^2} + \|\eta\|_{H^2} \right) \leq \frac{λ_2}{2} \left( \|y\|_{H^2} + \|\eta\|_{H^2} \right),

where \( λ_2 > 0 \) is an embedding constant and \( C_i > 0, i = 1, 2, \ldots, 5 \) are some constants.

Combining all complex estimates above and Equation (2.28), we can deduce that

\[ \frac{d}{dt} \left( \|y\|_{H^2} + \|\eta\|_{H^2} \right) \leq \beta \left( \|y\|_{H^2} + \|\eta\|_{H^2} \right), \tag{2.29} \]

where

\[ \beta = \max \left\{ \frac{|b|}{2} \|y\|_{H^2} + \frac{3}{2} \|u₂\|_{H^2} + \frac{1}{2} \left( \|y\|_{H^2} + \|\eta\|_{H^2} \right) \right\} \leq \frac{λ_2}{2} \left( \|y\|_{H^2} + \|\eta\|_{H^2} \right) \leq λ_2 \left( \|y\|_{H^2} + \|\eta\|_{H^2} \right). \]

Integrating inequality (2.29) with respect to \( t \) over \( [0, T] \), we have

\[ \left( \|y(t, x)\|_{H^2} + \|\eta(t, x)\|_{H^2} \right) \leq \left( \|y(0, x)\|_{H^2} + \|\eta(0, x)\|_{H^2} \right) \exp(βt), \tag{2.30} \]

where \( ∀t ∈ [0, T] \). It follows from \( \eta(0, x) = 0 \) that \( \|y(t, x)\|_{H^2} + \|\eta(t, x)\|_{H^2} = 0 \), which implies \( u₂(x) = u₂(t, x) \).

This completes the proof of uniqueness.

### 3. The Existence and Uniqueness of an Optimal Control

In this section, we will give the formulation of the quadratic cost optimal control problem for \( b \)-equation and investigate the existence and uniqueness of an optimal solution.

Let \( \mathcal{U} \) be a Hilbert space of control variables, and \( B ∈ \mathcal{L}(\mathcal{U}, L^2(0, T; V)) \) be an operator called a controller. We assume that the admissible set \( \mathcal{U}_{ad} \) be a bounded closed convex set, which has the non-empty interior with respect to \( \mathcal{U} \) topology, i.e. \( \text{int}_{L^2(0,T)} \mathcal{U}_{ad} ≠ \emptyset \).

We study the following nonlinear control system:

\[ \begin{cases} y_s(v; t, x) + c_0 u_s(v; t, x) + u(v; t, x) y_s(v; t, x) \\ + bu_s(v; t, x) y(v; t, x) + Γu_{ss}(v; t, x) = Bv, \\ u(v; t, x) + L = u(v; t, x), ∀x ∈ R, \forall t ∈ [0, T], \\ y(v; 0, x) = y_0(x) ∈ V, \end{cases} \tag{3.1} \]

where \( v ∈ \mathcal{U}_{ad} \) is a control. By virtue of Theorem 2.1 and Equation (3.1), we
can uniquely define the solution mapping \( v \rightarrow u(v; t, x) \) of \( U_{ad} \) into \( S(0, T) \).

The weak solution \( u(v; t, x) \) is called the state variable of the nonlinear control system (3.1).

The observation of the state is assumed to be given by

\[
z(v; t, x) = Cu(v; t, x),
\]

where \( C \in \mathcal{L}(S(0, T), \mathcal{M}) \) is an operator called the observer and \( \mathcal{M} \) is a Hilbert space of the observation variables.

We shall consider the following quadratic cost functional associated with the nonlinear control system (3.1):

\[
I(v) = \|Cu(v; t, x) - z\|_\mathcal{M}^2 + (Nv, v)_{\mathcal{M}},
\]

where \( z \) \( \notin \mathcal{M} \) is a desired value of \( u(v; t, x) \). \( N \in \mathcal{L}(\mathcal{U}, \mathcal{U}) \) is symmetric and positive definite, i.e., \( (Nv, v)_{\mathcal{M}} = (v, Nv)_{\mathcal{M}} \geq \lambda \|v\|_{\mathcal{M}}^2 \), where \( \lambda > 0 \) is some constant.

Hence, the discussed optimal control problem is to find an element \( v^* \in U_{ad} \) such that

\[
I(v^*) = \inf_{v \in \mathcal{U}_{ad}} \left\{ I(v) : \forall v \in \mathcal{U}_{ad} \right\},
\]

which subject to the controlled system (3.1) together with the control constraints.

Now, we shall discuss the existence and uniqueness of an optimal control \( v^* \) for the cost functional (3.3), which is the content of the following theorem.

**Theorem 3.1.** Let us suppose that the hypotheses of Theorem 2.1 are satisfied. Then there exists a unique optimal control \( v^* \in U_{ad} \) for the nonlinear control system (3.1) with the cost functional (3.3), such that \( I(v^*) = \inf_{v \in U_{ad}} I(v) \).

**Proof.** Because \( U_{ad} \neq \emptyset \) is a closed convex set, there exists a minimizing sequence \( \{v_n\}_{n \in \mathbb{N}} \) in \( U_{ad} \) such that

\[
\inf_{v \in U_{ad}} I(v) = \lim_{n \to \infty} I(v_n).
\]

We set

\[
\pi(v_1, v_2) = \left( C(u(v_1; t, x) - u(0; t, x)), C(u(v_2; t, x) - u(0; t, x)) \right)_{\mathcal{M}} + (Nv_1, v_2)_{\mathcal{M}},
\]

and

\[
L(v) = \left( z_d - Cu(0; t, x), C(u(v; t, x) - u(0; t, x)) \right)_{\mathcal{M}}.
\]

Then cost functional (3.3) can be rewritten as

\[
I(v) = \pi(v, v) - 2L(v) + \|z_d - Cu(0; t, x)\|_{\mathcal{M}}^2,
\]

where \( \pi(v_1, v_2) \) is a continuous symmetric bilinear form on \( \mathcal{U} \) and \( L(v) \) is a continuous linear form on \( \mathcal{U} \).

Obviously, \( \{I(v_n)\} \) is bounded in \( \mathbb{R}^+ \). So, the quadratic cost functional (3.3) implies that there exists a constant \( M_0 > 0 \) such that

\[
\lambda \|v_n\|_{\mathcal{M}}^2 \leq (Nv_n, v_n)_{\mathcal{M}} \leq I(v_n) \leq M_0,
\]
which indicates that \( \{v_n\}_{n \in \mathbb{N}^*} \) is bounded in \( U \). Because \( U_{ad} \) is closed and convex set, we can extract a subsequence \( \{v_{n_k}\} \subseteq \{v_n\}_{n \in \mathbb{N}^*} \) and find a \( v^* \in U_{ad} \) such that

\[
v_{n_k} \rightharpoonup v^* \quad \text{in} \quad U, \quad \text{as} \quad k \to \infty.
\] (3.6)

From now on, each state variable \( u_a(t,x) = u(v_n; t,x) \in S(0,T) \) corresponding to \( v_n \) is the solution of

\[
\begin{align*}
y_{n,t} + c_d u_{n,x} + bu_{n,x}y_{n} + \Gamma u_{n,xx} + Bv &= 0, \\
u_n(t,x + L) &= \Gamma_n (t,x), \quad \forall x \in R, \forall t \in [0,T], \\
y_n(0,x) &= y_0(x),
\end{align*}
\] (3.7)

where \( y_n = u_n - u_{n,xx} \).

From inequality (3.5), the right hand side of the first equation in Equation (3.7) can be estimated as

\[
\|Bv_n\|_{L^2(0,T;F)} \leq \|B\|_{L^2(U,L^2(0,T;F))} \|v_n\|_{L^2} \leq \|B\|_{L^2(U,L^2(0,T;F))} \sqrt{\lambda^{-1} M_0} \leq M,
\] (3.8)

where \( M > 0 \) is some constant.

Utilizing inequality (3.8), we can apply the same method used in Theorem 2.1 to deduce that \( \{v_n\}_{n \in \mathbb{N}^*} \) is bounded in \( W(0,T) \). Hence, by the extraction theorem of Rellich’s, we can extract a subsequence \( \{v_{n_k}\} \) of \( \{v_n\}_{n \in \mathbb{N}^*} \) and find a \( y = u - u_{xx} \in W(0,T) \) such that

\[
y_{n_k} \rightharpoonup y \quad \text{in} \quad W(0,T), \quad \text{as} \quad k \to \infty.
\] (3.9)

Using the fact that \( V \) embeds \( H \) compactly and the result of (3.9), we can refer to the conclusion of Aubin-Lions-Teman’s compact embedding theorem to verify that \( \{y_{n_k}\} \) is pre-compact in \( L^2(0,T;H) \). So we can also choose a subsequence (denoted again by \( \{y_{n_k}\} \)) of \( \{y_n\}_{n \in \mathbb{N}^*} \) such that

\[
y_{n_k} \rightharpoonup y, \quad \text{in} \quad L^2(0,T;H) \quad \text{as} \quad k \to \infty.
\] (3.10)

On the other hand, because \( W(0,T) \) embeds into \( C(0,T;H^2(\Omega)) \), we can infer that \( u_n \in C(0,T;H^2(\Omega)) \). And from (3.10), we can get a subsequence (denoted again by \( \{u_{n_k}\} \)) of \( \{u_n\}_{n \in \mathbb{N}^*} \) such that

\[
u_{n_k} \rightharpoonup u, \quad \text{in} \quad H^2(\Omega), \quad \text{as} \quad k \to \infty, \quad \text{for} \quad t \in [0,T] \quad \text{a.e.}
\] (3.11)

Combining (3.9)-(3.11) and the Lebesgue dominated convergence theorem, it is not difficult to obtain that

\[
u_{n_k,t} \rightharpoonup u_t \quad \text{in} \quad L^2(0,T;H), \quad \text{as} \quad k \to \infty;
\] (3.12)

\[
u_{n_k} y_{n_k,t} \rightharpoonup u_y \quad \text{in} \quad L^2(0,T;H), \quad \text{as} \quad k \to \infty;
\] (3.13)

\[
u_{n_k} y_{n_k} \rightharpoonup u_y \quad \text{in} \quad L^2(0,T;H), \quad \text{as} \quad k \to \infty;
\] (3.14)

\[
u_{n_k,xxx} \rightharpoonup u_{xxx} \quad \text{in} \quad L^2(0,T;H), \quad \text{as} \quad k \to \infty.
\] (3.15)

We replace \( u_a \) and \( v_a \) by \( u_{n_k} \) and \( v_{n_k} \) in Equation (3.7) respectively, and
take \( k \to \infty \). Then, by the standard arguments as in [83], we find that the limit \( u \) satisfies the following equations:

\[
\begin{aligned}
\begin{cases}
y + c \partial u + uy_x + bu_x, y + \Gamma u_{x|x} = Bv^*, \\
u(t, x + L) = u(t, x), \forall x \in \mathbb{R}, \forall t \in [0, T], \\
y(0, x) = y_0(x).
\end{cases}
\end{aligned}
\]  

(3.16)

in weak sense, where \( y = u - u_{ss} \). Moreover, by the uniqueness of weak solution of Equation (3.16) via Theorem 2.1 and Lemma 2.1, we can conclude that \( u = u(v^*; t, x) \in \mathcal{S}(0, T) \), which implies \( u(v^*; t, x) \) weakly \( \to u(v^*; t, x) \) in \( \mathcal{S}(0, T) \).

Because the mapping \( v \to \pi(v, v) \) is lower semi-continuous in the weak topology of \( \mathcal{U} \) and \( \mathcal{M} \) is also lower semi-continuous. The mapping \( v \to L(v) \) is continuous in the weak topology of \( \mathcal{U} \). Thus the mapping \( v \to I(v) \) is weakly lower semi-continuous.

So, we can deduce from cost functional (3.4) that

\[
\lim_{k \to \infty} I(v_k) \geq I(v^*).
\]  

(3.17)

At the same time, from inequality (3.17), we have

\[
\inf_{v \in \mathcal{U}_{ad}} I(v) = \lim_{k \to \infty} I(v_k) \geq I(v^*).
\]

Moreover, combining \( I(v^*) \geq \inf_{v \in \mathcal{U}_{ad}} I(v) \) by definition, we can obtain that

\[
I(v^*) = \inf_{v \in \mathcal{U}_{ad}} I(v).
\]  

(3.18)

Next, we will prove the uniqueness of \( v^* \in \mathcal{U}_{ad} \) in (3.18).

Because the mapping \( v \to \pi(v, v) \) is strictly convex and the mapping \( v \to L(v) \) is continuous. Hence the mapping \( v \to I(v) \) is also strictly convex.

Let \( v_1^* \in \mathcal{U}_{ad} \) and \( v_2^* \in \mathcal{U}_{ad} \) be two optimal controls, which satisfy \( I(v_1^*) = \inf_{v \in \mathcal{U}_{ad}} I(v) \) and \( I(v_2^*) = \inf_{v \in \mathcal{U}_{ad}} I(v) \) respectively. Because \( \mathcal{U}_{ad} \) is a bounded closed convex set, we can get that \( \frac{1}{2}(v_1^* + v_2^*) \in \mathcal{U}_{ad} \). We thus can deduce that

\[
I\left( \frac{1}{2}(v_1^* + v_2^*) \right) < \frac{1}{2}I(v_1^*) + \frac{1}{2}I(v_2^*) = \inf_{v \in \mathcal{U}_{ad}} I(v),
\]

which is a contradiction unless \( v_1^* = v_2^* \). This completes the proof.

From the above analysis, we can conclude that \( \left( u(v^*; t, x), v^* \right) \) of \( \mathcal{S}(0, T) \times \mathcal{U}_{ad} \) is a unique optimal solution to the optimal control problem investigated.

4. The Sufficient and Necessary Optimality Condition

In this section, we shall characterize the optimal control by giving the sufficient and necessary condition for optimality. We firstly give the following lemma according to optimal control theory.

**Lemma 4.1.** Assume that the mapping \( v \to I(v) \) is differentiable, strictly convex and \( \mathcal{U}_{ad} \) is bounded. Then the unique element (optimal control) \( v^* \) in \( \mathcal{U}_{ad} \) satisfying \( I(v^*) = \inf_{v \in \mathcal{U}_{ad}} I(v) \) can be characterized by
$I'(v')(v-v') \geq 0,$ \hspace{1cm} (4.1)

where $\forall v \in U_{ad}$ and $I'(v')$ denote the derivative of $I(v)$ at $v = v'$.  

**Proof.** Let $v'$ be the optimal control subject to Theorem 3.1. Then for $\forall v \in U_{ad}$ and $v' \in (0,1)$, we have

$$I(v') = I((1-\theta)v' + \theta v^*) \leq I((1-\theta)v' + \theta v). \hspace{1cm} (4.2)$$

From inequality (4.2), we can derive that

$$\theta I(v' + \theta(v-v^*)) - I(v^*) \geq 0. \hspace{1cm} (4.3)$$

Therefore, if we pass to the limit in inequality (4.3), we obtain that

$$I'(v')(v-v') \geq 0, \hspace{1cm} (4.4)$$

Alternatively, suppose inequality (4.1) remains true. Because the mapping $v \mapsto I(v)$ is strictly convex, we can get

$$I((1-\theta)v' + \theta v^*) < (1-\theta)I(v') + \theta I(v), \hspace{1cm} \forall \theta \in (0,1). \hspace{1cm} (4.5)$$

From inequality (4.4), we deduce that

$$\theta I(v' + \theta(v-v^*)) - I(v^*) < (v) - I(v'). \hspace{1cm} (4.6)$$

If we pass the limit in inequality (4.5), we can get

$$0 \leq I'(v')(v-v') = \lim_{\theta \to 0} I((v^* + \theta(v-v^*)) - I(v') \hspace{1cm} (4.7)$$

for $\forall v \in U_{ad}$, which completes the proof.

Conditions of the type (4.1) are usually termed as “first order sufficient and necessary condition”, in terminology of calculus of variations. In order to analyze inequality (4.1), we need to prove that the mapping $v \mapsto u(v; t, x)$ of $U_{ad} \to S(0,T)$ is differentiable at $v = v'$.

**Definition 4.1.** The solution mapping $v \mapsto u(v; t, x)$ of $U$ into $S(0,T)$ is said to be differentiable at $v = v'$ in any direction $w$, if $\forall w \in U$ and $\theta \in (0,1)$, there exists a $u'(v'; t, x) \in L(U, S(0,T))$ such that

$$\theta [u(v^* + \theta w; t, x) - u(v'; t, x)] \to u'(v'; t, x)w \text{ in } S(0,T), \hspace{1cm} \theta \to 0. \hspace{1cm} (4.8)$$

The function $u'(v'; t, x)w \in S(0,T)$ is called the directional derivative of $u(v; t, x)$, which plays crucial in the following discussion.

**Theorem 4.1.** The mapping $v \mapsto u(v; t, x)$ of $U_{ad}$ into $S(0,T)$ is derivative at $v = v'$ and such the derivative of $u(v; t, x)$ at $v = v'$ in the direction $w = v - v^* \in U_{ad}$, say $g = u'(v^*; t, x)w$, is a weak solution of the following equation:

$$
\begin{align*}
\mathcal{G} + c_{gg} g_x + c_{gy} y + c_{bu} (v^*; t, x)g + g_{xx} = Bw, \\
g(t, x + L) = g(t, x), \forall x \in R, \forall t \in [0,T], \\
\mathcal{G}(0, x) = 0,
\end{align*}
\hspace{1cm} (4.9)
$$
where \( y = u(v; t, x) - u_{\alpha}(v; t, x) \) and \( G = g - g_{\alpha} \).

**Proof.** Let \( \theta \in (-1, 0) \cup (0, 1) \). We set \( g_\theta = \theta^{-1}(u(v + \theta w; t, x) - u(v; t, x)) \) and \( G_\theta = g_\theta - g_{\theta, \alpha} \). Then \( g_\theta \) satisfies

\[
\begin{align*}
G_{\theta, t} + c_\theta G_{\theta, x} + u(v; t, x)G_{\theta, x} + bg_{\theta, x} y_\theta + bu_\theta (v; t, x) G_{\theta} + \Gamma g_{\theta, xx} & = Bw, \\
g_\theta (t, x + L) & = g_\theta (t, x), \forall x \in R, \forall t \in [0, T], \\
G_{\theta}(0, x) & = 0,
\end{align*}
\]  

(4.7)

where \( y_\theta = u(v + \theta w; t, x) - u_{\alpha}(v + \theta w; t, x) \).

In order to estimate \( G_\theta \), we multiply both sides of the first equation in Equation (4.7) by \( 2g_\theta \) and integrate it over \( \Omega \). Then we get

\[
\frac{d}{dt}(\|g_\theta\|_{L^2}^2 + \|g_\theta\|_{L^2}^2) = (4 - 2b) \int_{\Omega} v_\theta g_\theta g_{\theta, x} dx + (2 - 2b) \int_{\Omega} u_{\alpha}(v; t, x) g_\theta g_{\theta, x} dx
\]

(4.8)

Each item on the right hand of Equation (4.8) can be estimated as follows:

\[
(4 - 2b) \int_{\Omega} v_\theta g_\theta g_{\theta, x} dx \leq |4 - 2b| \|v_\theta\|_{L^2} \int_{\Omega} \|g_\theta\|_{L^2} \|g_{\theta, x}\|_{L^2} \leq \frac{|4 - 2b| m_1}{2} \left( \|g_\theta\|_{L^2}^2 + \|g_\theta\|_{L^2}^2 \right);
\]

(2 - 2b) \int_{\Omega} u_{\alpha}(v; t, x) g_\theta g_{\theta, x} dx \leq |2 - 2b| \|u_{\alpha}(v; t, x)\|_{L^2} \int_{\Omega} \|g_\theta\|_{L^2} \|g_{\theta, x}\|_{L^2} \leq \frac{|2 - 2b| m_1}{2} \left( \|g_\theta\|_{L^2}^2 + \|g_\theta\|_{L^2}^2 \right);
\]

(3 - 2b) \int_{\Omega} u_\theta (v; t, x) g_{\theta, x}^2 dx \leq |3 - 2b| \|u_\theta (v; t, x)\|_{L^2} \|g_{\theta, x}\|_{L^2} \leq |3 - 2b| m_1 \|g_{\theta, x}\|_{L^2}^2;
\]

(1 - 2b) \int_{\Omega} u_\theta (v; t, x) g_\theta^2 dx \leq |1 - 2b| \|u_\theta (v; t, x)\|_{L^2} \|g_\theta\|_{L^2} \leq |1 - 2b| m_1 \|g_\theta\|_{L^2}^2
\]

and

\[
2 \int_{\Omega} (Bw) g_\theta dx \leq \|Bw\|_{L^2}^2 + \|g_\theta\|_{L^2}^2 \leq \lambda_1^2 \|Bw\|_{L^2}^2 + \|g_\theta\|_{L^2}^2 ,
\]

where \( \lambda_1 > 0 \) is an embedding constant and \( m_i > 0, i = 1, 2, 3 \) are some constants.

Hence, Equation (4.8) can be changed into

\[
\frac{d}{dt}(\|g_\theta\|_{L^2}^2 + \|g_\theta\|_{L^2}^2) \leq \beta_1 \left( \|g_\theta\|_{L^2}^2 + \|g_\theta\|_{L^2}^2 \right) + \lambda_1^2 \|Bw\|_{L^2}^2 ,
\]

(4.9)

where

\[
\beta_1 = \left\{ \frac{|4 - 2b| m_1}{2} + \frac{|2 - 2b| m_1}{2} + |1 - 2b| m_1 + 1, \frac{|4 - 2b| m_1}{2} + \frac{|2 - 2b| m_1}{2} + |3 - 2b| m_1 \right\}
\]

It follows from inequality (4.9) and the Gronwall’s lemma that

\[
\|g_\theta(0, x)\|_{L^2}^2 + \|g_\theta(0, x)\|_{L^2}^2 \leq \exp(\beta_1 \int_0^T \left( \|g_\theta(0, x)\|_{L^2}^2 + \|g_\theta(0, x)\|_{L^2}^2 \right) + \lambda_1^2 \int_0^T \|Bw\|_{L^2}^2 \exp(-\beta_1 s) ds)
\]

\( \triangleq Z_1 \),

(4.10)
where \( \forall t \in [0, T] \).

Next, multiplying both sides of the first equation in Equation (4.7) by \(-2g_{\theta,x}\) and integrating it over \( \Omega \), which gives

\[
\frac{d}{dt} \left( \|g_{\theta}\|_{L^2}^2 + \|g_{\theta}\|_{H^1}^2 \right) = (2b - 2) \int_{\Omega} g_{\theta,x} g_{\theta,x,x} v_{0} dx - 2 \int_{\Omega} g_{\theta} g_{\theta,x,x} v_{0} dx + 2 \int_{\Omega} u_{x} \left( v_{0}; t, x \right) g_{\theta,x} g_{\theta,x} dx + 2b \int_{\Omega} u_{x} \left( v_{0}; t, x \right) g_{\theta} g_{\theta} dx + (1 - 2b) \int_{\Omega} u_{x} \left( v_{0}; t, x \right) g_{\theta,x}^2 dx - 2 \int_{\Omega} (Bw) g_{\theta,x} dx.
\]

Then, we estimate each item of the right hand of Equation (4.11) as follows:

\[
(2b - 2) \int_{\Omega} g_{\theta,x} g_{\theta,x,x} v_{0} dx \leq \frac{2b - 2}{2} \left( \|g_{\theta}\|_{L^2}^2 + \|g_{\theta}\|_{H^1}^2 \right);
\]

\[
-2 \int_{\Omega} g_{\theta} g_{\theta,x,x} v_{0} dx \leq \|v_{0}\|_{L^2} \int_{\Omega} 2g_{\theta} g_{\theta,x,x} dx = 0;
\]

\[
2 \int_{\Omega} u_{x} \left( v_{0}; t, x \right) g_{\theta,x} g_{\theta,xx} dx \leq \|u_{x} \left( v_{0}; t, x \right)\|_{L^2} \int_{\Omega} 2g_{\theta,x} g_{\theta,x} dx \leq m_{1} \left( \|g_{\theta}\|_{L^2}^2 + \|g_{\theta}\|_{H^1}^2 \right);
\]

\[
2b \int_{\Omega} u_{x} \left( v_{0}; t, x \right) g_{\theta} g_{\theta,x} dx \leq \|u_{x} \left( v_{0}; t, x \right)\|_{L^2} \left( \|g_{\theta}\|_{H^2}^2 + \|g_{\theta}\|_{H^1}^2 \right) \leq b m_{4} \left( \|g_{\theta}\|_{H^2}^2 + \|g_{\theta}\|_{H^1}^2 \right);
\]

\[
(1 - 2b) \int_{\Omega} u_{x} \left( v_{0}; t, x \right) g_{\theta,x}^2 dx \leq \left| 1 - 2b \right| \left( \|u_{x} \left( v_{0}; t, x \right)\|_{L^2} \|g_{\theta,x}\|_{L^2}^2 \right) \leq \left| 1 - 2b \right| m_{3} \left( \|g_{\theta}\|_{H^2}^2 \right)
\]

and

\[
-2 \int_{\Omega} (Bw) g_{\theta,x} dx \leq 2 \|Bw\|_{L^2} \|g_{\theta,x}\|_{L^2} \leq \|Bw\|_{L^2} \|g_{\theta}\|_{H^1}^2 \leq \lambda_{2} \|Bw\|_{L^2} + \|g_{\theta}\|_{H^1}^2,
\]

where \( \lambda_{2} > 0 \) is an embedding constant and \( m_{i} > 0, i = 1, 3, 4 \) are some constants.

By the above estimates, we can deduce from Equation (4.11) that

\[
\frac{d}{dt} \left( \|g_{\theta}\|_{L^2}^2 + \|g_{\theta}\|_{H^1}^2 \right) \leq \beta_{4} \left( \|g_{\theta}\|_{L^2}^2 + \|g_{\theta}\|_{H^1}^2 \right) + \lambda_{2} \|Bw\|_{L^2}^2,
\]

where

\[
\beta_{4} = \max \left\{ \frac{2b - 2}{2} m_{4} + \|m_{2}\|_{L^2} \lambda_{2}^2 + \|m_{3}\|_{L^2} \lambda_{2}^2 \right\}.
\]

Applying Gronwall’s lemma to inequality (4.12), which yields

\[
\|g_{\theta}\|_{L^2}^2 + \|g_{\theta}\|_{H^1}^2 \leq \exp \left( \beta_{4} t \right) \left( \|g_{\theta}(0, x)\|_{L^2}^2 + \|g_{\theta}(0, x)\|_{H^1}^2 \right) + \lambda_{2} \int_{0}^{t} \|Bw\|_{L^2}^2 \exp \left( -\beta_{4} s \right) ds \leq Z_{2},
\]

where \( \forall t \in [0, T] \).
Similarly, multiplying both sides of the first equation in Equation (4.7) by $2g_{\theta,xxx}$ and integrating it over $\Omega$, which gives

\[
\frac{d}{dt} \left( \| g_\theta \|_{H^2(\Omega)} + \| g_\theta \|_{H^3(\Omega)} \right)
= (2 - 2b) \int_\Omega g_{\theta,x} g_{\theta,xxx} \phi_\theta \, dx + 2 \int_\Omega g_{\theta} g_{\theta,xxx} \phi_\theta \, dx
- (2b + 3) \int_\Omega u_x (v'; t, x) g_{\theta,x} \phi_\theta \, dx
- (2b + 1) \int_\Omega u_x (v'; t, x) g_{\theta,x} g_{\theta,xxx} \phi_\theta \, dx + 2b \int_\Omega u_x (v'; t, x) g_{\theta} g_{\theta,xxx} \phi_\theta \, dx
- 2b \int_\Omega u_x (v'; t, x) g_{\theta,xxx} \phi_\theta \, dx + 2 \int_\Omega (Bw) g_{\theta,xxx} \phi_\theta \, dx.
\]  

(4.14)

We can also estimate each item of the right hand of Equation (4.14) as follows:

\[
(2 - 2b) \int_\Omega g_{\theta,x} g_{\theta,xxx} \phi_\theta \, dx \leq [2 - 2b] \| \phi_\theta \|_{L^2} \int_\Omega g_{\theta,x} g_{\theta,xxx} \phi_\theta \, dx = 0;
\]

\[
2 \int_\Omega g_{\theta} g_{\theta,xxx} \phi_\theta \, dx \leq \| \phi_\theta \|_{L^2} \int_\Omega 2g_{\theta} g_{\theta,xxx} \phi_\theta \, dx = 0;
\]

\[
- (2b + 3) \int_\Omega u_x (v'; t, x) g_{\theta,x} \phi_\theta \, dx \leq [2b + 3] \| u_x \|_{L^2} \int_\Omega g_{\theta,x} \phi_\theta \, dx \leq [2b + 3] m_3 \int_\Omega g_{\theta,xxx} \phi_\theta \, dx;
\]

\[
- (2b + 1) \int_\Omega u_x (v'; t, x) g_{\theta,x} g_{\theta,xxx} \phi_\theta \, dx \leq [2b + 1] \| u_x \|_{L^2} \int_\Omega g_{\theta,xxx} \phi_\theta \, dx \leq [2b + 1] m_3 \int_\Omega g_{\theta,xxx} \phi_\theta \, dx;
\]

\[
- (2b + 2) \int_\Omega u_x (v'; t, x) g_{\theta} g_{\theta,xxx} \phi_\theta \, dx \leq [2b + 2] m_2 \left( \lambda_2^2 \| g_{\theta} \|_{H^2(\Omega)}^2 + \| g_{\theta} \|_{H^3(\Omega)}^2 \right);
\]

\[
2b \int_\Omega u_x (v'; t, x) g_{\theta,xxx} \phi_\theta \, dx \leq [2b] \| u_x \|_{L^2} \int_\Omega \lambda_2^2 \| g_{\theta} \|_{H^2(\Omega)}^2 + \| g_{\theta} \|_{H^3(\Omega)}^2 \;
\]

\[
- 2b \int_\Omega u_x (v'; t, x) g_{\theta,xxx} \phi_\theta \, dx \leq [2b] \| u_x \|_{L^2} \int_\Omega \lambda_2^2 \| g_{\theta} \|_{H^2(\Omega)}^2 + \| g_{\theta} \|_{H^3(\Omega)}^2 \]
\]

and

\[
2 \int_\Omega (Bw) g_{\theta,xxx} \phi_\theta \, dx \leq 2 \int_\Omega (Bw) \phi_\theta \| g_{\theta,xxx} \|_{H^2(\Omega)} \leq 2 \int_\Omega (Bw) \| g_{\theta,xxx} \|_{H^2(\Omega)} + \| g_{\theta} \|_{H^3(\Omega)} \]

where $m_i > 0, i = 1, 2, 3$ are some constants and $\lambda_i > 0, i = 4, 5$ are some embedding constants.

Combining a series of complex estimates above and Equation (4.14), we can obtain that

\[
\frac{d}{dt} \left( \| g_\theta \|_{H^2(\Omega)} + \| g_\theta \|_{H^3(\Omega)} \right) \leq \beta_s \left( \| g_\theta \|_{H^2(\Omega)} + \| g_\theta \|_{H^3(\Omega)} \right) + \| Bw \|_{L^2}^2 ,
\]  

(4.15)

where

\[
\beta_s = \max \left\{ \left[ \left( \frac{2b + [2(\lambda_2^2 + 1)]}{2} \right) + [2b + 3] m_3 + [2b + 1] m_3 \right] \right. \}
\]

By applying the Gronwall’s lemma to inequality (4.15), we can get

\[
(\| g_\theta \|_{H^2(\Omega)} + \| g_\theta \|_{H^3(\Omega)}) \leq \exp (\beta_s t) \left[ \left( \| g_\theta (0, x) \|_{H^2(\Omega)}^2 + \| g_\theta (0, x) \|_{H^3(\Omega)}^2 \right) + \int_0^t \| Bw \|_{L^2}^2 \exp (-\beta_s \tau) \right] = Z_s ,
\]  

(4.16)
where \( \forall t \in [0,T] \).

Combining estimate inequality (4.13) and (4.16), we can deduce that
\[
\|g'_{\theta,x}\|_{L^2(\Omega_t)}^2 = \|g_{\theta,x} - g_{\theta,x,0}\|_{L^2(\Omega_t)}^2 + 2\|g_{\theta,x}\|_{H^1(\Omega_t)}^2 + \|g_{\theta}\|_{L^2(\Omega_t)}^2 \leq Z_2 + Z_3. \tag{4.17}
\]

Similarly, combining estimate inequality (4.10) and (4.13), we can obtain that
\[
\|g'_{\theta}\|_{L^2(\Omega_t)}^2 = \|g_{\theta} - g_{\theta,0}\|_{L^2(\Omega_t)}^2 + 2\|g_{\theta}\|_{H^1(\Omega_t)}^2 + \|g_{\theta}\|_{L^2(\Omega_t)}^2 \leq Z_1 + Z_2. \tag{4.18}
\]

From inequality (4.17), we derive that
\[
\|g'_{\theta}\|_{L^2(\Omega_t)}^2 \leq (Z_2 + Z_3)T, \tag{4.19}
\]
which indicates a uniformly \( L^2(0,T;V) \) bounded of \( G_{\theta} \).

Afterward, we will prove a uniformly \( L^2(0,T;V') \) bounded of \( G_{\theta,x} \).

From the first equation in Equation (4.7) and the Sobolev embedding theorem, we have
\[
\|g_{\theta,x}\|_{L^2(\Omega_t)} \leq \|B_{\theta}\|_{L^2} + c_0 \|g_{\theta}\|_{H^1(\Omega_t)} + \lambda_2 \|g_{\theta}\|_{L^2(\Omega_t)} \|u(v', t, x)\|_{L^2(\Omega_t)} + \|\theta\|_{L^2(\Omega_t)} \|v\|_{L^2(\Omega_t)} + \|\theta\|_{L^2(\Omega_t)} \|v\|_{L^2(\Omega_t)} \|u(v', t, x)\|_{L^2(\Omega_t)} + \|\theta\|_{L^2(\Omega_t)} \|v\|_{L^2(\Omega_t)} \|u(v', t, x)\|_{L^2(\Omega_t)}
\]
\[
\leq \lambda_3 \|B_{\theta}\|_{L^2} + c_0 Z_1^\frac{1}{2} + \lambda_2 m_5 Z_2^\frac{1}{2} + m_4 (Z_1 + Z_2)^\frac{1}{2}
\]
\[
+ \|\theta\|_{L^2(\Omega_t)} \lambda_4 m_6 (Z_2 + Z_3)^\frac{1}{2} + \|\theta\|_{L^2(\Omega_t)} \|v\|_{L^2(\Omega_t)} \|u(v', t, x)\|_{L^2(\Omega_t)} + \|\theta\|_{L^2(\Omega_t)} \|v\|_{L^2(\Omega_t)} \|u(v', t, x)\|_{L^2(\Omega_t)} \tag{4.20}
\]
where \( \lambda_i > 0, i = 2, 3 \) are some embedding constants and \( m_i > 0, i = 1, 4, 5, 6 \) are some constants.

Analogously, from inequality (4.20), we can get
\[
\|g_{\theta,x}\|_{L^2(\Omega_t)}^2 \leq \left[ \lambda_3 \|B_{\theta}\|_{L^2} + c_0 Z_1^\frac{1}{2} + \lambda_2 m_5 Z_2^\frac{1}{2} + m_4 (Z_1 + Z_2)^\frac{1}{2} + \|\theta\|_{L^2(\Omega_t)} \lambda_4 m_6 (Z_2 + Z_3)^\frac{1}{2} \right] T. \tag{4.21}
\]

Combining inequality (4.19) and (4.21), we can establish the boundedness of \( G_{\theta} \) in \( \mathcal{W}(0,T) \). Hence, from Lemma 2.1, we can deduce that
\[
\|g_{\theta}\|_{L^2(0,T)} \leq C \|g_{\theta}\|_{L^2(0,T)} < +\infty. \tag{4.22}
\]

From now on, we can infer that there exists a \( g \in S(0,T) \) and a sequence \( \{\theta_k\} \subset (-1,1) \) tending to 0 such that
\[
g_{\theta_k} \xrightarrow{weakly} g \text{ in } S(0,T), \text{ as } k \to \infty. \tag{4.22}
\]

Because the imbedding \( S(0,T) \) into \( L^2(0,T;H^2(\Omega)) \) is compact, then it can deduce from (4.22) that
\[
g_{\theta_k} \xrightarrow{strongly} g \text{ in } H^2(\Omega) \text{ a.e. } t \in [0,T], \tag{4.23}
\]
for some \( \{\theta_k\} \subset (-1,1) \) tending to 0 as \( k \to \infty \). Whence by (4.22) - (4.23), Theorem 2.1 and the Lebesgue dominated convergence theorem, we can easily obtain that
\[ g_{\theta} y_{\theta, x} \xrightarrow{\text{weakly}} g y_x \text{ in } L^2(0, T; H); \]  
\[ g_{\theta} y_{\theta} \xrightarrow{\text{strongly}} g_y \text{ in } L^2(0, T; H); \]  
\[ G_{\theta} \xrightarrow{\text{weakly}} G \text{ in } L^2(0, T; V); \]  
\[ g_{\theta, xxx} \xrightarrow{\text{weakly}} g_{xxx} \text{ in } L^2(0, T; H); \]  
as \( k \to \infty \), where \( G = g - g_{xx} \). And also we can derive from Equation (4.7) and inequality (4.21) that
\[ G_{\theta, x} \xrightarrow{\text{weakly}} G_{x} \text{ in } L^2(0, T; V^*), \text{ as } k \to \infty. \]  

Therefore, we can infer from (4.24) to (4.28) that
\[ g_{\theta} \xrightarrow{\text{weakly}} g = u^t(v^*; t, x)w \]  
in \( S(0, T) \) as \( \theta \to 0 \) in which \( g \) is a solution of Equation (4.6).

Consequently, the solution mapping \( v \to u(v; t, x) \) of \( \mathcal{U}_{ad} \) into \( S(0, T) \) is differentiable in the weak topology of \( S(0, T) \). This completes the proof.

The conclusion of Theorem 4.1 means that the cost \( I(v) \) is derivative at \( v^* \) in the direction \( v^* - v \). So, we can get that
\[ I'(v^*)(v^* - v) \]
\[ = \lim_{\theta \to 0} \theta \]
\[ = \lim_{\theta \to 0} \theta^{-1} \left( [Cu(v^* + \theta(v^* - v)) - z_d, Cu(v^* + \theta(v^* - v)) - z_d]_M \right. \]
\[ - [Cu(v^*) - z_d, Cu(v^*) - z_d]_M \]
\[ + \lim_{\theta \to 0} \theta^{-1} \left( [N(v^* + \theta(v^* - v)), v^* + \theta(v^* - v)]_H - (Nv^*, v^*)_H \right) \]
\[ = 2[Cu(v^*) - z_d, Cu(v^*) - z_d]_M + 2(Nv^*, v^*)_H \]

Then the sufficient and necessary optimality condition (4.1) can be rewritten as
\[ \left( Cu(v^*; t, x) - z_d, Cu'(v^*; t, x)(v^* - v) \right)_M + (Nv^*, v^*)_H \]
\[ = \left( C^* \Lambda_M \left( Cu(v^*; t, x) - z_d, u'(v^*; t, x)(v^* - v) \right)_M \right)_S(0, T) \]
\[ + (Nv^*, v^*)_H \geq 0, \]  
for \( \forall v \in \mathcal{U}_{ad} \), where \( \Lambda_M \) is the canonical isomorphism \( M \) onto \( M' \) and \( z_d \in M \) is desired value.

5. The Two Cases of Distributive Observations

In this section, we will characterize the optimal control by giving the sufficient and necessary optimality condition (4.29) for the following two cases of physical meaningful observations:

(I) We set \( M = L^2(0, T; H) \) and \( C \in L(S(0, T), M) \), then observe that
\( z(v; t, x) = Cu(v; t, x) = u(v; t, x) \in L^2(0, T; H) \).

(II) We set \( M = L^2(0, T; H) \) and \( C \in L(S(0, T), M) \), then observe that

\[
I(v) = \int_0^T \int_\Omega (u(v; t, x) - z_d(t, x))^2 + (Nv, v)_{\Omega_d} \, dt, \quad \forall v \in U_{ad} \subset U,
\]

where \( z_d(t, x) \in M \) is a desired value. Let \( v^* \) be the optimal control subject to Equation (3.1) and cost functional (5.1). Then the sufficient and necessary optimality condition (4.29) can be represented by

\[
\int_0^T \int_\Omega (u(v^*; t, x) - z_d(t, x)) gdxdt + (Nv^*, v - v^*)_{\Omega_d} \geq 0, \quad \forall v \in U_{ad},
\]

where \( g = u(v^*; t, x) (v - v^*) \) is the weak solution of Equation (4.6). Now we will introduce the adjoint system to describe the optimality condition (5.2):

\[
\begin{align*}
-\Psi_t (v^*; t, x) - c_0 \psi_x (v^*; t, x) - u(v^*; t, x) \Psi_x (v^*; t, x) \\
+ (3 - 2b)u_{ss} (v^*; t, x) \psi_x (v^*; t, x) + (3 - b)u_s (v^*; t, x) \psi_{ss} (v^*; t, x) \\
-bv^* (v^*; t, x) \psi_x (v^*; t, x) - \Gamma \psi_{sss} (v^*; t, x) = u(v^*; t, x) - z_d(t, x), \\
\psi (v^*; t, x + L) = \psi (v^*; t, x), \quad \forall x \in R, \forall t \in [0, T], \\
\Psi (v^*; T, x) = 0,
\end{align*}
\]

where

\[
\Psi (v^*; t, x) = \psi (v^*; t, x) - \psi_{ss} (v^*; t, x)
\]

and

\[
y (v^*; t, x) = u(v^*; t, x) - u_{ss} (v^*; t, x).
\]

Therefore, we can provide the characterization for the optimal control \( v^* \) of the quadratic cost functional (5.1) as follows:

**Theorem 5.1.** The optimal control \( v^* \) of the quadratic cost functional (5.1) is characterized by the following control system, adjoint system and inequality:

\[
\begin{align*}
y_t (v^*; t, x) + c_0 y_x (v^*; t, x) + u (v^*; t, x) y_x (v^*; t, x) \\
+ bu_x (v^*; t, x) y_t (v^*; t, x) + \Gamma u_{ss} (v^*; t, x) = Bv^*, \\
u (v^*; t, x + L) = u (v^*; t, x), \quad \forall x \in R, \forall t \in [0, T], \\
y (v^*; 0, x) = y_0 (x) \in V, \\
-\Psi_t (v^*; t, x) - c_0 \psi_x (v^*; t, x) - u (v^*; t, x) \Psi_x (v^*; t, x) \\
+ (3 - 2b)u_{ss} (v^*; t, x) \psi_x (v^*; t, x) + (3 - b)u_s (v^*; t, x) \psi_{ss} (v^*; t, x) \\
- bv^* (v^*; t, x) \psi_x (v^*; t, x) - \Gamma \psi_{sss} (v^*; t, x) = u (v^*; t, x) - z_d(t, x), \\
\psi (v^*; t, x + L) = \psi (v^*; t, x), \quad \forall x \in R, \forall t \in [0, T], \\
\Psi (v^*; T, x) = 0, \\
\int_0^T \int_\Omega \psi (v^*; t, x) B (v - v^*) dxdt + (Nv^*, v - v^*)_{\Omega_d} \geq 0, \quad \forall v \in U_{ad},
\end{align*}
\]
where
\[ y(v^*;t,x) = u(v^*;t,x) - u_{xx}(v^*;t,x) \]

and
\[ \Psi(v^*;t,x) = \psi(v^*;t,x) - \psi_{xx}(v^*;t,x). \]

**Proof.** Taking inner product of the first equation in Equation (5.3) by \( g \) over \( \Omega \), then integrating the result equation with respect to \( t \) on \([0,T]\), we get

\[
\int_0^T \int_\Omega -\Psi_g gdxdt - c_0 \int_0^T \int_\Omega \psi_g gdxdt - \int_0^T \int_\Omega u\psi_g gdxdt + (3-2b) \int_0^T \int_\Omega u_{xx}\psi gdxdt + (3-b) \int_0^T \int_\Omega u_x\psi_{xx} gdxdt - b \int_0^T \int_\Omega \psi gdxdt - \Gamma \int_0^T \int_{\partial \Omega} \psi_{xx} gdxdt = \int_0^T \int_{\Omega} (u-z_d) gdxdt. \tag{5.4}
\]

Combining Equation (4.6) and Equation (5.3) and taking integration by parts, the left hand side of Equation (5.4) yields

\[
\int_0^T \int_\Omega \psi \left( G + c_0 g_x + g_y + uG_x + bG_x y + buG + \Gamma g_{xx} \right) dxdt = \int_0^T \int_\Omega B \left( v - v^* \right) dxdt, \tag{5.5}
\]

where \( G = g - g_{xx} \). Therefore, utilizing Equation (5.4) and Equation (5.5), the sufficient and necessary optimality condition (5.2) is equivalent to

\[
\int_0^T \int_\Omega \psi \left( v^*;t,x \right) B \left( v - v^* \right) dxdt + \left( Nv^*, v - v^* \right)_{\Omega_t} \geq 0, \quad \forall \ v \in \mathcal{U}_{ad}. \]

Hence, the theorem is proved.

Secondly, we discuss the cost functional expressed by

\[
I(v) = \int_0^T \left[ \int_\Omega \left( v; t, x \right) - z_d(t,x) \right]^2 dxdt + \left( Nv, v \right)_{\Omega_t}, \quad \forall \ v \in \mathcal{U}_{ad} \subset \mathcal{U}, \tag{5.6}
\]

where \( z_d(t,x) \in \mathcal{M} \) is a desired value. Let \( v^* \) be the optimal control subject to Equation (3.1) and cost functional (5.6). Then the sufficient and necessary optimality condition (4.29) is represented by

\[
\int_0^T \int_\Omega \left( v; t, x \right) - z_d(t,x) \right) Gdxdt + \left( Nv^*, v - v^* \right)_{\Omega_t} \geq 0, \quad \forall \ v \in \mathcal{U}_{ad}, \tag{5.7}
\]

where \( G = g - g_{xx} \) and \( g = u'(v^*;t,x)w \) is the weak solution of Equation (4.6). Similarly, we formulate the adjoint system to describe the optimality condition (5.7):

\[
\begin{align*}
-\Psi_t(v^*;t,x) - c_0 \psi_{xt}(v^*;t,x) - u(v^*;t,x) \Psi_t(v^*;t,x) + (3-2b) u_{xx}(v^*;t,x) \psi_t(v^*;t,x) + (3-b) u_t(v^*;t,x) \psi_{xx}(v^*;t,x) - b \psi_t(v^*;t,x) - \Gamma \psi_{xx}(v^*;t,x) = (I - \partial^2_y) \left( y(v^*;t,x) - z_d(t,x) \right), \\
\psi(v^*;t,x + L) = \psi(v^*;t,x), \forall x \in R, \forall t \in [0,T], \\
\Psi(v^*;T,x) = 0,
\end{align*}
\tag{5.8}
\]
where
\[ y(v^*;t,x) = u(v^*;t,x) - u_{ss}(v^*;t,x) \]
and
\[ \Psi(v^*;t,x) = \psi(v^*;t,x) - \psi_{ss}(v^*;t,x). \]

Hence, we can give the following theorem.

**Theorem 5.2.** The optimal control \( v^* \) of the quadratic cost functional (5.7) is characterized by the following control system, adjoint system and inequality:

\[
\begin{cases}
    y(v^*;t,x) + c_0\mu_s(v^*;t,x) + u(v^*;t,x)y_s(v^*;t,x) \\
    + bu_r(v^*;t,x)y_r(v^*;t,x) + \Gamma u_{sss}(v^*;t,x) = Bv^*,
\end{cases}
\]

\[ u(v^*;t,x + L) = u(v^*;t,x), \forall x \in R, \forall t \in [0,T], \]

\[ y(v^*;0,x) = y_0(x) \in V, \]

\[ \Psi(v^*;t,x) - c_0\psi_s(v^*;t,x) - u(v^*;t,x)\psi_s(v^*;t,x) \\
    + (3 - 2b)u_{ss}(v^*;t,x)\psi_s(v^*;t,x) + (3 - b)u_{ssss}(v^*;t,x)\psi_{ss}(v^*;t,x) \\
    - by(v^*;t,x)\psi_s(v^*;t,x) - \Gamma \psi_{sss}(v^*;t,x) = (I - \partial_x^2)(y(v^*;t,x) - z_d(t,x)), \]

\[ \psi(v^*;t,x + L) = \psi(v^*;t,x), \forall x \in R, \forall t \in [0,T], \]

\[ \psi(v^*;T,x) = 0, \]

\[ \int_0^T \int_\Omega \psi(v^*;t,x)B(v - v^*)dxdt + \left( Nv^*, v - v^* \right)_{L^2(\Omega)} \geq 0, \forall v \in U_{ad}, \]

where
\[ y(v^*;t,x) = u(v^*;t,x) - u_{ss}(v^*;t,x) \]
and
\[ \Psi(v^*;t,x) = \psi(v^*;t,x) - \psi_{ss}(v^*;t,x). \]

**Proof.** As we did before, we multiply both sides of the first equation of Equation (5.8) by \( g \) and integrate it over \( [0,T] \times \Omega \). Then we have

\[
\int_0^T \int_\Omega \left[ -c_0\psi_s - u\psi_s + (3 - 2b)u_s\psi_s + (3 - b)u_{sss}\psi_s - by\psi_s - \Gamma \psi_{sss} \right] gdxdt
\]

\[ = \int_0^T \int_\Omega \left( I - \partial_x^2 \right)(y - z_d) gdxdt \]

\[ = \int_0^T \int_\Omega (y - z_d) Gdxdt, \]

where \( G = g - g_{xx} \).

Utilizing Equation (4.6), the integration by parts on the left hand side of Equation (5.9) yields

\[
\int_0^T \int_\Omega \psi(G + c_0g_x + gy_x + uG_x + bg_x + \psi + \Gamma g_{xx}) dxdt
\]

\[ = \int_0^T \int_\Omega \psi B(v - v^*) dxdt, \]

where \( G = g - g_{xx} \). Therefore, combining Equation (5.9) and Equation (5.10), the sufficient and necessary optimality condition (5.7) is equivalent to
\[
\int_0^T \int_\Omega \psi \left( v^\prime, t, x \right) B \left( v - v^\prime \right) \, dx \, dt + \left( N v^\prime, v - v^\prime \right)_{\Omega} \geq 0, \quad \forall v \in U_{ad},
\]
which completes the proof.

6. Conclusions

\textit{b-equation} is an important shallow water wave equation which has many practical meanings. In this paper, we aim at pursuing an in-depth study of the optimal control issue of the classical \textit{b-equation}. So, we investigate firstly the local existence and uniqueness of solution to the initial-boundary problem of the \textit{b-equation} with source term, and then discuss the formulation of the quadratic cost optimal control problem for the \textit{b-equation}, obtain the existence and uniqueness of an optimal control, establish the sufficient and necessary optimality condition of an optimal control in fixed final horizon case. Moreover, we give the specific sufficient and necessary optimality condition for two physical meaningful distributive observation cases by employing associate adjoint systems. Compared with other papers in similar directions, the weak solution analysis of \textit{b-equation} without relying on viscous item is one technical innovation, and the sufficient and necessary optimality condition of an optimal control which is not limited to the necessary condition is another novelty. However, much work remains to be done in this direction. For example, it is an optimal control problem of the distributed parameter system governed by the nonlinear partial differential equation, to obtain the numerical solutions for the optimal control-trajectory pair is not an easy job due to the tremendous calculation and possible model difficulties. We try to finish this non-trivial work in the follow-up research by optimizing numerical algorithm and carrying out numerical simulation, which can provide a basis for application in the engineering field.

Acknowledgements

Research was supported in part by the National Natural Science Foundation of China (No: 11371175, 11501253, 11571140).

References

[1] Escher, J. and Yin, Z. (2008) Well-Posedness, Blow-Up Phenomena, and Global Solutions for the \textit{b-Equation}. \textit{Journal für die reine und angewandte Mathematik}, \textbf{624}, 51-80. \url{https://doi.org/10.1515/CRELLE.2008.080}

[2] Dullin, H.R., Gottwald, G.A. and Holm, D.D. (2003) Camassa-Holm, Korteweg-de Vries-5 and Other Asymptotically Equivalent Equations for Shallow Water Waves. \textit{Fluid Dynamics Research}, \textbf{33}, 73-95. \url{https://doi.org/10.1016/S0169-5983(03)00046-7}

[3] Dullin, H.R., Gottwald, G.A. and Holm, D.D. (2004) On Asymptotically Equivalent Shallow Water Wave Equations. \textit{Journal of Physics D}, \textbf{190}, 1-14. \url{https://doi.org/10.1016/j.physd.2003.11.004}

[4] Holm, D.D. and Staley, M.F. (2003) Wave Structure and Nonlinear Balances in a Family of Evolutionary PDEs. \textit{SIAM Journal on Applied Dynamical Systems}, \textbf{2}, 323-380. \url{https://doi.org/10.1137/S1111111102410943
[5] Holm, D.D. and Staley, M.F. (2003) Nonlinear Balance and Exchange of Stability in Dynamics of Solitons, Peakons, Ramps/Cliffs and Leftons in a 1-1 Nonlinear Evolutionary PDE. *Physics Letters A*, 308, 437-444. https://doi.org/10.1016/S0375-9601(03)00114-2

[6] Mikhailov, A.V. and Novikov, V.S. (2002) Perturbative Symmetry Approach. *Journal of Physics A*, 35, 4779-4790. https://doi.org/10.1088/0305-4470/35/22/309

[7] Degasperis, A., Holm, D.D. and Hone, A.N.W. (2003) Integrable and Non-Integrable Equations with Peakons. In: Ablowitz, M.J., Boiti, M., Pempinelli, F. and Prinari, B., Eds., *Nonlinear Physics Theory and Experiment II*, World Scientific Publication, River Edge, NJ, 37-43. https://doi.org/10.1142/9789812704467_0005

[8] Degasperis, A. and Procesi, M. (1999) Asymptotic Integrability. In: Degasperis, A. and Gaeta, G., Eds., *Symmetry and Perturbation Theory*, World Scientific Publication, River Edge, NJ, 23-37.

[9] Constantin, A. (2006) The Trajectories of Particles in Stokes Waves. *Inventiones Mathematicae*, 166, 523-535. https://doi.org/10.1007/s00222-006-0002-5

[10] Toland, J.F. (1996) Stokes Waves. *Topological Methods in Nonlinear Analysis*, 7, 1-48. https://doi.org/10.12775/TMNA.1996.001

[11] Constantin, A. and Escher, J. (2007) Particle Trajectories in Solitary Water Waves. *Bulletin of the American Mathematical Society*, 44, 423-431. https://doi.org/10.1090/S0273-0979-07-01159-7

[12] Dullin, H.R., Gottwald, G.A. and Holm, D.D. (2001) An Integrable Shallow Water Equation with Linear and Nonlinear Dispersion. *Physical Review Letters*, 87, 4501-4504. https://doi.org/10.1103/PhysRevLett.87.194501

[13] McKean, H.P. (1979) Integrable Systems and Algebraic Curves. In: Grmela, M. and Marsden, J.E., Eds., *Global Analysis, Lecture Notes in Mathematics Vol. 755*, Springer, Berlin, 83-200. https://doi.org/10.1007/bfb0069806

[14] Bona, J.L. and Scott, R. (1976) Solutions of the Korteweg-de Vries Equation in Fractional Order Sobolev Spaces. *Duke Mathematical Journal*, 43, 87-99. https://doi.org/10.1215/S0012-7094-76-04309-X

[15] Kenig, C., Ponce, G. and Vega, L. (1993) Well-Posedness and Scattering Results for the Generalized Korteweg-de Vries Equation via the Contraction Principle. *Communications on Pure and Applied Mathematics*, 46, 527-620. https://doi.org/10.1002/cpa.3160460405

[16] Tao, T. (2002) Low-Regularity Global Solutions to Nonlinear Dispersive Equations. In: Hassell, A., Ed., *Surveys in Analysis and Operator Theory*, Centre for Mathematics and Its Applications, Mathematical Sciences Institute, The Australian National University, Canberra, 19-48.

[17] Whitham, G.B. (1974) Linear and Nonlinear Waves. John Wiley & Sons, Inc., New York.

[18] Camassa, R. and Holm, D.D. (1993) An Integrable Shallow Water Equation with Peakon Solitons. *Physical Review Letters*, 71, 1661-1664. https://doi.org/10.1103/PhysRevLett.71.1661

[19] Fuchssteiner, B. and Fokas, A.S. (1981) Symplectic Structures, Their Bäcklund Transformation and Hereditary Symmetries. *Journal of Physics D*, 4, 47-66. https://doi.org/10.1016/0167-2789(81)90004-X

[20] Johnson, R.S. (2002) Camassa-Holm, Korteweg-de Vries and Related Models for Water Waves. *Journal of Fluid Mechanics*, 455, 63-82. https://doi.org/10.1017/S0022112001007224

[21] Constantin, A. and Kolev, B. (2003) Geodesic Flow on the Diffeomorphism Group
of the Circle. *Commentarii Mathematici Helvetici*, **78**, 787-804. https://doi.org/10.1007/s00014-003-0785-6

[22] Constantin, A. and Lannes, D. (2009) The Hydrodynamical Relevance of the Camassa-Holm and Degasperis-Procesi Equations. *Archive for Rational Mechanics and Analysis*, **192**, 165-186. https://doi.org/10.1007/s00205-008-0128-2

[23] Constantin, A. and Kolev, B. (2002) On the Geometric Approach to the Motion of Inertial Mechanical Systems. *Journal of Physics A*, **35**, R51-R79. https://doi.org/10.1088/0305-4470/35/32/201

[24] Constantin, A. (1998) On the Inverse Spectral Problem for the Camassa-Holm Equation. *Journal of Functional Analysis*, **155**, 352-363. https://doi.org/10.1006/jfan.1997.3231

[25] Constantin, A. and McKea, H.P. (1999) A Shallow Water Equation on the Circle. *Communications on Pure and Applied Mathematics*, **52**, 949-982. https://doi.org/10.1002/(SICI)1097-0312(199908)52:8<949::AID-CPA3>3.0.CO;2-D

[26] Constantin, A. (2001) On the Scattering Problem for the Camassa-Holm Equation. *Proceedings of the Royal Society of London A*, **457**, 953-970. https://doi.org/10.1098/rspa.2000.0701

[27] Constantin, A., Gerdjikov, V.S. and Ivanov, R.I. (2006) Inverse Scattering Transform for the Camassa-Holm Equation. *Inverse Problems*, **22**, 2197-2207. https://doi.org/10.1088/0266-5611/22/6/017

[28] Camassa, R., Holm, D.D. and Hyman, J.M. (1994) A New Integrable Shallow Water Equation. *Advances in Applied Mechanics*, **31**, 1-33. https://doi.org/10.1016/S0065-2156(08)70254-0

[29] Constantin, A. and Strauss, W.A. (2000) Stability of Peakons. *Communications on Pure and Applied Mathematics*, **53**, 603-610. https://doi.org/10.1002/(SICI)1097-0312(20000515)53:5<603::AID-CPA3>3.0.CO;2-L

[30] Constantin, A. and Strauss, W.A. (2002) Stability of the Camassa-Holm Solitons. *Journal of Nonlinear Science*, **12**, 415-422. https://doi.org/10.1007/s00332-002-0517-x

[31] Constantin, A. and Escher, J. (1998) Global Existence and Blow-Up for a Shallow Water Equation. *Annali della Scuola Normale Superiore di Pisa - Classe di Scienze*, **26**, 303-328. http://www.numdam.org/item?id=ASNSP_1998_4_26_2_303_0

[32] Danchin, R. (2001) A Few Remarks on the Camassa-Holm Equation. *Differential and Integral Equations*, **14**, 953-988. http://projecteuclid.org/euclid.die/1356123175

[33] Li, Y.A. and Olver, P.J. (2000) Well-Posedness and Blow-Up Solutions for an Integrable Nonlinearly Dispersive Model Wave Equation. *Journal of Differential Equations*, **162**, 27-63. https://doi.org/10.1006/jdeq.1999.3683

[34] Constantin, A. and Escher, J. (1998) Well-Posedness, Global Existence, and Blowup Phenomena for a Periodic Quasi-linear Hyperbolic Equation. *Communications on Pure and Applied Mathematics*, **51**, 475-504. https://doi.org/10.1002/(SICI)1097-0312(19980515)51:5<475::AID-CPA2>3.0.CO;2-5

[35] Molinet, L. (2004) On Well-Posedness Results for Camassa-Holm Equation on the Line: A Survey. *Journal of Nonlinear Mathematical Physics*, **11**, 521-533. https://doi.org/10.2991/jnmp.2004.11.4.8

[36] Bressan, A. and Constantin, A. (2007) Global Conservative Solutions of the Camassa-Holm Equation. *Archive for Rational Mechanics and Analysis*, **183**, 215-239. https://doi.org/10.1007/s00205-006-0010-z

[37] Constantin, A. (2000) Existence of Permanent and Breaking Waves for a Shallow Water Equation: A Geometric Approach. *Annales de l'Institut Fourier (Grenoble),*
50, 321-362. https://doi.org/10.5802/aif.1757

[38] McKean, H.P. (1998) Breakdown of a Shallow Water Equation. Asian Journal of Mathematics, 2, 867-874. https://doi.org/10.4310/AJM.1998.v2.n4.a10

[39] Xin, Z. and Zhang, P. (2000) On the Weak Solutions to a Shallow Water Equation. Communications on Pure and Applied Mathematics, 53, 1411-1433. https://doi.org/10.1002/1097-0312(200011)53:11<1411::AID-CPA4>3.0.CO;2-5

[40] Constantin, A. and Molinet, L. (2000) Global Weak Solutions for a Shallow Water Equation. Communications in Mathematical Physics, 211, 45-61. https://doi.org/10.1007/s002200050801

[41] Holden, H. and Raynaud, X. (2007) Global Conservative Multipeakon Solutions of the Camassa-Holm Equation. Journal of Hyperbolic Differential Equations, 4, 39-63. https://doi.org/10.1142/S0219530507000857

[42] Holden, H. and Raynaud, X. (2007) Global Conservative Solutions of the Camassa-Holm Equation—A Lagrangian Point of View. Communications in Partial Differential Equations, 32, 1511-1549. https://doi.org/10.1080/03605300601088674

[43] Bressan, A. and Constantin, A. (2007) Global Dissipative Solutions of the Camassa-Holm Equation. Analysis and Applications, 5, 1-27. https://doi.org/10.1142/S0219530507000857

[44] Holden, H. and Raynaud, X. (2008) Global Dissipative Multipeakon Solutions of the Camassa-Holm Equation. Communications in Partial Differential Equations, 33, 2040-2063. https://doi.org/10.1080/03605300802501715

[45] Himonas, A., Misiołek, G., Ponce, G. and Zhou, Y. (2007) Persistence Properties and Unique Continuation of Solutions of the Camassa-Holm Equation. Communications in Mathematical Physics, 271, 511-522. https://doi.org/10.1007/s00220-006-0172-4

[46] Constantin, A. (2005) Finite Propagation Speed for the Camassa-Holm Equation. Journal of Mathematical Physics, 46, Article ID: 023506, 4 p.

[47] Degasperis, A., Holm, D.D. and Hone, A.N.W. (2002) A New Integral Equation with Peakon Solutions. Theoretical and Mathematical Physics, 133, 1463-1474. https://doi.org/10.1023/A:1021186408422

[48] Lundmark, H. (2007) Formation and Dynamics of Shock Waves in the Degasperis-Procesi Equation. Journal of Nonlinear Science, 17, 169-198. https://doi.org/10.1007/s00332-006-0803-3

[49] Lundmark, H. and Szmigielski, J. (2003) Multi-Peakon Solutions of the Degasperis-Procesi Equation. Inverse Problems, 19, 1241-1245. https://doi.org/10.1088/0266-5611/19/6/001

[50] Lenells, J. (2005) Traveling Wave Solutions of the Degasperis-Procesi Equation. Journal of Mathematical Analysis and Applications, 306, 72-82. https://doi.org/10.1016/j.jmaa.2004.11.038

[51] Yin, Z. (2003) Global Existence for a New Periodic Integrable Equation. Journal of Mathematical Analysis and Applications, 283, 129-139. https://doi.org/10.1016/S0022-247X(03)00250-6

[52] Yin, Z. (2003) On the Cauchy Problem for an Integrable Equation with Peakon Solutions. Illinois Journal of Mathematics, 47, 649-666. http://projecteuclid.org/euclid.ijm/1258138186

[53] Escher, J., Liu, Y. and Yin, Z. (2006) Global Weak Solutions and Blow-Up Structure for the Degasperis-Procesi Equation. Journal of Functional Analysis, 241, 457-485. https://doi.org/10.1016/j.jfa.2006.03.022
[54] Liu, Y. and Yin, Z. (2006) Global Existence and Blow-Up Phenomena for the Degasperis-Procesi Equation. *Communications in Mathematical Physics*, **267**, 801-820. https://doi.org/10.1007/s00220-006-0082-5

[55] Yin, Z. (2004) Global Solutions to a New Integrable Equation with Peakons. *Indiana University Mathematics Journal*, **53**, 1189-1210. https://doi.org/10.1512/iiumj.2004.53.2479

[56] Escher, J., Liu, Y. and Yin, Z. (2007) Shock Waves and Blow-Up Phenomena for the Periodic Degasperis-Procesi Equation. *Indiana University Mathematics Journal*, **56**, 87-117. https://doi.org/10.1512/iiumj.2007.56.3040

[57] Escher, J. (2007) Wave Breaking and Shock Waves for a Periodic Shallow Water Equation. *Philosophical Transactions of the Royal Society Series A*, **365**, 2281-2289. https://doi.org/10.1098/rsta.2007.2008

[58] Yin, Z. (2004) Global Weak Solutions to a New Periodic Integrable Equation with Peakon Solutions. *Journal of Functional Analysis*, **212**, 182-194. https://doi.org/10.1016/j.jfa.2003.07.010

[59] Henry, D. (2005) Infinite Propagation Speed for the Degasperis-Procesi Equation. *Journal of Mathematical Analysis and Applications*, **311**, 182-194. https://doi.org/10.1016/j.jmaa.2005.03.001

[60] Mustafa, O.G. (2005) A Note on the Degasperis-Procesi Equation. *Journal of Nonlinear Mathematical Physics*, **12**, 10-14. https://doi.org/10.2991/jnmp.2005.12.1.2

[61] Coclite, G.M. and Karlsen, K.H. (2006) On the Well-Posedness of the Degasperis-Procesi Equation. *Journal of Functional Analysis*, **233**, 60-91. https://doi.org/10.1016/j.jfa.2005.07.008

[62] Zhang, S. and Yin, Z. (2010) Global Solutions and Blow-Up Phenomena for the Periodic b-Equation. *Journal of the London Mathematical Society*, **82**, 482-500. https://doi.org/10.1112/jlms/jdq044

[63] Kodzha, M. (2012) On the Blow-Up of Solutions for the b-Equation. *Acta Mathematica Universitatis Comenianae*, **1**, 117-126. http://www.iam.fmph.uniba.sk/amuc/_vol-81/_no_1/_kodzha/kodzha.pdf

[64] Niu, W.S. and Zhang, S.H. (2011) Blow-Up Phenomena and Global Existence for the Non-Uniform Weakly Dissipative b-Equation. *Journal of Mathematical Analysis and Applications*, **374**, 166-177. https://doi.org/10.1016/j.jmaa.2010.08.002

[65] Gui, G., Liu, Y. and Tian, L. (2008) Global Existence and Blow-Up Phenomena for the Peakon b-Family of Equations. *Indiana University Mathematics Journal*, **57**, 1209-1234. https://doi.org/10.1512/iiumj.2008.57.3213

[66] Zhou, Y. (2010) On Solutions to the Holm-Staley b-Family of Equations. *Nonlinearity*, **23**, 369-381. https://doi.org/10.1088/0951-7715/23/2/008

[67] Christov, O. and Hakkaev, S. (2009) On the Cauchy Problem for the Periodic b-Family of Equations and of the Non-Uniform Continuity of Degasperis-Procesi Equation. *Journal of Mathematical Analysis and Applications*, **360**, 47-56. https://doi.org/10.1016/j.jmaa.2009.06.035

[68] Tang, H., Shi, S. and Liu, Z. (2015) The Dependences on Initial Data for the b-Family Equation in Critical Besov Space. *Monatshefte für Mathematik*, **177**, 471-492. https://doi.org/10.1007/s00605-014-0673-8

[69] Lakshmanan, M. (2007) Integrable Nonlinear Wave Equations and Possible Connections to Tsunami Dynamics. In: Kundu, A., Ed., *Tsunami and Nonlinear Waves*, Springer-Verlag, Berlin Heidelberg, 31-49. https://doi.org/10.1007/978-3-540-71256-5_2

[70] Segur, H. (2008) Integrable Models of Waves in Shallow Water. In: Pinsky, M. and...
Birnir, B., Eds., *Probability, Geometry and Integrable Systems*, Cambridge University Press, Cambridge, 345-371.

[71] Constantin, A. and Johnson, R.S. (2008) Propagation of Very Long Water Waves with Vorticity, over Variable Depth, with Applications to Tsunamis. *Fluid Dynamics Research*, **40**, 175-211. https://doi.org/10.1016/j.fluid.2007.06.004

[72] Zhang, B. (1990) Some Results for Nonlinear Dispersive Wave Equations with Applications to Control. Ph.D. Thesis, The University of Wisconsin, Madison.

[73] Glass, O. (2008) Controllability and Asymptotic Stabilization of the Camassa-Holm Equation. *Journal of Differential Equations*, **245**, 1584-1615. https://doi.org/10.1016/j.jde.2008.06.016

[74] Tian, L., Shen, C. and Ding, D. (2009) Optimal Control of the Viscous Camassa-Holm Equation. *Nonlinear Analysis: Real World Applications*, **10**, 519-530. https://doi.org/10.1016/j.nonrwa.2007.10.016

[75] Tian, L. and Shen, C. (2007) Optimal Control of the Viscous Degasperis-Procesi Equation. *Journal of Mathematical Physics*, **48**, Article ID: 113513, 16 p.

[76] Shen, C., Tian, L. and Gao, A. (2010) Optimal Control of the Viscous Dullin-Gottwald-Holm Equation. *Nonlinear Analysis: Real World Applications*, **11**, 480-491. https://doi.org/10.1016/j.nonrwa.2008.11.021

[77] Sun, B. (2013) An Optimal Distributed Control Problem of the Viscous Generalized Camassa-Holm Equation. *Transactions of the Institute of Measurement and Control*, **35**, 409-416. https://doi.org/10.1177/0142331212458520

[78] Sun, B. (2012) Maximum Principle for Optimal Distributed Control of the Viscous Dullin-Gottwald-Holm Equation. *Nonlinear Analysis: Real World Applications*, **13**, 325-332. https://doi.org/10.1016/j.nonrwa.2011.07.037

[79] Shen, C. and Gao, A. (2015) Optimal Distributed Control of the Fornberg-Whitham Equation. *Nonlinear Analysis: Real World Applications*, **21**, 127-141. https://doi.org/10.1016/j.nonrwa.2014.06.005

[80] Shen, C. (2014) Optimal Control of a Class of Nonlocal Dispersive Equations. *Nonlinear Analysis*, **108**, 99-113. https://doi.org/10.1016/j.na.2014.04.023

[81] Hwang, J. (2014) Optimal Control Problem with Necessary Optimality Conditions for the Viscous Dullin-Gottwald-Holm Equation. *Abstract and Applied Analysis*, **2014**, Article ID: 623129, 14 p.

[82] Lions, J.L. (1971) Optimal Control of Systems Governed by Partial Differential Equations. Springer-Verlag, Berlin Heidelberg.

[83] Dautray, R. and Lions, J.L. (2000) Mathematical Analysis and Numerical Methods for Science and Technology, Vol. 5 Evolution Problems I. Springer-Verlag, Berlin Heidelberg.
A Second Order Characteristic Mixed Finite Element Method for Convection Diffusion Reaction Equations

Tongjun Sun

School of Mathematics, Shandong University, Jinan, China
Email: tjsun@sdu.edu.cn

Abstract

A combined approximate scheme is defined for convection-diffusion-reaction equations. This scheme is constructed by two methods. Standard mixed finite element method is used for diffusion term. A second order characteristic finite element method is presented to handle the material derivative term, that is, the time derivative term plus the convection term. The stability is proved and the $L^2$-norm error estimates are derived for both the scalar unknown variable and its flux. The scheme is of second order accuracy in time increment, symmetric, and unconditionally stable.

Keywords

Mixed Finite Element Method, Characteristic Method, Second Order Accuracy, Convection Diffusion Reaction Equations

1. Introduction

Let $\Omega$ be a bounded domain in $\mathbb{R}^d$ ($d=2,3$) with Lipschitz boundary $\Gamma = \Gamma_D \cup \Gamma_N$, where $\Gamma_D \cap \Gamma_N = \emptyset$. Let $T$ be a positive constant. In this paper, we will consider the following linear convection-diffusion-reaction equations: find $\phi: \Omega \times (0,T) \to \mathbb{R}$ such that

$$
\begin{aligned}
\frac{\partial \phi(x,t)}{\partial t} + \mathbf{v}(x,t) \cdot \nabla \phi(x,t) - \text{div} (\mathbf{A}(x) \nabla \phi(x,t)) + r(x) \phi(x,t) &= f(x,t), & & \text{in } \Omega \times (0,T), \\
\phi(x,t) &= 0, & & \text{on } \Gamma_D \times (0,T), \\
\alpha \phi(x,t) + \mathbf{A}(x) \nabla \phi(x,t) \cdot \mathbf{n}(x) &= g(x,t), & & \text{on } \Gamma_N \times (0,T), \\
\phi(x,0) &= \phi^0(x), & & \text{in } \Omega.
\end{aligned}
$$

(1)
Here, \( \mathbf{v} : \overline{\Omega} \times [0,T] \rightarrow \mathbb{R}^d \) is the convection vector field; \( r : \overline{\Omega} \times [0,T] \rightarrow \mathbb{R} \) is the reaction function; \( f : \overline{\Omega} \times [0,T] \rightarrow \mathbb{R} \) and \( g : \Gamma_r \times [0,T] \rightarrow \mathbb{R} \) are given scalar functions; \( \alpha > 0 \) is a constant and \( \mathbf{n} \) is the outward unit normal vector to \( \Gamma \); \( A : \overline{\Omega} \rightarrow \mathcal{S}_d \) denotes the diffusion matrix function, where \( \mathcal{S}_d \) is the space of symmetric \( d \times d \) matrices, such that

\[
a_{ij} \| \xi \|^2 \leq \sum_{i,j=1}^{d} A_{ij} \xi_i \xi_j \leq a_{i} \| \xi \|^2, \quad \forall \xi \in \mathbb{R}^d,
\]

where \( a_{ij}, a_{i} \) are positive constants.

Typically, these Equations (1) express the general mathematical model incorporating different types of transport phenomena in engineering and applied sciences, such as the dispersal of a pollutant through a moving viscous medium (e.g., a river or the atmosphere, [1]), currents in semiconductor devices [2], and airflow past an aerofoil (see [3], for example). When the diffusive term is smaller than the convective one, these equations are the so-called convection dominated problems (see [4]).

In many practical convection-diffusion processes, convection essentially dominates diffusion (e.g., in some financial models [5]), and although the governing differential equation is parabolic, it displays several characteristics of hyperbolic problems. When applied to these problems, standard finite element and finite difference methods usually exhibit some combination of nonphysical oscillation and excessive numerical dispersion [6] [7]. It is therefore logical to design numerical procedures that incorporate the parabolic/hyperbolic nature of these problems. One such method is the modified method of characteristics (MMOC) which was first formulated for a scalar parabolic equation by J. Douglas and T. F. Russell in [8] and then extended by Russell [9] to nonlinear coupled systems in two and three spatial dimensions. Similar schemes had been defined by Pironneau [10] for the incompressible Navier-Stokes equations and by Süli [11] and Morton, Priestley, and Süli [12] for first-order hyperbolic equations, with the latter technique being referred to as the Euler characteristic Galerkin method. The intent of the method is to obtain accurate approximations to convection-dominated problems. Basically, in the modified method of characteristics, the time derivative and the convection term are combined as a directional derivative. In other words, the procedure involves time stepping along the characteristics, allowing us to use large, accurate time steps.

Mixed finite element method has been proven to be an effective numerical method for solving fluid problems. It has an advantage to approximate the unknown variable and its diffusive flux simultaneously. There are many research articles on this method ([13] [14] [15] [16]). An algorithm combining the mixed finite-element method and the modified method of characteristics was first applied to the miscible displacement problem in porous media by Ewing, Russell, and Wheeler [17]. Then, the scheme had been extended by Wheeler and Dawson [18] to advection-diffusion-reaction problems. Numerical results have verified that large, accurate time steps are possible, and sharp fronts have been resolved.
Arbogast and Wheeler [19] defined a characteristics-mixed method to approximate the solution of an advection-dominated transport problem. It used a characteristic approximation that is similar to that of MMOC-Galerkin method to handle advection in time and a lowest order mixed finite element spatial approximation for diffusion term. Piecewise constants were in the space of test function, so mass is conserved element-by-element. It was proved finally that the method was optimally convergent to order 1 in time and at least suboptimally convergent to order 3/2 in space. In [20], we have considered a combined numerical approximation for incompressible miscible displacement in porous media. Standard mixed finite element was used for Darcy velocity equation and a characteristics-mixed finite element method was presented for approximating the concentration equation. Characteristic approximation was applied to handle the convection term, and a lowest order mixed finite element spatial approximation was adopted to deal with the diffusion term. Thus, the scalar unknown concentration and the diffusive flux can be approximated simultaneously. This approximation conserves mass globally. The optimal $L^2$-norm error estimates were derived. Then, we extended this method to the slightly compressible miscible displacement problem in [21].

It should be pointed out that the works mentioned above which involved the characteristic method all gave one order accuracy in time increment $\Delta t$. That is to say, the first order characteristic method in time was analyzed. As for higher order characteristic method in time, Rui and Tabata [22] used the second order Runge-Kutta method to approximate the material derivative term for convection-diffusion problems. The scheme presented was of second order accuracy in time increment $\Delta t$, symmetric and unconditionally stable. Optimal error estimates were proved in the framework of $L^2$-theory. Numerical analyses of convection-diffusion-reaction problems with higher order characteristic/finite elements were analyzed in [23] [24], which extended the work [22]. The $L^\infty (L^2)$ error estimates of second order in time increment $\Delta t$ were obtained.

The goal of this paper is to present a second order characteristic mixed finite element method in time increment to handle the material derivative term of (1). It is organized as follows. In Section 2, we formulate an approximate scheme that combines the second order characteristic finite element method for the material derivative term and mixed finite element method for the diffusion term. In Section 3, we prove the stability of the combined approximate scheme. In Section 4, we derive the $L^2$-norm error estimates for both the unknown scalar variable and its flux. The scheme is of second order accuracy in time increment, symmetric, and unconditionally stable.

2. Formulation of the Method

In this paper, we adopt notations and norms of usual Sobolev spaces. Moreover, we adopt some notations for the functional spaces involved, which were used in...
For a Banach space $X$ and a positive integer $m$, spaces $C^{m}([0,T],X)$ and $H^{m}((0,T),X)$ are abbreviated as $C^{m}(X)$ and $H^{m}(X)$, respectively, and endowed with norms

$$
\|\phi\|_{C^{m}(X)} := \max_{i \in [0,T]} \left\{ \max_{j \in \mathbb{N}} \|\phi^{(j)}(t)\|_{X} \right\}, \quad \|\phi\|_{H^{m}(X)} := \left( \int_{0}^{T} \sum_{j=0}^{m} \|\phi^{(j)}(t)\|_{X}^{2} \, dt \right)^{1/2},
$$

where $\phi^{(j)}$ denotes the $j$-th derivative of $\phi$ with respect to time. The Banach space $Z^{m}$ is defined by

$$
Z^{m} = \left\{ f \in C^{j}(H^{m-j}(\Omega)) : j \in \mathbb{N}, \right\},
$$

equipped with the norm $\|\phi\|_{Z^{m}} := \max_{j \in \mathbb{N}} \|\phi^{(j)}_{(H^{m-j})}\|_{0 \leq j \leq m}$. Similar spaces and norms are considered for the boundary sets $\Gamma_{R}$ and $\Gamma_{D}$. Denote by $H^{1}_{\delta}(\Omega)$ the closed subspace of $H^{1}(\Omega)$ defined by

$$
H^{1}_{\delta}(\Omega) = \left\{ \phi \in H^{1}(\Omega), \phi\big|_{\delta} = 0 \right\},
$$

and [25]

$$
H(\text{div};\Omega) = \left\{ v \in \left( L^{2}(\Omega) \right)^{d}, \text{div} v \in L^{2}(\Omega) \right\}.
$$

### 2.1. The Characteristic Lines

Now, we define the characteristic lines associated with vector field $v$ and recall some classical properties satisfied by them. Thus, for given $(x,t) \in \Omega \times [0,T]$, the characteristic line through $(x,t)$ is the vector function $X_{x}(x,t;\cdot)$ solving the initial value problem

$$
\frac{\partial X_{x}}{\partial \tau}(x,t;\tau) = v(X_{x}(x,t;\tau),\tau), \quad X_{x}(x,t;\tau) = x. \quad (3)
$$

Next, assuming they exist, we denote by $F_{x}$ (respectively, by $L$) the gradient of $X_{x}$ (respectively, of $v$) with respect to the space variable $x$, i.e.,

$$
(F_{x})_{x}(x,t;\tau) := \frac{\partial (X_{x})}{\partial x}(x,t;\tau), \quad L_{x}(x,t) := \frac{\partial v}{\partial x}(x,t). \quad (4)
$$

We adopted some propositions and lemmas from [23].

**Proposition 1.** If $v \in C^{n}(\Omega)$, then $X_{x} \in C^{n}(\Omega \times [0,T])$ and it is $C^{n}$ with respect to the $x$ variable.

Proposition 2. If $v \in C^{n}(\Omega)$, then

$$
\|F_{x}(x,t;\tau)\|_{L^{2}(\Omega \times [0,T])} \leq e^{M_{n-1}(\cdot)}(\delta), \quad \forall x \in \Omega, t, \tau \in [0,T]. \quad (6)
$$
Proposition 3. If \( \mathbf{v} \in C^0 \left( C^2 \left( \overline{\Omega} \right) \right) \cap C^1 \left( C^1 \left( \overline{\Omega} \right) \right) \), then \( F_e \) satisfies the Taylor expansion

\[
F_e(x,t;s) = I + (s-t)L(x,t) + \int_t^s (r-t)\nabla \left( \frac{\partial F}{\partial t} + LV \right) \left( X_e(x,t;r), r \right) F_e(x,t;r) \, dr,
\]

and its inverse, \( F_e^{-1} \), satisfies the Liouville’s theorem

\[
F_e^{-1}(x,t;s) = I + (t-s)L \left( X_e(x,t;s), s \right) + \int_t^s (r-t)\nabla \left( \frac{\partial F}{\partial t} + LV \right) \left( X_e(x,t;r), r \right) F_e(x,t;s), r \, dr.
\]

Proposition 4. If \( \mathbf{v} \in C^0 \left( C^2 \left( \overline{\Omega} \right) \right) \), then

\[
det F_e^{-1}(x,t;r) \leq e^{H||c| |n| \|f\|_{L^1}} \quad \forall x \in \Omega, t, r \in [0,T].
\]

Proposition 5. If \( \mathbf{v} \in C^0 \left( C^2 \left( \overline{\Omega} \right) \right) \cap C^1 \left( C^1 \left( \overline{\Omega} \right) \right) \), then \( det F_e^{-1} \) satisfies

\[
det F_e^{-1}(x,t;s) = 1 - (s-t)\text{div}(x,t) + \int_t^s (r-t)\frac{\partial^2}{\partial r^2} \left( det F_e^{-1} \right)(x,t;r) \, dr.
\]

2.2. Variational Formulation

From the definition of the characteristic lines and by using the chain rule, it follows that

\[
\frac{d\phi}{dr} \left( X_e(x,t;r), r \right) = \frac{\partial \phi}{\partial t} \left( X_e(x,t;r), r \right) + \mathbf{v} \left( X_e(x,t;r), r \right) \cdot \nabla \phi \left( X_e(x,t;r), r \right).
\]

By introducing the flux \( \sigma = \mathbf{A} \nabla \phi \) and using (12), we rewrite equation (1.1) at point \( X_e(x,t;r) \) and time \( r \) as follows

\[
\frac{d\phi}{dr} \left( X_e(x,t;r), r \right) - \text{div}\sigma \left( X_e(x,t;r), r \right) = f \left( X_e(x,t;r), r \right),
\]

Before giving a weak formulation of (13), we adopted a lemma from [23], which can be considered as a Green’s formula.

Lemma 6. Let \( X : \Omega \to \overline{X(\Omega)}, X \in C^2 \left( \overline{\Omega} \right) \) be an invertible vector valued function. Let \( F = \nabla X \) and assume that \( F^{-1} \in C^1 \left( \overline{\Omega} \right) \). Then
\[
\int_{\Omega} \nabla \cdot \mathbf{w}(X(x)) \psi(x) \, dx = \int_{\Omega} F^{-1}(x) \mathbf{n}(x) \cdot \mathbf{w}(X(x)) \psi(x) \, d\Gamma - \int_{\Omega} F^{-1}(x) \mathbf{w}(X(x)) \cdot \nabla \psi(x) \, dx - \int_{\Omega} \nabla \cdot \left[ F^{-1}(x) \mathbf{w}(X(x)) \psi(x) \right] \, dx,
\]

with \( \mathbf{w} \in H^1(X(\Omega)) \) a vector valued function and \( \psi \in H^1(\Omega) \) a scalar function.

Now, we can multiply (13) by test functions \( \psi \in H^1(\Omega) \) and \( \chi \in H(div;\Omega) \), integrate in \( \Omega \) respectively, and apply the usual Green’s formula and (14) with \( X(x) = X_e(x,t;\tau) \), obtaining

\[
\left\{ \begin{array}{l}
\int_{\partial \Omega} \frac{\partial}{\partial \tau} \left( X_e(x,t;\tau), \tau \right) \psi(x) \, dx + \int_{\Omega} F^{-1}(x,t;\tau) \sigma(X_e(x,t;\tau),\tau) \cdot \nabla \psi(x) \, dx \\
+ \int_{\partial \Omega} \nabla F^{-1}(x,t;\tau) \cdot \sigma(X_e(x,t;\tau),\tau) \psi(x) \, dx \\
+ \int_{\Omega} \nabla \cdot \left[ F^{-1}(x,t;\tau) \mathbf{n}(x) \cdot \sigma(X_e(x,t;\tau),\tau) \psi(x) \right] \, d\Gamma \\
- \int_{\Omega} \left[ \mathbf{n}(x) \cdot \sigma(X_e(x,t;\tau),\tau) \psi(x) \right] \, d\Gamma \\
= \int_{\Omega} \mathbf{f}(X_e(x,t;\tau),\tau) \psi(x) \, dx, \\
(\sigma(X_e(x,t;\tau),\tau),\chi) = (A(X_e(x,t;\tau)) \nabla \phi(X_e(x,t;\tau),\tau),\chi).
\end{array} \right.
\]

**Lemma 7.** [23] Let \( X: \overline{\Omega} \rightarrow \overline{X(\Omega)}, X \in C^2(\overline{\Omega}) \) be an invertible vector valued function satisfying \( X(x) = x, \forall x \in \Gamma \). Let \( F = \nabla X \) and assume that \( F^{-1} \in C^1(\overline{\Omega}) \). Then

\[
\int_{\partial \Omega} F^{-1}(x) \mathbf{n}(x) \cdot \mathbf{w}(x) \psi(x) \, d\Gamma = \int_{\partial \Omega} \mathbf{n}(x) \cdot \mathbf{w}(x) \psi(x) \, d\Gamma,
\]

with \( \mathbf{w} \in H^1(X(\Omega)) \) and \( \psi \in H^1(\Omega) \), where \( \mathbf{n} \) is the outward unit normal vector to \( \Gamma \).

Now, replacing in (15) formula (16) with \( X(x) = X_e(x,t;\tau) \), and replacing the Robin condition, we have

\[
\left\{ \begin{array}{l}
\int_{\partial \Omega} \frac{\partial}{\partial \tau} \left( X_e(x,t;\tau), \tau \right) \psi(x) \, dx + \int_{\Omega} F^{-1}(x,t;\tau) \sigma(X_e(x,t;\tau),\tau) \cdot \nabla \psi(x) \, dx \\
+ \int_{\partial \Omega} \nabla F^{-1}(x,t;\tau) \cdot \sigma(X_e(x,t;\tau),\tau) \psi(x) \, dx \\
+ \int_{\Omega} \nabla \cdot \left[ F^{-1}(x,t;\tau) \mathbf{n}(x) \cdot \sigma(X_e(x,t;\tau),\tau) \psi(x) \right] \, d\Gamma \\
- \int_{\Omega} \left[ \mathbf{n}(x) \cdot \sigma(X_e(x,t;\tau),\tau) \psi(x) \right] \, d\Gamma \\
= \int_{\Omega} \mathbf{f}(X_e(x,t;\tau),\tau) \psi(x) \, dx + \int_{\partial \Omega} g(X_e(x,t;\tau),\tau) \psi(x) \, d\Gamma, \\
(\sigma(X_e(x,t;\tau),\tau),\chi) = (A(X_e(x,t;\tau)) \nabla \phi(X_e(x,t;\tau),\tau),\chi).
\end{array} \right.
\]

### 2.3. The Combined Approximate Scheme

We now present our time-stepping procedure for (17). Let \( N \) be the number of time steps, \( \Delta t = T/N \) be the time step, and \( t^n = n\Delta t \) for \( n = 0,1,2,3/2,\cdots,N \). We will use the notation \( \psi^n(x) := \psi(x,t^n) \) for a function. Moreover, for \( n = 0,1,\cdots \), we define

\[
X_e^n(x) := X_e(x,t^{n+1};t^n), \quad F_e^n(x) := X_e(x,t^{n+1};t^n), \\
X_e^{n-1}(x) := X_e(x,t^{n+1};t^{n-\frac{1}{2}}), \quad F_e^{n-1}(x) := X_e(x,t^{n+1};t^{n-\frac{1}{2}}),
\]

(18)
By fixing \( t = \tau^{n+1}, n = 0, 1, \cdots, N-1 \) in (17) and using a Crank-Nicolson method \[26\] with respect to \( \tau \). Thus, from (18) we have

\[
\begin{align*}
\int_{\Omega} \frac{\phi^{n+1}(x) - \phi^n(x)}{\Delta t} \psi(x) \, dx &+ \frac{1}{2} \int_{\Omega} \sigma^{n+1}(x) \cdot \nabla \psi(x) \, dx \\
&+ \frac{1}{2} \int_{\Omega} \left( F^n \right)^{-1}(x) \sigma^n(X^n(x)) \cdot \nabla \psi(x) \, dx \\
&+ \frac{1}{2} \int_{\Omega} \text{div} \left( F^n \right)^{-T}(x) \cdot \sigma^n(X^n(x)) \psi(x) \, dx \\
&+ \frac{1}{2} \int_{\Delta x} \left( r(x) \phi^{n+1}(x) + r(x) \psi(x) \right) \, d\Gamma \\
&= \frac{1}{2} \int_{\Omega} \left( f^{n+1}(x) + f^n(X^n(x)) \right) \psi(x) \, dx \\
&+ \frac{1}{2} \int_{\Delta x} \left( g^{n+1}(x) + g^n \left( 1 + \Delta t \text{div} v^{n+1} \right) \right) \psi(x) \, d\Gamma,
\end{align*}
\]

(19)

By using (8) and (11), we see that

\[
\begin{align*}
\left( F^n \right)^{-1}(x) &= I(x) + \Delta t \left( X^n(x) \right) + O\left( \left( \Delta t \right)^2 \right), \\
\text{det} \left( F^n \right)^{-1}(x) &= 1 + \Delta t \text{div} v^{n+1}(x) + O\left( \left( \Delta t \right)^2 \right), \\
\text{div} \left( F^n \right)^{-T}(x) &= \Delta t \text{div} v^n \left( X^n(x) \right) + O\left( \left( \Delta t \right)^2 \right).
\end{align*}
\]

(20)

Taking (20) into (19), we can obtain

\[
\begin{align*}
\int_{\Omega} \frac{\phi^{n+1} - \phi^n(X^n(x))}{\Delta t} \psi \, dx &+ \frac{1}{2} \int_{\Omega} \sigma^{n+1} + \sigma^n(X^n(x)) \cdot \nabla \psi \, dx \\
&+ \frac{\Delta t}{2} \int_{\Omega} L^n(X^n(x)) \sigma^n(X^n(x)) \cdot \nabla \psi \, dx \\
&+ \frac{\Delta t}{2} \int_{\Omega} \nabla \text{div} v^n \left( X^n(x) \right) \cdot \sigma^n(X^n(x)) \psi \, dx \\
&+ \int_{\Omega} \frac{r \phi^{n+1} + r(X^n(x)) \phi^n(X^n(x))}{2} \psi \, dx \\
&+ \int_{\Delta x} \frac{\phi^{n+1} + \phi^n(X^n(x))}{2} \psi \, d\Gamma \\
&= \int_{\Omega} \frac{f^{n+1} + f^n(X^n(x))}{2} \psi \, dx + \int_{\Delta x} \frac{g^{n+1} + g^n \left( 1 + \Delta t \text{div} v^{n+1} \right)}{2} \psi \, d\Gamma,
\end{align*}
\]

(21)

\[
\begin{align*}
\left( A^{-1} \sigma^{n+1}, \chi \right) &= \left( \nabla \phi^{n+1}, \chi \right).
\end{align*}
\]

We propose two explicit numerical schemes to approximate \( X^n(x) \):

\[
\begin{align*}
X^n_E(x) &= x - \Delta t \text{div} v^{n+1}(x) \quad \text{(Euler scheme),} \\
X^{n+1}_{\text{RK}}(x) &= x - \Delta t \left( \frac{1}{2} \text{div} v^{n+1} \right) \left( x - \frac{\Delta t}{2} v^{n+1} \right) \quad \text{(Runge-Kutta scheme).}
\end{align*}
\]

(22)

A similar notation to the one in Section 2.2 is used for the Jacobian of \( X^n_E \),
namely, 
\[ F^n_k(x) := \nabla X^n_k(x) = I(x) - \Delta L^{n+1}(x). \]

Now, we restate three lemmas concerning properties of the characteristic line approximations. For this, we require the time step to be bounded and the velocity to satisfy the following assumption.

**Claim 1.** The velocity field \( v \in C^0 \left( W^{1,\infty}(\Omega) \right) \) and satisfies \( v = 0 \) on \( \Gamma \).

**Lemma 8.** \[ 23 \] Under Claim 1, if \( \| v \|_{C^0(\Omega)} \Delta t < 1/2 \), we can see that
\[
X^n_k(\Omega) = X^n_k(\Omega).
\]

**Lemma 9.** \[ 23 \] Under Claim 1, if \( \| v \|_{C^0(\Omega)} \Delta t < 1/2 \), we have
\[
\left( F^n_k \right)^{-1}(x) = I + \Delta L^{n+1}(x) + (\Delta t)^2 \left( L^{n+1}(x) \right)^2 + O \left( (\Delta t)^3 \right).
\]

**Corollary 1.** Under the assumptions of Lemma 9, \( \forall x \in \Omega \), we have
\[
\left| \left| \det(F^n_k)^{-1}(x) \right| \right| \leq 1 + c \Delta t \| v \|^2, \text{ for } n = 0, \cdots, N \text{ and } i = E, RK,
\]
where \( v \circ X^n_k = v \left( X^n_k \right) \).

Thus, in the case where the characteristic lines and their gradients are not explicitly known, we propose the following time approximation of (21)
\[
\begin{align*}
\int_{\Omega} \phi^{n+1} - \phi^n \circ X^n \frac{\Delta t}{2} \psi dx + \int_{\Omega} \sigma^{n+1} + \sigma^n \circ X^n \nabla \psi dx \\
+ \frac{\Delta t}{2} \int_{\Omega} \left( L^{n+1} \sigma^n \right) \circ X^n \cdot \nabla \psi dx + \frac{\Delta t}{2} \int_{\Omega} \left( \nabla \text{div} \sigma^n \cdot \sigma^n \right) \circ X^n \psi dx \\
+ \int_{\Gamma} \left( \phi^{n+1} + r \phi^n \circ X^n \right) \psi d\Gamma + \int_{\Gamma} \left( \phi^{n+1} + \phi^n \left( 1 + \Delta t \text{div} \sigma^{n+1} \right) \right) \psi d\Gamma \\
= \int_{\Gamma} f^{n+1} + f^n \circ X^n \psi d\Gamma + \int_{\Gamma} g^{n+1} + g^n \left( 1 + \Delta t \text{div} \sigma^{n+1} \right) \psi d\Gamma,
\end{align*}
\]

The time difference approximation (27) will be combined with a standard Galerkin finite element and mixed finite element in the space for \( \phi(x) \) and \( \sigma(x) \), respectively \[ 27 \] \[ 28 \]. We discrete \( \phi(x) \) in space on a quasi-uniform finite element mesh \( T_h \) of \( \Omega \) with maximal element diameter \( h \). For \( \sigma(x) \), we denote as \( h > 0 \) and \( T_h \) similarly. Let \( V_h^k \times W_h^k \subset H^1(\Omega) \times H \left( \text{div}; \Omega \right) \) be finite element spaces with index \( k \) and \( l \), respectively.

We define a bilinear form \( A^{n+1/2} \) on \( V_h^k \times W_h^k \times V_h^k \) and a linear form \( F^{n+1/2} \) on \( V_h^k \) for \( n = 0, \cdots, N - 1 \) by
\[
\begin{align*}
A^{n+1/2}_h (\phi_h, \sigma_h, \psi_h) &= \left( \frac{\phi_h^{n+1} - \phi_h^n \circ X^n_E}{\Delta t} \cdot \psi_h^n \right) + \left( \frac{\sigma_h^{n+1} + \sigma_h^n \circ X^n_E}{2} \cdot \nabla \psi_h^n \right) \\
&+ \frac{\Delta t}{2} \left( (L^\star \sigma_h^n) \circ X^n_E, \nabla \psi_h^n \right) + \frac{\Delta t}{2} \left( \nabla \text{div} \nabla \cdot \sigma_h^n \circ X^n_E, \psi_h^n \right) \\
&+ \left( \frac{r \phi_h^{n+1} + (r \phi_h^n) \circ X^n_E}{2}, \psi_h^n \right) + \alpha \left( \frac{\phi_h^{n+1} + \phi_h^n (1 + \Delta t \text{div} \nabla^{n+1})}{2}, \psi_h^n \right), \\
F^{n+1/2}_h (\psi_h) &= \left( \frac{f^{n+1} + f^n \circ X^n_E}{2}, \psi_h^n \right) + \left( \frac{g^{n+1} + g^n (1 + \Delta t \text{div} \nabla^{n+1})}{2}, \psi_h^n \right),
\end{align*}
\]

where $\phi_h \in V^h, \sigma_h \in W^l_h, \psi_h \in V^h$.

Then, the fully discrete scheme reads: Given $\phi_0^n \in V^h$, find $\{\phi_h^n, \sigma_h^n\}_{n=1}^N \in V^h \times W^l_h$ such that

\[
\begin{align*}
A^{n+1/2}_h (\phi_h, \sigma_h, \psi_h) &= F^{n+1/2}_h (\psi_h), \quad \forall \psi_h \in V^h, \\
(A^{-1} \sigma_h^{n+1}, \chi_h) &= (\nabla \phi_h^{n+1}, \chi_h), \quad \forall \chi_h \in W^l_h.
\end{align*}
\]

Throughout the analysis, $K$ will denote a generic positive constant, independent of $h, h_n, \Delta t$. Similarly, $\varepsilon$ will denote a generic small positive constant.

3. Stability of the Approximate Scheme

In this section, we derive the stability of the approximate scheme (29). In order to develop the stability, some assumptions on the different terms of (1) are required.

Claim 2. The velocity field $v \in C^0 (W^{2,\infty}(\Omega))$ and satisfies $v = 0$ on $\Gamma$.

Remark. Throughout this paper $c_1$ denotes the maximum between the positive constant appearing in Lemma 10 and the norm of the velocity in $C^0 (W^{2,\infty}(\Omega))$.

Claim 3. The diffusion matrix coefficients, $A_{ij}$, belong to $W^{1,\infty}(\Omega)$. Moreover, $A$ is a positive definite symmetric $d \times d$ matrix and there exists a strictly positive constant $\delta$ which is a uniform lower bound for the eigenvalues of $A^{-1}$.

As a consequence of Claim 3, there exists a unique positive definite symmetric $d \times d$ matrix function $C$, such that $A^{-1} = C^2$ and $C_{ij} \in W^{1,\infty}(\Omega)$. Let us introduce the constant $c_2 := \max_{i,j} \left\| C_{ij} \right\|_{W^{1,\infty}(\Omega)}$. Clearly, Claim 3, we have

\[
\delta \left\| v \right\|_w^2 \leq (A^{-1}v, w) = \left\| Cw \right\|_C^2 \leq c_2 \left\| w \right\|_E^2, \quad \forall w \in \mathbb{R}^d.
\]

Claim 4. The reaction function, $r \in W^{1,\infty}(\Omega)$, satisfies $0 < \gamma \leq r(x)$ in $\Omega$, where $\gamma$ is a constant.

Under the previous claims, let $c_3 := \left\| C \right\|_{W^{1,\infty}(\Omega)}$.

Claim 5. The source function $f \in C^0 \left( L^2(\Omega) \right)$. In Robin boundary condition, $g \in C^0 \left( L^2(\Gamma_g) \right)$ and $\alpha > 0$.

For a given series of functions $\{\psi^n\}_{n=0}^N$, we define the following norms
Similar definitions are considered for functional spaces $L^n(\Gamma_R)$ and $L^n(\Gamma_s)$ associated to the Robin boundary condition. Moreover, we define

$$\|\varphi\|_{L^n(\Gamma_R)} = \left\{ \Delta t \sum_{n=0}^{N+1} \left\| \varphi^n \right\|_{L^n(\Gamma_R)}^2 \right\}^{1/2}.$$

**Lemma 11.** Let the above Claims 2 - 5 and $c_i \Delta t \leq 1/2$ be assumed. If $\{\phi_n, \sigma_n\}_{n=1}^N$ be the solution of (29). Then it holds that

$$A_n^{\alpha+1}(\phi_n, \sigma_n, \phi_n^{\alpha+1}) \geq \partial_t \left( \frac{1}{2} \left\| \phi_n \right\|^2 + \frac{\Delta t}{4} \left\| \sigma_n \right\|^2 + \frac{\Delta t}{4} \left\| \sqrt{\sigma_n} \right\|^2 + \frac{\Delta t}{4} \left\| \phi_n \right\|_{L^n(\Gamma_R)}^2 \right)$$

$$+ \frac{1}{2\Delta t} \left( \phi_n^{\alpha+1} - \phi_n^{\alpha+1} \circ X_n^{l}\|_{H}^2 + \frac{\delta}{4} \sigma_n^{\alpha+1} + \frac{\sigma_n^{\alpha+1}}{4} X_n^{l}\|_{L^n(\Gamma_R)}^2 \right)$$

$$+ \frac{1}{4} \left( \left\| \phi_n \right\| + \left\| \phi_n^{\alpha+1} \right\| + \left\| \sigma_n \right\| \right) + c\Delta t \left\| \phi_n \right\|_{L^n(\Gamma_R)}^2 + c\Delta t \left\| \nabla \phi_n \right\|_{L^n(\Gamma_R)}^2$$

$$+ c\Delta t \left( \left\| \phi_n \right\| + \left\| \phi_n^{\alpha+1} \right\| + \alpha \left\| \phi_n \right\|_{L^n(\Gamma_R)}^2 \right),$$

the constant $c$ is given by

$$c = \max \left\{ 1, \frac{\delta}{4} - 2c_2, c_3, c_4 \right\}.$$

**Proof.** Substituting $\psi_n = \phi_n^{\alpha+1}$ into (28), we have

$$A_n^{\alpha+1/2}(\phi_n, \sigma_n, \phi_n^{\alpha+1})$$

$$= \left( \phi_n^{\alpha+1} - \phi_n^{\alpha+1} \circ X_n^{l}, \phi_n^{\alpha+1} \right) + \left( \sigma_n^{\alpha+1} + \sigma_n^{\alpha+1} \circ X_n^{l}, A \Phi \sigma_n^{\alpha+1} \right)$$

$$+ \frac{\Delta t}{2} \left( \left( L \sigma_n \right) \circ X_n^{l}, V \Phi \sigma_n^{\alpha+1} \right) + \frac{\Delta t}{2} \left( \left( \nabla \Phi n \cdot \sigma_n \right) \circ X_n^{l}, \Phi \sigma_n^{\alpha+1} \right)$$

$$+ \left( \frac{r \Phi \sigma_n^{\alpha+1} + \frac{r \Phi \sigma_n^{\alpha+1}}{2} \circ X_n^{l}, \Phi \sigma_n^{\alpha+1} \right) + \alpha \left( \Phi \sigma_n^{\alpha+1} + \Phi \sigma_n^{\alpha+1} \circ X_n^{l}, \Phi \sigma_n^{\alpha+1} \right)$$

$$= I_1 + I_2 + I_3 + I_4 + I_5 + I_6.$$

Lemma 4.1 in [23] implies that

$$I_1 = \frac{1}{2\Delta t} \left( \left\| \Phi \sigma_n^{\alpha+1} \right\| - \left\| \Phi \sigma_n^{\alpha+1} \circ X_n^{l} \right\| \right) + \frac{1}{2\Delta t} \left\| \Phi \sigma_n^{\alpha+1} - \Phi \sigma_n^{\alpha+1} \circ X_n^{l} \right\| \right]$$

$$\geq \partial_t \left( \frac{1}{2} \left\| \Phi \sigma_n \right\|^2 - \frac{c_1}{2} \left\| \Phi \sigma_n \right\|^2 + \frac{1}{2\Delta t} \left\| \Phi \sigma_n^{\alpha+1} - \Phi \sigma_n^{\alpha+1} \circ X_n^{l} \right\|^2 \right),$$

and
\[ I_2 = \frac{1}{4} \left( \left\| \sigma_h^{n+1} \right\|_{1, \Gamma_h}^2 - \left\| \sigma_h \circ X_n \right\|_{1, \Gamma_h}^2 \right) + \frac{1}{4} \left\| \sigma_h^{n+1} + \sigma_h \circ X_n \right\|_{1, \Gamma_h}^2 \]

\[ \geq \frac{1}{4} \left( \delta \left\| \sigma_h^{n+1} \right\|^2 - (1 + c_1 \Delta t) c_2 \left\| \sigma_h \right\|^2 \right) + \frac{1}{4} \left\| \sigma_h^{n+1} + \sigma_h \circ X_n \right\|_{1, \Gamma_h}^2 \]

\[ \geq \partial_i \left( \frac{\Delta t}{4} \left\| \sigma_h \right\|^2 \right) + \left( \frac{\delta}{4} - 2c_2 \right) \left\| \sigma_h \right\|^2 + \frac{1}{4} \left\| \sigma_h^{n+1} + \sigma_h \circ X_n \right\|_{1, \Gamma_h}^2. \] (34)

Next, by using \( c_1 \Delta t < 1 \), we obtain

\[ I_3 \leq \frac{\Delta t}{2} \sqrt{1 + c_1 \Delta t} \left\| \nabla \sigma_h \right\| \left\| \nabla \phi_h^{n+1} \right\| \leq \frac{c_1 \Delta t}{2} \left( \left\| \sigma_h \right\|^2 + \left\| \nabla \phi_h^{n+1} \right\|^2 \right). \]

Then when \( I_3 \geq 0 \) and \( I_3 < 0 \), we have

\[ I_3 \geq -\frac{c_1 \Delta t}{2} \left( \left\| \sigma_h \right\|^2 + \left\| \nabla \phi_h^{n+1} \right\|^2 \right). \] (35)

Similarly, for \( I_4 \) we obtain the estimate

\[ I_4 \geq -\frac{c_1 \Delta t}{2} \left( \left\| \sigma_h \right\|^2 - \frac{1}{4} \left\| \phi_h^{n+1} \right\|^2 \right). \] (36)

Analogous computations to term \( I_2 \) give

\[ I_5 \geq \partial_i \left( \frac{\Delta t}{4} \left\| \sqrt{r} \phi_h \right\|^2 \right) + \frac{\Delta t}{4} \left\| \sqrt{r} \phi_h^{n+1} \right\|^2 + \left( \frac{\delta}{4} - 2c_2 \right) \left\| \sqrt{r} \phi_h \right\|^2 \]

\[ \geq -\max \{c_1, c_2, c_3, \gamma\} \frac{\Delta t}{2} \left( \left\| \sqrt{r} \phi_h \right\|^2 + \left\| \sqrt{r} \phi_h^{n+1} \right\|^2 \right). \] (37)

For the boundary integral term \( I_6 \), we first use some properties of the inner product in the space \( L^2(\Gamma_h) \) and the inequality \( (1 + c_1 \Delta t)^2 \leq 1 + 3c_1 \Delta t \) to get the estimate

\[ \left\| \psi \left( 1 + \Delta t \text{div} \phi^{n+1} \right) \right\|_{L^2(\Gamma)} \leq (1 + c_1 \Delta t)^2 \left\| \psi \right\|_{L^2(\Gamma)} \leq (1 + 3c_1 \Delta t) \left\| \psi \right\|_{L^2(\Gamma)} \]

for \( \psi \in L^2(\Gamma_h) \). Thus, we obtain

\[ I_6 \geq \partial_i \left( \frac{\alpha \Delta t}{4} \left\| \phi_h \right\|^2_{L^2(\Gamma)} \right) + \frac{\alpha \Delta t}{4} \left\| \phi_h^{n+1} \right\|^2_{L^2(\Gamma)} + \psi \left( 1 + \Delta t \text{div} \phi^{n+1} \right) \left\| \phi_h \right\|^2_{L^2(\Gamma)} \]

\[ \geq \frac{3}{4} c_1 \Delta t \left\| \phi_h \right\|^2_{L^2(\Gamma)}. \] (38)

Then, by summing up from (33) to (38), inequality (31) follows.

By using Lemma 11 and following the arguments to Lemmas 5.6, 5.7 and Theorem 5.8 in [23], we can get the following stability theorem:

**Theorem 12 (Stability Theorem)** Let the above Claims 2 - 5 assumed, and \( \left\{ \phi_h, \sigma_h \right\}_{h \in \mathcal{H}} \) be the solution of (29) subject to the initial value \( \phi_h^0 \). Then there exist two positive constants \( c \) and \( \delta = d \left( c, c_2, c_3, \gamma \right) \), such that, for \( \Delta t < d \), we have

\[ \frac{1}{\sqrt{2}} \left\| \phi_h \right\|^2_{L^2(\Gamma)} + \frac{\Delta t}{4} \left\| \sigma_h \right\|^2_{L^2(\Gamma)} + \sqrt{\frac{\Delta t}{16}} \left\| \sqrt{r} \phi_h \right\|^2_{L^2(\Gamma)} \]

\[ \leq c \left\{ \frac{1}{2} \left\| \phi_h \right\|^2 + \frac{\Delta t}{4} \left\| \sigma_h \right\|^2 + \sqrt{\frac{\Delta t}{16}} \left\| \sqrt{r} \phi_h \right\|^2 + \left\| f \right\|_{L^2(\Gamma)} + \left\| g \right\|_{L^2(\Gamma)} \right\}. \] (39)
4. Error Estimate Theorem

Now, we turn to derive a priori error estimate in $L^2$-norm for the solutions of (29). In order to state error estimates, we need two following Lagrange interpolation operators ([29] [30])

\[ \Pi: C^0(\Omega) \to V_h, \quad P: C^0(\Omega) \to W^{l}. \]

**Lemma 13.** There exist positive constants $K_1$ and $K_2$, independent of $h_0$ and $h_0$ respectively, such that

\[ \| \Pi_0 \psi - \psi \|_{L^2(\Omega)} \leq K_1 h_0^{1-s}, \quad s = 0, 1, \ \forall \psi \in H^{k+1}(\Omega) \cap C^0(\Omega), \]  

and

\[ \| P_0 \sigma - \sigma \|_{L^2(\Omega)} \leq K_2 h_0^{1-m}, \quad m = 0, 1, \ \forall \sigma \in H^{k+1}(\Omega) \cap C^0(\Omega). \]

Let $e_0^n = \phi^n - \Pi_0 \phi^n$, $\eta_0^n = \phi^n - \Pi_0 \phi^n$, $\theta_0^n = \sigma^n - P_0 \sigma^n$, $\rho_0^n = \sigma^n - P_0 \sigma^n$.

Corresponding to $A^{n+1/2}$ and $F^{n+1/2}$, we introduce a bilinear form $A^{n+1/2}$ on $H^1(\Omega) \times \mathcal{H}(\Omega) \times H^1(\Gamma)$ and a linear form $F^{n+1/2}$ on $H^1(\Omega)$ for $n = 0, \cdots, N$ as follows

\[ A^{n+1/2}(\phi, \sigma, \psi) = \left( \frac{d\phi}{dt} \right)^{n+1/2} \circ X^{n+1/2}_e, \sigma^{n+1/2} \circ X^{n+1/2}_e, \nabla \psi \right) + \left( \left( \nabla \phi^{n+1/2} \right)^T \sigma^{n+1/2} \circ X^{n+1/2}_e, \psi \right) + \left( r \phi^{n+1/2} \circ X^{n+1/2}_e, \psi \right) \left( \det \left( F^{n+1/2}_e \right) \phi^{n+1/2} \circ X^{n+1/2}_e, \psi \right) \left( \Gamma, \right) \]

\[ F^{n+1/2}(\psi) = \left( f^{n+1/2} \circ X^{n+1/2}_e, \psi \right) + \left( \det \left( F^{n+1/2}_e \right) g^{n+1/2} \circ X^{n+1/2}_e, \psi \right) \left( \Gamma, \right) \]

If $\phi$ and $\sigma = \mathbf{A} \nabla \phi$ are the solutions of (1), we have for $n = 0, \cdots, N - 1$

\[ A^{n+1/2}(\phi, \sigma, \psi) = F^{n+1/2}(\psi), \ \forall \psi \in H^1(\Omega). \]  

We decompose $e_0^n, \theta_0^n$ as

\[ A^{n+1/2}_h(e, \sigma, \psi) = \left( \left( \nabla \phi^{n+1/2}_h \right)^T \sigma^{n+1/2} \circ X^{n+1/2}_e, \psi \right) + A^{n+1/2}_h(\eta, \rho, \psi), \]

where $\psi \in V^{2}_h$.

In order to estimate the terms on the right-hand side of (44), we adopt the following lemmas.

**Lemma 14.** [23] Assume the above Claims 2 - 5 hold, and that the coefficients of the problem (1) satisfy $v \in C\left( W^{3,\infty}(\Omega) \right) \cap C\left( W^{2,\infty}(\Omega) \right) \cap C\left( L^\infty(\Omega) \right)$, $v|_{\Gamma} = 0$, $A \in W^{3,\infty}(\Omega)$, $r \in W^{2,\infty}(\Omega)$ and that $\|v\|_{W^{3,\infty}(\Omega)} \Delta t < 1/2$. Let the solution of (43) satisfy $\phi \in Z^3$, $\nabla \phi \in Z^3$, $\phi|_{\Gamma} \in C\left( L^2(\Gamma) \right)$. Then, for each $n = 0, 1, \cdots, N - 1$, there exist two functions $\varepsilon^{n+1/2}_1: \Omega \to R$ and $\varepsilon^{n+1/2}_2: \Gamma \to R$, such that
\[
(A_{n+1/2} - A_{n-1/2})(\phi, \sigma, \psi) = \left( \xi_1^{-1/2}, \psi \right)_{\mathcal{T}_h} + \left( \xi_2^{-1/2}, \psi \right)_{\mathcal{T}_h},
\]
\(\forall \psi \in H^1_0(\Omega)\). Moreover, \(\xi_1^{-1/2} \in L^2(\Omega)\), \(\xi_2^{-1/2} \in L^2(\Gamma_h)\) and the following estimates hold:
\[
\begin{align*}
\left\| \xi_1^{-1/2} \right\|_{L^2(\Omega)} &\leq \tilde{c}_1 (\Delta t)^{1/2} \left\| \sigma \right\|_{L^2(\Omega)} + \left\| \phi \right\|_{L^2(\Omega)} , \\
\left\| \xi_2^{-1/2} \right\|_{L^2(\Gamma_h)} &\leq \tilde{c}_1 (\Delta t)^{1/2} \left\| \sigma \cdot n \right\|_{L^2(\Gamma_h)} + \alpha \left\| \phi \right\|_{H^1(\Gamma_h)} ,
\end{align*}
\]
where \(\tilde{c}_1\) denotes a constant independent of \(\Delta t\).

**Lemma 15.** [23] Assume the above Claims 2 - 5 hold, and that the coefficients of the problem (1) satisfy \(v \in C^0\left( W^{2,\infty}(\Omega) \right) \cap C^1\left( W^{1,\infty}(\Omega) \right), \ f \in Z^2, \ g \in C^2\left( L^2(\Gamma_h) \right)\) and \(\left\| x \right\|_{C^0(\Omega)} \Delta t < 1/2\). Then, for each \(n = 0, 1, \ldots, N - 1,\) there exist functions \(\xi_f^{-1/2} : \Omega \rightarrow R\) and \(\xi_g^{-1/2} : \Gamma_h \rightarrow R\), such that
\[
(F_{n+1/2} - F_{n-1/2})(\psi) = \left( \xi_f^{-1/2}, \psi \right)_{\mathcal{T}_h} + \left( \xi_g^{-1/2}, \psi \right)_{\mathcal{T}_h}, \forall \psi \in H^1(\Omega).
\]
Moreover, \(\xi_f^{-1/2} \in L^2(\Omega), \xi_g^{-1/2} \in L^2(\Gamma_h)\) and the following estimates hold:
\[
\begin{align*}
\left\| \xi_f^{-1/2} \right\|_{L^2(\Omega)} &\leq \tilde{c}_1 (\Delta t)^{1/2} \left\| \phi \right\|_{L^2(\Omega)} , \\
\left\| \xi_g^{-1/2} \right\|_{L^2(\Gamma_h)} &\leq \tilde{c}_1 (\Delta t)^{1/2} \left\| \sigma \cdot n \right\|_{L^2(\Gamma_h)} ,
\end{align*}
\]
where \(\tilde{c}_1\) denotes a constant independent of \(\Delta t\).

Now, we turn to bound the third term on the right-hand side of (44).

**Lemma 16.** Assume the above Claims 2 - 5 hold, and that the coefficients of the problem (1) satisfy \(\phi \in C^0\left( C^0(\Omega) \right) \cap C^0\left( H^{k+1}(\Omega) \right) \cap H^1\left( H^k(\Omega) \right)\) and \(c_1 \Delta t < 1/2\). There exists \(A_{n+1/2}\)
\[
A_{n+1/2}(\eta_h, \rho_h, \bar{e}_h) \\
\leq \frac{1}{8} \left\| \theta^{n+1} + (C\theta^n) \cdot X^n \right\|_{L^2(\Omega)}^2 + \partial_t \left( \frac{\Delta t}{2} \left( C\rho^n, C\theta^n \right) \right) \\
+ \frac{1}{8} \left\| \nabla e^{n+1} + (\nabla e^n) \cdot X^n \right\|_{L^2(\Omega)}^2 + \partial_t \left( \frac{\Delta t}{2} \left( \nabla \eta^n, \nabla e^n \right) \right) \\
+ \frac{\alpha}{8} \left\| \bar{e}_h \right\|^2 + e^{n+1} \left( 1 + \Delta t \text{div} e^{n+1} \right) \left\| \nabla e^{n+1} \right\|^2 + \partial_t \left( \frac{\alpha \Delta t}{2} \left( \eta^n, \bar{e}_h \right) \right) \\
+ c \left\| e^{n+1} \right\|^2 + c \Delta t \left( \left\| \phi \right\|^2 + \left\| \phi^{n+1} \right\|^2 + \left\| \nabla e^n \right\|^2 + \left\| \nabla e^{n+1} \right\|^2 + \left\| \bar{e}_h \right\|^2 \right) \\
+ \tilde{c}^2 K_h \left( \frac{1}{\Delta t} \left\| \bar{e}_h \right\|^2 + \frac{1}{\Delta t} \left\| \phi \right\|^2 + \Delta t \left\| \psi \right\|^2 \right) \\
+ \tilde{c}^2 K_h \left( \left\| \sigma \right\|^2 + \tilde{c} \Delta t \left\| \bar{e}_h \right\|^2 \right).
\]
with \( c = \max \left\{ \frac{1}{4}(c_1 + 1)/4c_2, c_1/4, c_1 \right\} \), \( \tilde{c} \) a positive constant.

**Proof.** From the definition of \( A_{n+1/2} \), we have

\[
A_{n+1/2}(\eta_h, \rho_h, e_{n+1}^h)
= \left( \eta_h^{n+1} - \eta_h^n \circ X_h^n, e_{n+1}^h \right) + \left( \frac{\rho_h^{n+1} + \rho_h^n \circ X_h^n}{2}, A^{-1} \frac{e_{n+1}^h}{\Delta t} \right) \\
+ \frac{\Delta t}{2} \left( \left( L^n \rho_h^n \circ X_h^n, \nabla e_{n+1}^h \right) + \frac{\Delta t}{2} \left( \left( \nabla \text{div} \rho_h^n \circ X_h^n, \rho_h^n \circ X_h^n, e_{n+1}^h \right) \\
+ \left( \eta_h^{n+1} + \frac{\eta_h^n}{2} \right) \frac{\eta_h^{n+1} + \frac{\eta_h^n}{2}}{1 + \Delta t \text{div} \rho_h^{n+1}} \right) \right)
\]

(50)

\[
= I_1 + I_2 + I_3 + I_4 + I_5 + I_6.
\]

From the arguments of Lemma 4.1 in [24], we get similarly the bounds of the above terms \( I_1, I_2, I_6 \) as the followings:

\[
I_1 \leq \frac{(1 + c_1 \Delta t)(1 + 2c_1)^2}{2\Delta t} K^2_h \|
\eta_h^{n+1} \|_I + \left\| \frac{\partial \phi}{\partial t} \right\|_{L^2(\Omega, \mathbb{R}^d, \mathbb{R}^d)}
\]

(51)

\[
+ \frac{1}{2} \left\| e_{n+1}^h \right\|_{H^1},
\]

\[
I_5 \leq \partial_t \left( \frac{\Delta t}{2} \left( \frac{\sqrt{\eta_h^n} \cdot \sqrt{e_{n+1}^h}}{2} \right) \right) + \left( \frac{c_1 \Delta t}{4} + (1 + c_1 \Delta t) c_1 \right) \|
\eta_h^{n+1} \|_I + \frac{(2c_1 + c_1 \Delta t) c_1 \|
\eta_h^{n+1} \|_I}{\Delta t} \\
+ \frac{1}{8} \left\| \sqrt{e_{n+1}^h} \right\|_I + \frac{1}{8} \left\| \sqrt{e_{n+1}^h} \right\|_I \right)
\]

(52)

\[
I_6 \leq \partial_t \left( \frac{c_1 \Delta t}{2} \left( \eta_h^{n+1}, e_{n+1}^h \right) \right) + \left( \frac{c_1 \Delta t}{4} \right) \|
\eta_h^{n+1} \|_I + \|
\eta_h^{n+1} \|_I \right) \right) \\
+ \frac{1}{8} \left( \frac{1}{8} \left\| e_{n+1}^h \right\|_I \left( 1 + \Delta t \text{div} \rho_h^{n+1} \right) \right) \right)
\]

(53)

To estimate \( I_2 \), we divide it into three parts

\[
I_2 = \frac{1}{2} \left[ \left( (C \rho_h^{n+1}, C \theta_h^{n+1}) - (C \rho_h^n, C \theta_h^n) \right) \right] + I_{21} + I_{22},
\]

(54)

where

\[
I_{21} = \frac{1}{2} \left[ \left( C \rho_h^n, C \theta_h^n \right) - \frac{1}{2} \left( (C \rho_h^n) \circ X_h^n, (C \theta_h^n) \circ X_h^n \right) \right],
\]

\[
I_{22} = \frac{1}{2} \left[ \left( (C \rho_h^n) \circ X_h^n, C \circ X_h^n \right) \left( \theta_h^{n+1} + \theta_h^n \circ X_h^n \right) \right].
\]

Using the transformation \( y = X_h^n(x) \), we have
\[
I_{21} = \frac{1}{2} \int_{\Omega} \left( 1 - \text{det} \left( \left( F_{E}^{n} \right)^{-1}(x) \right) \right) \left( C \rho_{h}^{n} ight)(x) \cdot \left( C \theta_{h}^{n} \right)(x) \, dx
\]
\[
\leq \frac{c_{2}c_{1} \Delta t}{4} \left( K_{2}^{n} \rho_{h}^{n} \right) \| \sigma \|_{k+1}^{2} + \| \theta_{h} \|^{2}.
\]  
(55)

Now, we replace in \( I_{22} \) equality \( C \left( X_{E}^{n}(x) \right) = C(x) - D^{n}(x) \), where
\[
D_{b}^{n}(x) := \int_{\gamma} \nabla C_{0} \left( Y_{E}^{n}(x,s) \right) \cdot \mathbf{v}^{n+1}(x) \, ds, \ a.e. \ x \in \Omega,
\]  
(56)

with \( |D_{b}^{n}(x)| \leq c_{1} \sqrt{c_{2} \Delta t} \), and the function \( Y_{E}^{n}(x,s) : [t^{n}, t^{n+1}] \rightarrow \Omega \) is defined by \( Y_{E}^{n}(x,s) := x - \left( t^{n+1} - s \right) \mathbf{v}^{n+1}(x) \). Thus, we get
\[
I_{22} = \frac{1}{2} \left( \left( C \rho_{h}^{n} \right) \circ X_{E}^{n}, C \theta_{h}^{n+1} + \left( C \theta_{h}^{n} \right) \circ X_{E}^{n} \right)
\]
\[
\leq \left\| \left( C \rho_{h}^{n} \right) \circ X_{E}^{n} \right\|^{2} + \frac{1}{8} \left\| C \theta_{h}^{n+1} + \left( C \theta_{h}^{n} \right) \circ X_{E}^{n} \right\|^{2} + \frac{1}{8} \left\| D \theta_{h}^{n+1} \right\|^{2}.
\]  
(57)

Moreover, we have
\[
\left\| \left( C \rho_{h}^{n} \right) \circ X_{E}^{n} \right\|^{2} \leq \left( 1 + c_{1} \Delta t \right) c_{1} K_{2}^{n} \rho_{h}^{n} \| \sigma \|_{k+1}^{2},
\]
\[
\left\| D \theta_{h}^{n+1} \right\| \leq c_{2}^{2} \left( \Delta t \right)^{2} \left\| \theta_{h}^{n+1} \right\|^{2}.
\]  
(58)

By considering together from (54) to (58), we can state
\[
I_{2} \leq \frac{\Delta t}{2} \left( C \rho_{h}^{n}, C \theta_{h}^{n} \right) + \left( \frac{c_{2} \Delta t}{4} + (1 + c_{1} \Delta t) c_{1} K_{2}^{n} \rho_{h}^{n} \| \sigma \|_{k+1}^{2} + \frac{1}{8} \left\| C \theta_{h}^{n+1} + \left( C \theta_{h}^{n} \right) \circ X_{E}^{n} \right\|^{2}.
\]  
(59)

Similar arguments, i.e. Lemma 5.4 in [23] lead to
\[
\left\| \left( \nabla \cdot \mathbf{v}^{n} \cdot \rho_{h}^{n} \right) \circ X_{E}^{n} \right\|^{2} \leq \left( 1 + c_{1} \Delta t \right) c_{2} \Delta t \rho_{h}^{n} \| \sigma \|_{k+1}^{2},
\]
\[
\left\| \left( \nabla \cdot \mathbf{v}^{n} \cdot \rho_{h}^{n} \right) \circ X_{E}^{n} \right\|^{2} \leq \left( 1 + c_{1} \Delta t \right) c_{2} \Delta t \rho_{h}^{n} \| \sigma \|_{k+1}^{2}.
\]  
(60)

Using these inequalities, \( I_{3} \) and \( I_{4} \) can be bounded as follows
\[
I_{3} \leq \frac{(1 + c_{1} \Delta t) c_{1}^{2} \rho_{h}^{n} \| \sigma \|_{k+1}^{2}}{4 \Delta t},
\]
\[
I_{4} \leq \frac{(1 + c_{1} \Delta t) c_{1}^{2} \rho_{h}^{n} \| \sigma \|_{k+1}^{2}}{4 \Delta t}.
\]  
(61)

Gathering (50), (51), (52), (53), (59), and (61) together, we complete the proof.

We now turn to estimate \( e^{n+1}_{h} \) and \( \sigma^{n+1}_{h} \). From the definition of \( A_{b}^{n+1/2} \) and \( A^{n+1/2} \), we see
\[
A_{b}^{n+1/2}(\epsilon_{h}^{n+1}, \sigma_{h}^{n+1}) = A_{b}^{n+1/2}(\eta_{h}, \rho_{h}, \epsilon_{h}^{n+1}) + (X_{h}^{n+1/2} - X_{h}^{n+1/2})(\epsilon_{h}^{n+1})
\]
\[
+ (A^{n+1/2} - A_{b}^{n+1/2})(\phi, \sigma, \epsilon_{h}^{n+1}).
\]  
(62)

From Lemma 11, we obtain the lower bound for (62)
By jointly considering the lower bound (63), the upper bounds given in Lemmas 14 - 16, we deduce

\[
\begin{aligned}
&\delta (\frac{1}{2} e_g^2 + \frac{\Delta t}{4} \| \theta^g_n \| + \frac{\Delta t}{4} \| \sqrt{\tau} e_{\phi}^2 + \frac{\alpha_\Delta}{4} \| e_\phi^2 \|_{_{0,T_R}} ) \\
&+ \frac{1}{2\Delta t} \| e_{\phi}^{n+1} - e_g^0 \cdot X_{R_k}^n \|^2 + \frac{\delta}{4} \| \theta_{\phi}^{n+1} + \theta_{\phi}^0 \cdot X_{R_k}^n \|^2 \\
&+ \frac{1}{4} \| \sqrt{\tau} e_{\phi}^{n+1} + (\sqrt{\tau} e_{\phi}^0) \cdot X_{R_k}^n \|^2 + \frac{\alpha}{4} \| e_{\phi}^{n+1} + e_{\phi}^0 (1 + \Delta t \text{div} e_{\phi}^{n+1}) \|_{_{0,T_R}}^2 \\
&- \frac{c}{2} (e_{\phi}^2 + \| e_{\phi}^{n+1} \|^2 + \| \theta_{\phi}^2 \| + c\Delta t \| e_{\phi}^{n+1} \| + c\Delta t \| \nabla e_{\phi}^{n+1} \|^2 \\
&+ c\Delta t (\| \sqrt{\tau} e_{\phi}^{n+1} + \| \sqrt{\tau} e_{\phi}^0 \|^2 + \| \sqrt{\tau} e_{\phi}^{n+1} + \| e_{\phi}^2 \|_{_{0,T_R}}^2 ) \\
&+ \frac{\alpha\Delta t}{2} (\eta_{\phi}^{n+1}, e_{\phi}^0)_{_{T_R}} + \frac{\alpha\Delta t}{2} (\eta_{\phi}^0, e_{\phi}^n)_{_{T_R}} \\
&+ c\Delta t (\| \theta_{\phi}^2 \| + \| e_{\phi}^{n+1} \|^2 + \| \sqrt{\tau} e_{\phi}^0 \|^2 + \| \sqrt{\tau} e_{\phi}^{n+1} + \| e_{\phi}^2 \|_{_{0,T_R}}^2 ) \\
&+ \hat{c} K h_{\phi}^2 \left( \frac{1}{\Delta t} \| \phi ' \|_{_{L^2 (\tau_{j+1}, T)}}^2 + \frac{2}{\Delta t} \| \phi ' \|_{_{L^2 (\tau_j, T_{j+1})}}^2 + \Delta t \| \phi ' \|_{_{L^2 (\tau_j, T_{j+1})}}^2 \\
&+ \hat{c} K h_{\phi}^2 \| \sigma ' \|_{_{L^2 (\tau_j, T_{j+1})}}^2 + \Delta t \| e_{\phi}^{n+1} \|^2 + c \| e_{\phi}^{n+1} \|^2 \\
&+ \hat{c} \| e_{\phi}^{n+1} \|^2 + \| e_{\phi}^{n+1} \|^2 + \| e_{\phi}^{n+1} \|^2 + \| e_{\phi}^{n+1} \|^2 \right).
\end{aligned}
\]

Analogous computations to those developed in Lemma 16 give

\[
\begin{aligned}
&\left\{ \frac{\Delta t}{2} (C \rho_{\phi}^0, C \theta_{\phi}^0) \leq \frac{c_2 K h_{\phi}^2 \Delta t}{2} \| \phi ' \|_{_{L^2 (\tau_{j+1}, T)}}^2 + \frac{\Delta t}{8} \| C \theta_{\phi} \|^2 \\
&\Delta t (\sqrt{\tau} \eta_{\phi}^0, \sqrt{\tau} e_{\phi}^0) \leq \frac{c_1 K h_{\phi}^2 \Delta t}{2} \| \phi ' \|_{_{L^2 (\tau_j, T_{j+1})}}^2 + \frac{\Delta t}{8} \| \sqrt{\tau} e_{\phi} \|^2, \\
&\frac{\alpha \Delta t}{2} (\eta_{\phi}^0, e_{\phi}^n)_{_{T_R}} \leq \alpha c_2 K h_{\phi}^2 \Delta t \| \phi ' \|_{_{L^2 (\tau_j, T_{j+1})}}^2 + \frac{\alpha \Delta t}{16} \| e_{\phi} \|^2_{_{L^2 (0,T_R)}}, \quad \forall \ j = 0, N.
\end{aligned}
\]

Putting (65) into (64), multiplying (64) by \( \Delta t \), summing it about time form \( t = 0 \) and \( t = T_N \), taking the initial values \( \phi_0^0 = \Pi \phi_0^0, \sigma_0^0 = P_{\delta} \sigma_0^0 \) and using discrete Gronwall’s inequality, we can derive the following error estimate:

**Theorem 17 (Error Estimate)** Let Claims 2 - 5 be assumed. Let \( \phi \in Z^1 \cap C^0 (H^{k+1} (\Omega)) \cap H^1 (H^k (\Omega)) \) be the solution of (1) with \( \nabla \phi \in Z^2 \), \( \phi_{xR} \in Z^1 (\Gamma_R) \), \{ \phi_n^0, \sigma_n^0 \}_{n=1}^N \) be the solution of (29) subject to the initial values
\( \phi_0^0 = \Pi_\nu \phi^0, \sigma_0^0 = P_\sigma \sigma^0 \). Then there exists a positive constant \( c \) independent of \( h_\nu, h_\sigma \) and \( \Delta t \) such that

\[
\begin{align*}
&\sqrt{\frac{1}{2}} \| \phi^0 - \phi \|_{L^2([\tau^1, \tau^2])} + \sqrt{\frac{\alpha \Delta t}{8}} \| \sigma^0 - \sigma \|_{L^2([\tau^1, \tau^2])} \\
&+ \sqrt{\frac{\Delta t}{8}} \| \sqrt{\rho} \phi^0 - \sqrt{\rho} \phi \|_{L^2([\tau^1, \tau^2])} + \sqrt{\frac{\alpha \Delta t}{16}} \| \phi^0 - \phi \|_{L^2([\tau^1, \tau^2])} \\
&\leq c h^k_\nu \left( \| \phi^0 \|_{H^1(\Omega)} + \| \nabla \sigma^0 \|_{H^1(\Omega)} \right) + c h^k_\sigma \left( \| \phi^0 \|_{H^1(\Omega)} \right) \\
&+ c (\Delta t)^2 \left( \| \phi^0 \|_{L^2(\Omega)} + \| \sigma^0 \|_{L^2(\Omega)} + \| \sqrt{\rho} \phi^0 \|_{L^2(\Omega)} + \| \nabla \phi^0 \|_{L^2(\Omega)} + \| \nabla \sigma^0 \|_{L^2(\Omega)} \right).
\end{align*}
\]

### Acknowledgement

This research was supported by the NSF of China (Grant Nos. 11271231 and 11301300).

### References

[1] Frolkovic, P. and De Schepper, H. (2000) Numerical Modelling of Convection Dominated Transport Coupled with Density Driven Flow in Porous Media. *Advances in Water Resources*, 24, 63-72. [https://doi.org/10.1016/S0309-1708(00)00025-7]

[2] Markowich, P.A. and Szmolyan, P. (1989) A System of Convection-Diffusion Equations with Small Diffusion Coefficient A rising in Semiconductor Physics. *Journal of Differential Equations*, 81, 234-254. [https://doi.org/10.1016/0022-0396(89)90122-8]

[3] Morton, K.W. (1996) Numerical Solution of Convection-Diffusion Problems. Chapman & Hall, London.

[4] Ewing, R.E. and Wang, H. (2001) A Summary of Numerical Methods for Time-Dependent Advection-Fominated Partial Differential Equations. *Journal of Computational and Applied Mathematics*, 128, 423-445. [https://doi.org/10.1016/S0377-0427(00)00522-7]

[5] Wilmott, P., Dewynne, J. and Howison, S. (1993) Option Pricing, Mathematical Models and Computation. Oxford Financial Press, Oxford.

[6] Ewing R. (1983) The Mathematics of Reservoir Simulation. SIAM, Philadelphia.

[7] Johnson, C. (1987) Numerical Solution of Partial Differential Equations by the Finite Element Method. Cambridge University Press, Cambridge.

[8] Douglas, Jr.J. and Russell, T.F. (1982) Numerical Methods for Convection-Domininated Diffusion Problems Based on Combining the Method of Characteristics with Finite Element or Finite Difference Procedures. *SIAM Journal on Numerical Analysis*, 19, 871-885. [https://doi.org/10.1137/0719063]

[9] Russell, T.F. (1985) Time-Stepping along Characteristics with Incomplete Iterationfora Galerkin Approximation of Miscible Displacement in Porous Media. *SIAM Journal on Numerical Analysis*, 22, 970-1013. [https://doi.org/10.1137/0722059]

[10] Pironneau, O. (1982) On the Transport-Diffusion Algorithm and Its Application to the Navier-Stokes Equations. *Numerische Mathematik*, 38, 309-332. [https://doi.org/10.1007/BF01396435]

[11] Süli, E. (1986) Convergence Analysis of the Lagrange-Galerkin Method for the
Navier-Stokes Equations, Report 86/3. Oxford University Computing Laboratory, Oxford.

[12] Morton, K.W., Priestley, A. and Süli, E. (1986) Stability Analysis of the Lagrange-Galerkin Method with Non-Exact Integration, Report 86/14. Oxford University Computing Laboratory, Oxford.

[13] Johnson, C. and Thomée, V. (1981) Error Estimates for Some Mixed Finite Element Methods for Parabolic Type Problems. *RAIRO Analyse Numérique*, 15, 41-78. https://doi.org/10.1051/m2an/1981150100411

[14] Raviart, P.A. and Thomas, J.M. (1977) A Mixed Finite Element Method for Second Order Elliptic Problems, Mathematical Aspects of the Finite Element Method. Lecture Notes in Mathematics 606, Springer, Berlin.

[15] Nédélec, J.C. (1980) Mixed Finite Elements in $R^3$. *Numerische Mathematik*, 35, 315-341. https://doi.org/10.1007/BF01396415

[16] Douglas, Jr.J. and Roberts, J.E. (1985) Global Estimates for Mixed Methods for Second Order Elliptic Equations. *Mathematics of Computation*, 44, 39-52. https://doi.org/10.1090/S0025-5718-1985-0771029-9

[17] Ewing, R.E., Russell, T.F. and Wheeler, M.F. (1984) Convergence Analysis of an Approximation of Miscible Displacement in Porous Media by Mixed Finite Elements and a Modified Method of Characteristics. *Computer Methods in Applied Mechanics and Engineering*, 47, 73-92. https://doi.org/10.1016/0045-7825(84)90048-3

[18] Wheeler, M.F. and Dawson, C.N. (1984) An Operator-splitting Method for Advection Diffusion-Reaction Problems. *The Mathematics of Finite Elements and Applications*, V, 122-144.

[19] Arbogast, T. and Wheeler, M.F. (1995) A Characteristic-Mixed Finite Element Methods for Advection-Dominated Transport Problems. *SIAM Journal on Numerical Analysis*, 32, 404-424. https://doi.org/10.1137/0732017

[20] Sun, T.J. and Yuan, Y.R. (2009) An Approximation of Incompressible Miscible Displacement in Porous Media by Mixed Finite Element Method and Characteristics-mixed Finite Element Method. *Journal of Computational and Applied Mathematics*, 228, 391-411. https://doi.org/10.1016/j.cam.2008.09.029

[21] Sun, T.J. and Yuan, Y.R. (2015) Mixed Finite Element Method and the Characteristics-mixed Finite Element Method for a Slightly Compressible Miscible Displacement Problem in Porous Media. *Mathematics and Computers in Simulation*, 107, 24-45. https://doi.org/10.1016/j.matcom.2014.07.005

[22] Rui, H.X. and Tabata, M. (2002) A Second Order Characteristic Finite Element Scheme for Convection-Diffusion Problems. *Numerische Mathematik*, 92, 161-177. https://doi.org/10.1007/s002110100364

[23] Bermúdez, A., Nogueiras, M.R. and Vázquez, C. (2006) Numerical Analysis of Convection-Diffusion-Reaction Problems with Higher Order Characteristic/Finite Elements, Part I: Time Discretization. *SIAM Journal on Numerical Analysis*, 44, 1829-1853. https://doi.org/10.1137/040612014

[24] Bermúdez, A., Nogueiras, M.R. and Vázquez, C. (2006) Numerical Analysis of Convection-Diffusion-Reaction Problems with Higher Order Characteristic/Finite Elements, Part II: Fully Discretized Scheme and Quadrature Formulas. *SIAM Journal on Numerical Analysis*, 44, 1854-1876. https://doi.org/10.1137/040615109

[25] Brezzi, F. (1974) On the Existence, Uniqueness and Approximation of Saddle-Point Problems Arising from Lagrangian Multipliers. *RAIRO Analyse Numérique*, 2, 129-151. https://doi.org/10.1051/m2an/197408R201291
[26] Wheeler, M.F. (1973) A Priori $L_2$ Error Estimates for Galerkin Approximations to Parabolic Partial Differential Equations, SIAM Journal on Numerical Analysis, 10, 723-759. https://doi.org/10.1137/0710062

[27] Arnolds, D.N., Scott, L.R. and Vogelus, M. (1998) Regular Inversion of the Divergence Operator with Dirichlet Boundary Conditions on a Polygonal. Annali della Scuola Normale Superiore di Pisa-Classe di Scienze, 15, 169-192.

[28] Ciarlet, P.G. (1991) Basic Error Estimates for Elliptic Problems, Handbook of Numerical Analysis, Volume II. North-Holland Publishing Company, Amsterdam, 19-351.

[29] Russell, T.F. (1986) Finite Elements with Characteristics for Two-component Incompressible Miscible Displacement, SPE 10500. Proceedings of 6th SPE Symposium on Reservoir Simulation, New Orleans, 123-135.

[30] Ciarlet, P.G. (1978) The Finite Element Method for Elliptic Problems. North-Holland Publishing Company, Amsterdam.
A-Equation and Its Connections to Nonlinear Integrable System

Yilian Zhang

Department of Mathematical Science, University of South Carolina, Aiken, SC, USA
Email: yilianz@usca.edu

Abstract

A novel approach to inverse spectral theory for Schrödinger Equation operators on a half-line was first introduced by Barry Simon and actively studied in recent literatures. The remarkable discovery is a new object A-function and integro-differential Equation (called A-Equation) it satisfies. Inverse problem of reconstructing potential is then directly connected to finding solutions of A-Equation. In this work, we present a large class of exact solutions to A-Equation and reveal the connection to a class of arbitrarily large systems of nonlinear ordinary differential Equations. This non-linear system turns out to be C-integrable in the sense of F. Calogero. Integration scheme is proposed and the approach is illustrated in several examples.

Keywords

Integrable System, A-Equation

1. Introduction

Several years ago, Barry Simon investigated a new approach to inverse spectral theory for the half-line Schrödinger operator, \(-\frac{d^2}{dx^2} + q(x)\) in \(L^2(0, \infty)\) in [3].

A new A-function introduced in [1] [2] [3], is related to Weyl-Titchmarsh function by the following relation:

\[
m(x, -\kappa^2) = -\kappa - \int_0^\infty A(\alpha, x) e^{2\kappa \alpha} d\alpha
\]

where \(A(\alpha, x) \in L^1(0, a)\) for all \(a\).

In [3], the key discovery is that \(A(\alpha, x)\) satisfies the following integro-differential Equation:

\[
\frac{\partial A(\alpha, x)}{\partial \alpha} = \frac{\partial A(\alpha, x)}{\partial \alpha} + \int_0^\alpha A(\beta, x) A(\alpha - \beta, x) d\beta.
\]
Given the fact that
\[
\lim_{\alpha \downarrow 0} A(\alpha, x) = q(x)
\]  (3)
(at least in the \(L^1\) sense), \(q(x)\) can be determined directly from \(A(\alpha, x)\).
And \(A(\alpha, x)\) can be calculated from \(A(\alpha,0)\) (which is essentially the inverse Laplace transform of the data), by solving an equation which does not involve \(q(x)\). Thus the inverse problem to determine \(q\) from \(m\), becomes a problem to solve the integro-differential Equation (2). Properties of (2) are discussed in [4] [5] [6] [7]. To construct numerical solvers to this integro-differential equation, one needs to study sets of exact analytic solutions.

In this paper, we study a larger class of analytic solutions of (2), which is of the form
\[
A(\alpha, x) = \sum_{j=1}^{n} f_j(x) e^{-2\gamma_j(x)}.
\]  (4)

This ansatz is motivated by the explicit example in [1], where \(A(\alpha,0)\) is calculated for Bargmann potentials using inverse scattering theory (which is valid only under restrictive assumptions). Our aim is to determine the behavior of such solutions for all \((\alpha, x)\) and to do so using only (2).

Substituting (4) in (2), we find that \(f_j(x) = -2\gamma'_j(x)\), and \(\gamma_j\) satisfy the nonlinear equations:
\[
\gamma_j'' = -2\gamma_j'\gamma_j' + \sum_{r,j} 2\gamma_r'\gamma_j'^r, \quad 1 \leq j \leq n.
\]  (5)

Then we give a method for solving (5) explicitly in Section 3. The idea is to introduce new variables \(c_j\), the symmetric functions of \(\gamma_j\) \((j = 1, \ldots, n)\), that is \(c_j = \sum_{\{i_1, \ldots, i_j\} \subseteq \{1, \ldots, n\}} \gamma_{i_1}\gamma_{i_2}\cdots\gamma_{i_j}\) . Via this “change of variable”, (5) yields a new nonlinear system:
\[
2c_j' = 2c_{j+1}' + c_{j+1}', \quad 1 \leq j \leq n+1.
\]  (6)
\[
c_0 = 1, \quad c_j = 0 \quad \text{when } j > n.
\]

This nonlinear system turns out to be solvable. Calogero proved that a certain family of n-body problems is solvable in a 2004 J. Math. Phys. paper and his model includes system (6), and the method we use in this method is different from his approach. Our method also shows some insightful connection to scattering problems. In Section 3, first we find \(n\) constants of motion for the system (6) which allow us to reduce it to a first order nonlinear system. Explicitly we will prove

**Theorem 1.** (i) Supposing that for any \(x\) in an open interval \(I\), \(c_j\) are solutions of the second order nonlinear system (6). Then on \(I\), \(c_j\) solves the first order system
\[
\sum_{k=1}^{n} (-1)^{j-k} c_{2j-4-k} (c_k' - c_{k+1}') = \mu_j
\]  (7)
for \(j = 0, \ldots, n\). Here \(\mu_j (j \neq 0)\) are constants and \(\mu_0 = 1\).

(ii) Conversely, if \(c_j(x)\) is solutions of (7) with \(c_k(x) \neq 0\) and \(\gamma_j^2(x) \neq 0\)
\( \gamma_i^2(x) \) for \( x \in I \), then (6) holds.

The latter is then solved by finding a nonlinear analogue of the method of integrating factors (Theorem 13).

We note that \( \gamma_j \) is zeros of polynomials with coefficients \( c_j \). Calogero pointed out in [8] [9] that some nonlinear systems can be linearized by nonlinear mapping between coefficients of polynomial and its zero, and thus is integrable. The novelty in this paper is that the nonlinear mapping from \( \gamma_j \) to \( c_j \) relates the system (5) to a solvable yet still nonlinear system. Interestingly, a system similar to (6) arises ([10] [11]) if one seeks potentials for which the large frequency WKB series is finite and yields solutions of the corresponding Schrödinger equations (with no error).

Section 4 shows how we obtain analytic examples of (2) by following this systematic procedure.

2. The \( \gamma \) Equation

As described in introduction, we relate a large class of exact solution of A-Equation (2) to a second order non-linear system (5).

Without loss generality, we assume \( \gamma_i \neq \gamma_j \) for all \( i \neq j \). Then the following proposition can be followed by direct calculation.

\textbf{Proposition 2.} If \( A(\alpha, x) \) is of the form (4), and satisfies (2), then \( f_j(x) = -2\gamma_j'(x) \), and \( \gamma_j(x) \) satisfy (5). Conversely, if \( \gamma_j(x) \) satisfy (5), then the function \( A(\alpha, x) = -2\sum_{j=1}^{n} \gamma_j e^{2\alpha \gamma_j} \) solves (2).

Our goal is to solve (5) explicitly. To begin with, we need some notations. Let \( \delta_l \) be the \( l \)th symmetric function on \( \gamma_k \), \( k \neq j \):

\[
\delta_l = \sum_{i, j} \gamma_i^l \gamma_j^l \cdots \gamma_n^l. \tag{8}
\]

\textbf{Lemma 3.} If \( \gamma_1^2, \cdots, \gamma_n^2 \) are distinct, and \( \delta_l^j (0 \leq l \leq n-1) \) are the symmetric functions on \( \gamma_k^2 \), \( k \neq j \), \( 1 \leq k \leq n \), then the matrix

\[
\begin{pmatrix}
\delta_0^j & \delta_1^j & \cdots & \delta_n^j \\
\delta_0 & \delta_1^j & \cdots & \delta_n^j \\
\vdots & \vdots & \ddots & \vdots \\
\delta_0^j & \delta_1 & \cdots & \delta_n^j
\end{pmatrix}
\tag{9}
\]

is invertible.

\textbf{Proof.} Suppose the matrix is not invertible; then there exists a non-zero vector \((a_1, a_2, \cdots, a_n)^T\), such that

\[
\sum_{j=1}^{n} \delta_l^j a_j = 0 \ \forall l. \tag{10}
\]

Then

\[
\sum_{j=1}^{n} a_j \prod_{i \neq j} (z - \gamma_i^2) = -\sum_{j=1}^{n} (-1)^l \sum_{l=1}^{n} \delta_l^j a_j z^{n-l} = 0 \ \forall z. \tag{11}
\]

Evaluate the above at \( z = \gamma_i^2 \) (\( j_0 = 1, \cdots, n \)):
\[ a_{i_0} \prod_{m \neq i_0} \left( y_{i_0}^2 - y_m^2 \right) = 0. \]  

(12)

We assumed \( y_{i_0}^2 \) are distinct, so \( a_{i_0} = 0 \) for all \( 1 \leq i_0 \leq n \). This contradicts our assumption, and proves the given matrix must be invertible.

3. A Transformed System and Explicit Solutions

3.1. Non-Linear Integrable Equation

To solve Equation (5) explicitly, we construct a nonlinear mapping from \( \gamma \) to new dependent variables \( c \). We take \( c_j \) to be the \( j \)-th symmetric function of \( \gamma_j \),

\[ c_j = \sum_{k < \cdots < k_j} \gamma_k \cdots \gamma_j. \]  

(13)

For convenience, we define \( c_0 = 1, c_n = 0 \) for \( k < 0 \) or \( k > n \).

**Proposition 4.** If \( \{ \gamma_j \} \) satisfy Equation (5), then \( \{ c_j \} \) as defined by (13), satisfy the system:

\[ 2c_j' = 2c_j c_j - c_j, \quad 1 \leq j \leq n + 1. \]  

(14)

\[ c_0 = 1, \quad c_j = 0 \quad \text{when} \quad j > n. \]  

(15)

**Proof.** It follows directly by calculation, that for every \( 1 \leq j \leq n \),

\[ 2c_j' - 2c_j c_j - c_j = -\sum_{k < -j} \sum_{j \in \gamma_{k} \cdots \gamma_{j}} \frac{2y_k y_j}{\gamma_k - \gamma_j} \prod_{m \in \gamma_{k+1}} y_m = 0. \]  

(16)

And for \( j = n + 1 \), we have

\[ -2c_n' - c_n = -2 \sum_{i=1}^{n} y_i c_n - \sum_{i=1}^{n} \left( \sum_{j \in \gamma_i} y_j + \sum_{j \notin \gamma_i} y_j y_j' \right) c_n \]

\[ = -c_n \sum_{i=1}^{n} \left( \gamma_i - \gamma_i' \right) = 0 \]  

(17)

Conversely, we have

**Proposition 5.** If \( c_j \) satisfy the system (14), and \( \gamma_1 \cdots \gamma_n \) are the distinct roots of the polynomial with coefficients \( (-1)^j c_j \), then \( \gamma_j \) satisfy the system (5).

**Proof.** As in the previous calculations, for every \( 1 \leq j \leq n + 1 \), we have

\[ 2c_j' - 2c_j c_j - c_j = \sum_{k < -j} \sum_{j \in \gamma_{k} \cdots \gamma_{j}} \left( \gamma_k - 2 \gamma_k y_j' + \sum_{l \in \gamma_i} 2y_l y_j' \right) \neq \sum_{l \in \gamma_i} \gamma_l \cdots \gamma_j = 0 \]  

(18)

By assumption, \( \gamma_i \neq \gamma_j, \forall i \neq j \). The proof of Lemma 3 then shows that the matrix \( \left( \delta_{ij} \right) \) ( \( j = 2, \cdots, n + 1, m = 1, \cdots, n \) ), where
\[ \delta_j^m = \sum_{k \leq j \leq j_k} \gamma_k \cdots \gamma_j, \]

is invertible. Thus

\[ -\gamma_m^m - 2\gamma_m^m \gamma_j + \sum_{j \leq n} 2\gamma_j \gamma_m^m = 0 \quad \text{for} \ m = 1, \cdots, n; \]

and \( \gamma_j \) satisfy (5).

### 3.2. Second Order Nonlinear System to First Order Nonlinear System

We have identified \( n \) constants of motion for the system (14). This will allow us to reduce the second order system to a first order system.

**Proposition 6.** The nonlinear system (14) has the following constants of motion:

\[ \sum_{k \leq 0} (-1)^{j-k} \begin{vmatrix} c_{j, k+1} & c_{j, k} & c_{j, k+1} \\ c_{j, k+1} & c_{j, k} & c_{j, k+1} \\ c_{j, k+1} & c_{j, k} & c_{j, k+1} \end{vmatrix} = \text{Const} \quad (19) \]

for all the \( 0 \leq j \leq n \). Here \( c_j = 0 \), when \( j > n \) or \( j < 0 \).

**Proof.** Since \( 2c_j' = 2c_j' c_{j+1} + c_{j+1}' \), \( 1 \leq j \leq n+1 \), we can write

\[ c_j'' = 2c_j' - 2c_j c_{j+1}, \quad (20) \]

\[ c_j' = 2c_j' - 2c_j c_j. \quad (21) \]

Multiplying the first of these equations by \( c_j \) and the second equation by \( c_{j+1} \), and subtracting, we have

\[ c_j c_j'' - c_j c_j'' = 2c_j c_j' - 2c_j c_{j+1}, \]

so that,

\[ 2c_j' c_{j+1}' = \left( c_{j+1}' c_{j+1}' - c_{j+1}' c_j + c_j' \right). \quad (22) \]

Similarly, from

\[ c_{j+1}' = 2c_{j+1}' - 2c_j c_{j+1}, \]

and

\[ c_j'' = 2c_j' - 2c_j c_{j+1}, \]

we find

\[ c_j c_j'' - c_j c_{j+1}' = 2c_j c_{j+1} + c_j c_{j+1}' - c_{j+1} c_{j+1}'. \]

Using this equation we obtain (compare with (22))

\[ 2c_j' c_{j+1}' = -2c_j c_{j+2} - \left( 2c_j c_j + c_{j+2} c_{j+1}' - c_{j+1} c_{j+1}' \right). \]

It follows by induction that

\[ 2c_j' c_{j+1}' = \sum_{k=0}^{j} (-1)^{j-k} 2c_k c_j + \sum_{k=0}^{j} (-1)^{j-k} \left( c_{j-k} c_{j+k} - c_{j-k} c_{j+k} \right). \]

This identity, together with (22) shows that
\[
\sum_{k=0}^{n-j} (-1)^{k+1} (c_{j-k} c'_{j+k} - c'_{j-k} c_{j+k}) + c'_j - \sum_{k=1}^{n-j} (-1)^{k+1} 2c_{j+k} c_{j-k} = \text{Const},
\]

for \(0 \leq j \leq n\).

We can also write (19) as

\[
\sum_{k=1}^{j} (-1)^{j-k} c_{2j-k-1}(c'_k - c_{k+1}) = \mu_j
\]

(23)

for \(j = 0, \ldots, n\). Here \(\mu_j (j \neq 0)\) are constants and \(\mu_0 = 1\).

Theorem 1 presents the equivalence of the first order system (23) and the second order system (14).

**Proof.** of Theorem 1

(i) this result follows directly from Proposition 6.

(ii) let \(T_j = c'_j - 2c'_{j+1} + 2c'_j\). Differentiate (23), for \(1 \leq j \leq n\),

\[
0 = \left(\sum_{k=1}^{j} (-1)^{j-k} c_{2j-k-1}(c'_k - c_{k+1})\right)' = \sum_{k=j}^{\min(j+1,n-j)} c_{2j-k-1}T_k.
\]

If we write the above equation in matrix form, we have

\[
\begin{pmatrix}
-c_0 & 0 & \cdots & 0 & 0 \\
-c_2 & -c_1 & \cdots & 0 & 0 \\
-c_4 & c_3 & \cdots & 0 & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
(-1)^{n-1}c_n & (-1)^{n-1}c_{n-1} & \cdots & (-1)^{n-1}c_2 & (-1)^n c_1 \\
0 & 0 & \cdots & -c_{n-2} & c_{n-1} \\
0 & 0 & \cdots & c_n & -c_{n-1}
\end{pmatrix}
\begin{pmatrix}
T_1 \\
T_2 \\
T_3 \\
\vdots \\
T_{n-1} \\
T_n
\end{pmatrix} = 0
\]

(24)

The coefficient matrix of above Equation (24) is a Sylvester resultant matrix. A well known theorem from linear algebra then expands the determinant of the Sylvester resultant matrix as the resultant of the two polynomials,

\[
a(s) = -c_0 s^{n-1} + c_1 s^{n-2} - \cdots + (-1)^{n} c_{n-1},
\]

\[
b(s) = c_0 s^n - c_1 s^{n-1} + c_2 s^{n-2} - \cdots + (-1)^n c_n.
\]

The coefficient matrix of (24) is nonsingular if and only if \(a(s)\) and \(b(s)\) are coprime for \(x \in I\). Let \(Q_i(s)\) be the polynomial

\[
Q_i(s) = \sum_{j=0}^{n} (-1)^j c_j s^{n-j}.
\]

We observe that \((-1)^{n} (b(-s^n) - sa(-s^n)) = Q_i(s)\), and since \(c_j\) is j-th symmetric function of \(\gamma_j\), we have

\[
(-1)^{n} (b(-s^n) - sa(-s^n)) = Q_i(s) = \prod_{j=1}^{n} (s - \gamma_j).
\]

If \(a(s)\) and \(b(s)\) are not coprime, they have a common root \(s_0\), such that \(s_0 \neq 0\). Let \(s_1, s_2\) be the two distinct square roots of \(-s_0\). Substitute \(s = s_1\)}


and \( s = s_2 \) in (26); this yields \( Q_1(s_1) = 0 \) and \( Q_2(s_2) = 0 \) respectively. Thus there exist \( \gamma_0, \gamma_0^\prime \) such that \( \gamma_0 = s_1, \gamma_0^\prime = s_2 \).

Since \( \gamma_1, \cdots, \gamma_n^\prime \) are assumed distinct, we obtain \( s_1^2 \neq s_2^2 \), which contradicts the fact that \( s_1^2 = s_2^2 = -s_0 \).

Therefore \( a(s) \) and \( b(s) \) must be coprime for \( s \in I \), and (24) has the unique trivial solution:

\[
T_j = c_j^* - 2c_j^* + 2c_j^* c_j = 0 \quad \text{for} \quad j = 1, \cdots, n.
\]

Thus \( c_j \) solve the second order system (14).

### 3.3. Method of Integrating Factor

We have reduced the second order non-linear system (14) to the first order non-linear system (23). To solve the latter system explicitly, we begin by writing it in matrix form. Let

\[
S = \begin{pmatrix}
-c_0 & 0 & 0 & \cdots & 0 & 0 \\
c_2 & -c_1 & c_0 & \cdots & 0 & 0 \\
-c_4 & c_3 & -c_2 & \cdots & 0 & 0 \\
c_6 & -c_5 & c_4 & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
(-1)^{\frac{n}{2}-1} c_n & (-1)^{\frac{n}{2}} c_{n-1} & (-1)^{\frac{n}{2}-1} c_{n-2} & \cdots & (-1)^{\frac{n}{2}} c_0 & 0 \\
0 & 0 & (-1)^{\frac{n}{2}} c_n & \cdots & (-1)^{\frac{n}{2}-1} c_2 & (-1)^{\frac{n}{2}} c_1 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \cdots & -c_n & c_{n-1}
\end{pmatrix}
\]

We will assume \( n \) is even from now on. When \( n \) is odd, we obtain similar results. The Equation (23) can be written as a matrix equation.

\[
S \begin{pmatrix}
-c_0 \\
-c_1 \\
c_1 - c_2 \\
\vdots \\
c_n' - c_{n+1} \\
c_n'
\end{pmatrix} = \begin{pmatrix}
\mu_0 \\
\mu_1 \\
\mu_2 \\
\vdots \\
\mu_n \\
\mu_n
\end{pmatrix}
\]

for \( j = 0, \cdots, n \).

(28)

We will show that the nonlinear system (28) can be solved explicitly. Let

\[
C = \begin{pmatrix}
c_0 \\
c_1 \\
\vdots \\
c_n
\end{pmatrix}, \quad C^\prime = \begin{pmatrix}
c_0' \\
c_1' \\
\vdots \\
c_n'
\end{pmatrix}
\]

The nonlinear system (28) can be written as

\[
S(c) J(c) C^\prime - S(c) J(c) c C = \mu,
\]

where
Our goal is to find an integrating factor $M$ such that after multiplication on the left by $M$, (30) takes the form

$$MS(c)J_1 C = MS(c) J_2 C = (NC)' = M \mu = 0.$$  

(33)

Thus we would like to find an $n \times (n+1)$ matrix $M$ and an $n \times (n+1)$ matrix $N$ such that

$$MS(c) J_1 = N;$$

(34)

$$MS(c) J_2 = -N';$$

(35)

$$M \mu = 0.$$  

(36)

This leads to

$$MS(c) = (B,N) = (-N'B_{n+2}).$$

(37)

$B_1$ is the first column vector of $MS(c)$ and $B_{n+2}$ is the last column vector of $MS(c)$.

For any $n \times n + 1$ matrix $N = (a_{ij})$ which satisfies (37), and $B_i = (\beta_{1i}, \beta_{2i}, \cdots, \beta_{ni})^T$, $B_{n+2} = (\beta_{1n+2}, \beta_{2n+2}, \cdots, \beta_{nn+2})^T$,

$$\begin{bmatrix}
\begin{array}{cccc}
a_{11}' & a_{12}' & \cdots & a_{1n+1}' \\
a_{21}' & a_{22}' & \cdots & a_{2n+1}' \\
& & \ddots & \vdots \\
& & & a_{n1}' & a_{n2}' & \cdots & a_{nn+1}' \\
\end{array}
\end{bmatrix} =
\begin{bmatrix}
\begin{array}{cccc}
\beta_{11} & a_{11} & \cdots & a_{1n+1} \\
\beta_{21} & a_{21} & \cdots & a_{2n+1} \\
& & \ddots & \vdots \\
& & & \beta_{n1} & a_{n1} & \cdots & a_{nn+1} \\
\end{array}
\end{bmatrix} = MS.$$  

(38)

We must have $a_{uv}' = -a_{v(u+1)}$, $a_{u1}' = -\beta_{u1}$, for $u = 1, \cdots, n; v = 2, \cdots, n+1$.

Let $f_u = a_{uu+1}$, then these conditions show that $N$ must be of the form

$$\begin{bmatrix}
\begin{array}{cccc}
f_1^{(n)} & -f_1^{(n-1)} & \cdots & f_1 \\
f_2^{(n)} & -f_2^{(n-1)} & \cdots & f_2 \\
& & \ddots & \vdots \\
& & & f_n^{(n)} & -f_n^{(n-1)} & \cdots & f_n \\
\end{array}
\end{bmatrix},$$

(38)

and moreover $\beta_{u1} = -f_u^{(n+1)}$, for $u = 1, \cdots, n$.

To find $M$, we now rewrite (34) and (36) in matrix form,
Thus each of the \( n \) rows of \( M \) solves an over-determined linear system, consisting of \( n + 3 \) equations and \( n + 1 \) unknowns.

Studying the structure of the matrix \( S(c) \), we notice the following algebraic identity,

**Lemma 7.** \( S(c)J_C = 0 \)

As an immediate corollary, we have the following

**Lemma 8.** Given a nontrivial solution \( \{c_i\}(i = 1, \cdots, n) \) of (28), the overdetermined system (39) is solvable only if \( N = 0 \). Moreover, we also have \( N^*C = 0 \).

For each overdetermined system

\[
S(c)^T M_i^T = \begin{cases}
-f_i^{(n+1)} \\
f_i^{(n)} \\
- \cdots \\
f_i \\
\end{cases}, \quad i = 1, \cdots, n.
\]  

where \( M_i^T \) is the \( i \)-th column vector of \( M^T \), Lemma 7 and Lemma 8 show that the rank of the augmented matrix is less than \( n + 2 \). The over-determined system has at most \( n + 1 \) linear independent equations.

Let \( S_1(c)^T \) denote the sub-matrix of \( S(c)^T \) obtained by deleting the first row, the \( n \)-th row, the \((n + 1)\)-th row, first column and last column of \( S(c)^T \). We observe that \( S_1(c)^T \) is also a Sylvester resultant matrix. The determinant of a Sylvester resultant matrix is the resultant of the two polynomials

\[
a(s) = -c_1 s^{n-1} + c_2 s^{n-2} - \cdots + (-1)^n c_{n-1},
\]

\[
b(s) = c_0 s^n - c_2 s^{n-2} + c_4 s^{n-4} - \cdots + (-1)^n c_{n}.
\]

Two cases need to be considered here. (1) \( a(s) \) and \( b(s) \) are coprime: Then \( S_1(c)^T \) is nonsingular. The augmented matrix of (40) has the same rank \( n + 1 \) as the corresponding coefficient matrix. Thus (40) is solvable. (2) \( a(s) \) and \( b(s) \) are noncoprime: We then use the following result of Laidacker [12]: let \( d(s) \) be the greatest common divisor of two polynomials \( a(s), b(s) \), then the rank of the Sylvester resultant matrix is \( n - \deg(d(s)) - 1 \). Thus the rank of the augmented matrix should be also \( n + 1 - \deg(d(s)) \), if (40) is to be solvable.

We obtain an algebraic fact about the Sylvester matrix.

**Lemma 9.** Suppose that \( c_d t^n - c_1 t^{n-1} + c_2 t^{n-2} - c_3 t^{n-3} \cdots + c_n = 0 \), and that \( a(-t^2) \), and \( b(-t^2) \) do not vanish simultaneously. Then the following algebraic system is always solvable.
Proof. Let \( \eta_j(s) \) be the polynomial with coefficients consisting of the \( i \)-th row of \( S_i(c)^T \). Let \( e_i(s) \) be the polynomial with coefficients consisting of the \( i \)-th row of \( S_i(c)^T \) in echelon form. It should be noted that \( e_i(s) \) is a linear combination of the polynomials of \( \eta_j(s) \). The algebraic system is solvable if and only if each zero row of \( S_i(c)^T \) in echelon form corresponds to a zero row of the augmented matrix in echelon form.

Suppose the \( i \)-th row of \( S_i(c)^T \) in echelon form is zero, that is \( e_i(s) = 0 \). From the structure of \( S_i(c)^T \),

\[
e_i(s) = \left( k_0 s^{n-1} - k_2 s^{n-2} + \cdots + (-1)^{k_n} k_{n-2} \right) a(s) + \left( k_1 s^{n-2} - k_3 s^{n-3} + \cdots + (-1)^{k_{n-2}} k_{n-3} \right) b(s) = 0.
\]

Let \( d(s) \) be the greatest common divisor of the two polynomials \( a(s) \) and \( b(s) \). Let \( a(s) = d_a(s) d(s), b(s) = d_b(s) d(s) \), where \( \gcd(d_a(s), d_b(s)) = 1 \). The above shows that there exists a polynomial \( d_a(s) \) such that

\[
k_0 s^{n-1} - k_2 s^{n-2} + \cdots + (-1)^{k_n} k_{n-2} = d_a(s) d_a(s) \quad (41)
\]

where \( k_2, \ldots, k_{n-2} \) are arbitrarily distinct and non-zero constants.

Lemma 8 also shows that (39) is solvable only if \( N \) satisfies \( NC = 0 \).
To calculate $NC$, we first introduce some new notation. For each $\kappa_i$, we define $O_j, E_i$ as follows,

$$O_j = \kappa^{n}_i c_0 + \kappa^{n-2}_i c_2 + \cdots + c_n,$$

$$E_i = \kappa^{n+1}_i c_1 + \kappa^{n+3}_i c_3 + \cdots + \kappa c_{n+1}.$$  \hfill (42)

Then, we have

$$NC = \begin{pmatrix}
\kappa_1^n f_1 c_0 - \kappa_1^{n-2} f_1' + \cdots + f_1 c_n \\
\kappa_2^n f_2 c_0 - \kappa_2^{n-2} f_2' + \cdots + f_2 c_n \\
\vdots \\
\kappa_n^n f_n c_0 - \kappa_n^{n-2} f_n' + \cdots + f_n c_n
\end{pmatrix} = \begin{pmatrix}
O_1 f_1 - \frac{1}{\kappa_1} E_j f_1' \\
O_2 f_2 - \frac{1}{\kappa_2} E_j f_2' \\
\vdots \\
O_n f_n - \frac{1}{\kappa_n} E_j f_n'
\end{pmatrix}. \hfill (44)$$

$NC = 0$ only if $O_j f_j - \frac{1}{\kappa_j} E_j f_j' = 0$ for $j = 1, \ldots, n$, i.e. $f_j' = \kappa_j \frac{O_j}{E_j} f_j$.

The following proposition proves that $N$ satisfies both $NC = 0$ and $N^*C = 0$.

**Proposition 10.** If (28) is satisfied and $\kappa_j, O_j, E_j$ are defined as above, then

$$\frac{1}{\kappa_j} \left( O_j' E_j - O_j E'_j \right) + O_j^2 = E_j^2.$$  \hfill (45)

**Proof.** Straight forward calculation.

Recall that our objective is to construct an integrating factor to reduce (23) to an algebraic system. As described above, let

$$f_j(x) = e^{\int_{E_j}^{O_j} f_j dt} \text{ for } j = 1, \ldots, n. \hfill (46)$$

Thus

$$f_j' = \kappa_j \frac{O_j}{E_j} f_j, \hfill (47)$$

$$f_j'' = \kappa_j \left( \frac{O_j}{E_j} \right) f_j + \kappa_j^2 \left( \frac{O_j}{E_j} \right)^2 f_j = \kappa_j^2 f_j. \hfill (48)$$

We assume $E_j \neq 0$ in this transformation.

**Lemma 11.** $f_j^{(l)} = \kappa_j^l \frac{O_j}{E_j} f_j$ if $l$ is odd and $f_j^{(l)} = \kappa_j^l f_j$ if $l$ is even.

$N$ can be rewritten as

$$\begin{pmatrix}
\kappa_1^n f_1 - \kappa_1^{n-2} \frac{O_1}{E_1} f_1 & \kappa_1^{n-2} f_1 & \cdots & f_1 \\
\kappa_2^n f_2 - \kappa_2^{n-2} \frac{O_2}{E_2} f_2 & \kappa_2^{n-2} f_2 & \cdots & f_2 \\
\vdots & \vdots & \ddots & \vdots \\
\kappa_n^n f_n - \kappa_n^{n-2} \frac{O_n}{E_n} f_n & \kappa_n^{n-2} f_n & \cdots & f_n
\end{pmatrix}. \hfill (49)$$
With \( f_j \) as given above, we can define a matrix \( M_1 \),
\[
M_1 = \begin{pmatrix}
k_1^2 f_1 & -k_1^{2n-2} f_1 & \cdots & (-1)^n f_1 \\
k_2^2 f_2 & -k_2^{2n-2} f_2 & \cdots & (-1)^n f_2 \\
\vdots & \vdots & \ddots & \vdots \\
k_n^2 f_n & -k_n^{2n-2} f_n & \cdots & (-1)^n f_n
\end{pmatrix}.
\] (50)

The following proposition shows that \( M = \frac{1}{E_1 E_2 \cdots E_n} M_1 \) is a solution of (39), i.e. \( M \) is an integrating factor of (30).

**Proposition 12.** Given \( M = \frac{1}{E_1 E_2 \cdots E_n} M_1 \), \( f_j \) by (46), and \( \kappa_j^2 \) \((j = 1, \ldots, n)\) be distinct roots of \( \Psi(z) = \sum_{i=0}^n (-1)^i \mu_i z^{-i} \), then \( M \) solves (34), (35), (36).

**Proof.** Straightforward calculation shows
\[
M_1 S(c) J_1 = \begin{pmatrix}
-k_1^2 f_1 O_1 & k_1^{2n} f_1 E_1 & \cdots & -f_1 O_1 \\
-k_2^2 f_2 O_2 & k_2^{2n} f_2 E_2 & \cdots & -f_2 O_2 \\
\vdots & \vdots & \ddots & \vdots \\
-k_n^2 f_n O_n & k_n^{2n} f_n E_n & \cdots & -f_n O_n
\end{pmatrix},
\]
thus, \( MS(c) J_1 = N \). Similar calculation proves \( MS(c) J_2 = N' \). At the same time, we have
\[
M \mu = \frac{1}{E_1 E_2 \cdots E_n} \begin{pmatrix}
\sum_{i=0}^n (-1)^i \mu_i \left( \kappa_1^2 \right)^{n-i} f_1 \\
\vdots \\
\sum_{i=0}^n (-1)^i \mu_i \left( \kappa_n^2 \right)^{n-i} f_n
\end{pmatrix} = 0.
\]

**Theorem 13.** If \( S(c'_i - c_{i+1}) = (\mu_i) \), then \( M_1 \cdot S(c'_i - c_{i+1}) = 0 \). Conversely, if \( M_1 \cdot S(c'_i - c_{i+1}) = 0 \) and \( f_j \neq 0 \), then \( S(c'_i - c_{i+1}) = (\mu_i) \), moreover, \( c_k \) satisfy the linear system,
\[
NC = \sum_{i=0}^n f_j^{(n-k)} (-1)^i c_k = 0, \quad 0 < j \leq n.
\] (51)

The system (28) is integrable and equivalent to this linear system.

**Remark 14.** \( f_j \) can be solved directly from Equation (48) as \( f_j = \sinh(j + \delta) \). The algebraic system (51) will then lead to the solution \( c_k \).

**Proof.** Multiplying Equation (30) by \( M_1 \), and using Proposition 12, yields \( M_1 \cdot S(c'_i - c_{i+1}) = 0 \). Further, multiplying both sides by \( \frac{1}{E_1 E_2 \cdots E_n} \), we have
\[
M \cdot S(c'_i - c_{i+1}) = (NC)' = 0.
\] (52)

Thus, \( \sum_{i=0}^n (-1)^j f_j^{(n-k)} c_k = \text{Const} \). The fact that
\[ MSJ_i C = \sum_{k=0}^{n} (-1)^k f_j^{(n-k)} c_k = 0 \]

forces the constant to be zero.

Conversely if \( M_1 \cdot S(c'_j - c_{k,1}) = 0 \), since \( M_1 \) is a \( n \times (n+1) \) matrix and \( \kappa_j^2 \) are assumed distinct, \( f_j \neq 0 \), so the dimension of the kernel is 1. The solution \( \mu \) which satisfies \( M_1 \cdot \mu = 0 \) and \( \mu_0 = 1 \) is unique. Since the first entry of \( S(c'_j - c_{k,1}) \) is \(-c_0) \cdot (-c_0) = 1\), it follows that \( S(c'_j - c_{k,1}) = (\mu_j) \).

This proves that (28) is integrable and provides a procedure to obtain explicit solutions from the linear system (51).

4. Exact Analytic Examples

We will illustrate the procedure of section 2 and section 3 by a simple exact analytic example in this section.

We explicitly discuss the case, where \( n = 2 \) in (4). Then \( A(\alpha, x) = -2y'_1(x) e^{-2y_1(x)} - 2y'_2(x) e^{-2y_2(x)} \). We construct the non-linear mapping from \( \gamma \) to \( c \),

\[
\begin{align*}
c_1 &= \gamma_1 + \gamma_2, \\
c_2 &= f_1 f_2'.
\end{align*}
\]

Then \( c_1, c_2 \) satisfy (14). To solve for \( c \), we reduce the second order system (14) to the first order system (30):

\[
\begin{pmatrix}
-c_0 & 0 & 0 & 0 \\
c_0 & -c_1 & 0 & 0 \\
0 & -c_2 & c_1 & c_2' \\
0 & 0 & c_2 & c_2'
\end{pmatrix}
\begin{pmatrix}
-c_0 \\
-c_1 \\
c_1 - c_2 \\
c_1 - c_2'
\end{pmatrix} =
\begin{pmatrix}
\mu_0 \\
\mu_1 \\
\mu_2 \\
\mu_2
\end{pmatrix}.
\]

Given \( \kappa_1, \kappa_2 \) where \( \kappa_1 \neq \kappa_2 \), we construct \( M \) of the form (50)

\[
M = \begin{pmatrix}
\kappa_2^{-1} f_1 & -\kappa_2^{-1} f_1 & f_1 \\
\kappa_2^{-1} f_2 & -\kappa_2^{-1} f_2 & f_2
\end{pmatrix}
\]

where \( f_1, f_2 \) are solution of (48):

\[
\begin{align*}
f_1' &= k_1^2 f_1, \\
f_2' &= k_2^2 f_2.
\end{align*}
\]

We can write

\[
f_i = \sinh \left( \kappa_i \left( x + \delta_i \right) \right) \quad \text{for} \quad i = 1, 2.
\]

with \( \delta_1, \delta_2 \neq 0 \).

After multiplying (54) by \( M \) on the left, the first order system is solved explicitly. Indeed, \( c_j \) satisfy linear system (51). Let

\[
z_i = \kappa_i \coth \left( \kappa_i \left( x + \delta_i \right) \right).
\]

Then (51) takes the form

\[
\begin{align*}
z_1 c_1 - c_2 &= \kappa_1^2, \\
z_2 c_2 - c_2 &= \kappa_2^2.
\end{align*}
\]
with solution:

\[ c_1(x) = \frac{\kappa_2^2 - \kappa_1^2}{z_2 - z_1}, \quad (59) \]

\[ c_2(x) = \frac{\kappa_2^2 z_1 - \kappa_1^2 z_2}{z_2 - z_1}. \quad (60) \]

To invert the mapping (53) we find \( \gamma_1(x) \), \( \gamma_2(x) \) as the roots of the equation

\[ s^2 - c_1(x)s + c_2(x) = 0. \]

This gives the following exact solutions of \( A \)-equation:

\[ q(x) = -2\gamma'_1(x) - 2\gamma'_2(x) = -2c'_1(x) \]

\[ A(\alpha, x) = -2\gamma'_1(x)e^{-2\alpha y(x)} - 2\gamma'_2(x)e^{-2\alpha y(x)}. \]

**Theorem 15.** For any distinct non-zero complex \( \kappa_1^2, \ldots, \kappa_n^2 \) and \( d_1^2, \ldots, d_n^2 \), there exists a solution \( A(\alpha, x) \) of the \( A \)-equation with the form

\[ A(\alpha, x) = \sum_{j=1}^{n} -2\gamma'_1(x)e^{-2\alpha y_j(x)} \quad \text{where} \quad \gamma'_j(0) = d_j \]

\[ \gamma'_j(0) = \prod_{m=1}^{n} (\kappa_m^2 - d_j^2) \prod_{i \neq j} (d_i^2 - d_j^2)^{-1} \quad \text{for} \quad 1 \leq j \leq n. \]

**Proof.** This theorem is a direct corollary of the results in section 3.

**Remark 16.** This theorem does not cover all solutions of the form

\[ A(\alpha, x) = \sum_{j=1}^{n} -2\gamma'_1(x)e^{-2\alpha y_j(x)}. \]

Consider an example from [1],

\[ A(\alpha) = -\frac{c_0}{k_0} e^{2\alpha k_0} + \frac{c_0}{k_0} e^{-2\alpha k_0}, \quad (61) \]

\[ q(x) = -2 \frac{d^2}{dx^2} \ln \left[ 1 + \frac{c_0}{k_0} \int_{0}^{x} \sinh^2 \left( k_0 y \right) dy \right]. \quad (62) \]

Working through the procedure in section 3, we get

\[ \gamma'_1(0) = -\gamma'_2(0) = -k_0, \]

\[ \gamma'_1(0) = -\gamma'_2(0) = \frac{c_0}{2k_0}, \]

\[ c_1(0) = 0, \quad c'_1(0) = 0, \]

\[ c_2(0) = -k_0^2, \quad c'_2(0) = c_0. \]

Here the values of the constants are:

\[ \sigma_1 = \mu_1 = 2k_0^2 \]

\[ \sigma_2 = \mu_2 = k_0^4 \]

\[ \kappa_1^2 = \kappa_2^2 = k_0^2 \]

so that \( \kappa_1^2 \) and \( \kappa_2^2 \) are not distinct.

This leads to \( O_i = 0, E_i = 0 \). Proposition 10 holds, but the nonlinear transformation (46) is not defined.
5. Conclusion

A large class of exact Equations to A-Equation was found in this work. Techniques used in our approach include non-linear transformation between coefficient of a polynomial and its zero, constants of motion, and an interesting integrating factor method. The nonlinear system studied here is of interest not only for its connection to inverse problems. It represents a larger category of integrable system than C-integrable system and is worth further investigation.

Acknowledgement

A special thanks goes to reviewers for their valuable suggestions.

References

[1] Gesztesy, F. and Simon, B. (2000) A New Approach to Inverse Spectral Theory. II. General Real Potentials and the Connection to the Spectral Measure. *Annals of Mathematics*, **152**, 593-643. [https://doi.org/10.2307/2661393](https://doi.org/10.2307/2661393)

[2] Ramm, A. and Simon, B. (2000) A New Approach to Inverse Spectral Theory. III. Short-Range Potentials. *Journal d'Analyse Mathématique*, **80**, 319-334. [https://doi.org/10.1007/BF02791540](https://doi.org/10.1007/BF02791540)

[3] Simon, B. (1999) A New Approach to Inverse Spectral Theory. I. Fundamental Formalism. *Annals of Mathematics*, **150**, 1029-1057. [https://doi.org/10.2307/121061](https://doi.org/10.2307/121061)

[4] Gesztesy, F. (2007) Inverse Spectral Theory as Influenced by Barry Simon. *Proceedings of Symposia in Pure Mathematics*, American Mathematical Society, **76/2**, 741-820.

[5] Remling, C. (2003) Inverse Spectral Theory for One-Dimensional Schrödinger Operators: The A Function. *Mathematische Zeitschrift*, **245**, 597-617. [https://doi.org/10.1007/s00209-003-0559-2](https://doi.org/10.1007/s00209-003-0559-2)

[6] Remling, C. (2002) Schrödinger Operators and de Branges Spaces. *Journal of Functional Analysis*, **196**, 323-394. [https://doi.org/10.1016/S0022-1236(02)00007-1](https://doi.org/10.1016/S0022-1236(02)00007-1)

[7] Zhang, Y. (2006) Solvability of a Class of Integro-Differential Equations and Connections to One Dimensional Inverse Problems. *Journal of Mathematical Analysis and Applications*, **321**, 286-298. [https://doi.org/10.1016/j.jmaa.2005.08.016](https://doi.org/10.1016/j.jmaa.2005.08.016)

[8] Calogero, F. (1986) A Class of Solvable Dynamical Systems. *Physica*, **18D**, 280-302. [https://doi.org/10.1016/0167-2789(86)90189-2](https://doi.org/10.1016/0167-2789(86)90189-2)

[9] Calogero, F. (1994) A Class of C-Integrable PDEs in Multidimensions. *Inverse Problems*, **10**, 1231-1234. [https://doi.org/10.1088/0266-5611/10/6/004](https://doi.org/10.1088/0266-5611/10/6/004)

[10] Varley, E. and Seymour, B.R. (1988) A Method of Obtaining Exact Solutions to Partial Differential Equations with Variable Coefficients. *Studies in Applied Mathematics*, **78**, 183-205. [https://doi.org/10.1002/sapm1988783183](https://doi.org/10.1002/sapm1988783183)

[11] Varley, E. and Seymour, B.R. (1998) A Simple Derivation of the N-Soliton Solutions to the Korteweg-Devries Equation. *SIAM Journal on Applied Mathematics*, **58**, 904-911. [https://doi.org/10.1137/S0036139996303270](https://doi.org/10.1137/S0036139996303270)

[12] Laidacker, M.A. (1969) Another Theorem Relating Sylvester's Matrix and the Greatest Common Divisor. *Mathematics Magazine*, **42**, 126-128. [https://doi.org/10.2307/2689124](https://doi.org/10.2307/2689124)
Approximate and Invariant Solutions of a Mathematical Model Describing a Simple One-Dimensional Blood Flow of Variable Density

Ranis Ibragimov¹, Nail Ibragimov²,³, Pirooz Mohazzabi¹

¹Department of Mathematics and Physics, University of Wisconsin-Parkside, Kenosha, WI, USA
²Center ALGA, Department of Mathematics and Natural Sciences, Blekinge Institute of Technology, Karlskrona, Sweden
³Laboratory, “Group Analysis of Mathematical Models in Natural and Engineering Sciences,” Ufa State Aviation Technical University, Ufa, Russia
Email: ibragimo@uwsp.edu

Abstract

We examine governing equations that could be used to model a one-dimensional blood flow within a pulsating elastic artery that is represented by a tube of varying cross-section. The model is considered from two perspectives. The first represents a singular perturbation theory providing explicit approximate solutions to the model, and the second is based on group theoretical modeling that provides exact solutions in implicit form. The main goal is to compare these two approaches and lay out the advantages and disadvantages of each approach.

Keywords

Blood Flow, Variable Density, Approximate, Invariant, Solution

1. Introduction

The mathematical modelling and the numerical simulations have become important tools for better understanding of the human cardiovascular system in recent years. One of the main goals in investigating the flow in the aortic system is to understand atherosclerosis and the related phenomena as well as their dependence on a blood flow structure. In particular, the aorta and arteries have a low resistance to blood flow compared with the arterioles and capillaries. When the ventricle contracts, a volume of blood is rapidly ejected into the arterial vessels. Since the outflow to the arteriole is relatively slow because of their high re-
sistance to flow, the arteries are inflated to accommodate the extra blood volume. During diastole, the elastic recoil of the arteries forces the blood out into the arterioles. Thus, the elastic properties of the arteries help to convert the pulsatile flow of blood from the heart into a more continuous flow through the rest of the circulation. Hemodynamics is a term used to describe the mechanisms that affect the dynamics of blood circulation [1] [2] [3]. In reality, an accurate model of blood flow in the arteries would include the following realistic features: 1) the flow is pulsatile, with a time history containing major frequency components up to the eighth harmonic of the heart period; 2) the arteries are elastic and tapered tubes; 3) the geometry of the arteries is complex and includes tapered, curved, and branching tubes and 4) in small arteries, the viscosity depends upon vessel radius and shear rate [4]. Such a complex model has never been accomplished. But each of the features above has been “isolated,” and qualitative if not quantitative models have been derived. As is so often the case in the engineering analysis of a complex system, the model derived is a function of the major phenomena one wishes to illustrate.

Our goal is to model and examine the general trend of possible solutions associated with the governing equations describing a simple one-dimensional blood flow that would depict a blood progressing within a thin and elastic pulsating artery. In reality, for many flow situations, the changes of density due to changes in pressure associated with the flow are very small but not zero. In our simulations, the density is assumed variable for the following reason: hereafter we treat blood not as a homogeneous fluid but a suspension of particles (red cells, white cells, platelets) in fluid called plasma. Blood particles must be taken into account in the rheological model in smaller arterioles and capillaries since their size becomes comparable to that of the vessel [5] [6] [7] [8]. In particular, as has been discussed in [9], red blood cells (RBCs) exhibit a unique deformability, which enables them to change shape reversibly in response to an external force. Human RBCs have the ability to undergo large deformations when subjected to external stresses, which allows them to pass through capillaries that are narrower than the diameter of a resting RBC. In fact, RBCs are more deformable than any other biomaterial. RBCs are biconcave discs, typically 6 - 8 μm in diameter and 2 μm thick, and their deformation can involve a change in cell curvature, a uniaxial deformation, or an area expansion. In mammals, RBCs are non-nucleated and consist of a concentrated hemoglobin solution enveloped by a highly flexible membrane. The deformability of RBCs plays an important role in their main function, the transport of gases (O₂ and CO₂) via blood circulation (see also [10] and [9]).

To put the deformability of blood due to pressure in perspective, consider a multi-component system of total volume $V$, with

$$V = \sum_i V_i$$

(1)

where $V_i$ is the subvolume of component $i$ in the system. The (isothermal) compressibility of the system is
\[
\kappa = -\frac{1}{V} \left( \frac{\partial V}{\partial P} \right)_T = -\frac{1}{V} \sum \left( \frac{\partial V_i}{\partial P} \right)_T
\]

(2)

But the compressibility of each component is

\[
\kappa_i = -\frac{1}{V_i} \left( \frac{\partial V_i}{\partial P} \right)_T
\]

(3)

Therefore, (2) reduces to

\[
\kappa = \frac{1}{V} \sum_i V_i \kappa_i
\]

(4)

Finally, denoting the volume fraction of each component of the system by \( \alpha_i \), we have

\[
\kappa = \sum_i \alpha_i \kappa_i
\]

(5)

In our case, the main contribution to the compressibility and deformability of the blood is coming from RBCs.

The method of obtaining general solution of the governing equations for the given model is considered from two prospective points of view. The first approach represents a singular perturbation theory providing explicit approximate solutions to the model and the second one is based on group theoretical modeling that provides the exact solutions written in an implicit form. The main goal is to compare these two approaches and fetch out the efficiency and deficiency of each proposed approach.

The range of developed models or models being developed extends from lumped models to complicated three-dimensional fluid-structure models [11] and [12]. In this article we consider a simple one-dimensional model of blood flow in a vessel. The blood flow in the vessel is described by this and generally by all one-dimensional models is not suitable for describing blood flow in complicated morphological regions as stenosis or bifurcations. However, these situations can also be covered to certain extent and, from one hand, can be used as an alternative to the more complex three-dimensional fluid-structure models or in conjunction with them in a geometrical multiscale fashion, as explained in [13]. On the other hand computational complexity of one-dimensional models is several orders of magnitude lower than that of multi-dimensional models. Few decades ago, a multi scale approach has attracted wider interest. Namely, in a multiscale approach, one-dimensional models may be coupled on the one hand with lumped-parameter models [12] based on a system of ordinary differential equations [11] [14], or to three-dimensional fluid-structure models, as discussed in [15] and [16]. In the latter case they may also provide a way of implementing more realistic boundary conditions for 3D calculations; or, they can be used for the numerical acceleration of a three-dimensional Navier-Stokes solver in a multilevel-multiscale scheme. Additionally, one-dimensional models give a good description of the propagation of pressure waves in arteries [17] [18], hence they can be successfully used to investigate the effects on pulse waves of the geometrical and mechanical arterial modification, due e.g. to the presence of stenoses,
or to the placement of stents or prostheses [13].

2. Modeling

In order to describe a problem in mathematical terms, one must make use of the basic laws that govern the elements of the problem. Within the frame of the present modeling, we start with conservation laws for mass and momentum and consider a perfect compressible fluid propagating along a tube with longitudinal coordinate \( x \) and slowly varying cross-section \( a(x,t) \). Because of the pressure gradient in the blood, the artery wall must deform. The elastic restoring force in the wall makes it possible for waves to propagate. In terms of one-dimensional modeling, we assume that the artery radius \( r(x,t) \) varies from the constant mean \( r_0 \) in time and along the artery (in \( x \)). We denote the local cross sectional area be \( a = \pi r^2 \), and the averaged velocity be \( v(x,t) \). Consider a fixed geometrical volume between \( x \) and \( x + dx \), through which fluid moves in and out. Conservation of mass requires

\[
\frac{\partial a}{\partial t} + \frac{\partial (va)}{\partial t} = 0. \tag{6}
\]

We next assume that the time rate of momentum change in the volume is balanced by the net influx of momentum through the two ends and the pressure force acting on all sides. The rate of momentum change \( M \) is given by

\[
M = \frac{\partial (\rho va)}{\partial t}, \tag{7}
\]

where \( \rho(x,t) \) is the density of fluid associated with the mixture density of the blood consisting of blood plasma and red blood cells (RBC). In reality, RBC fraction may include large viscosity variations, stressing the importance of accounting for the non-Newtonian effects (see e.g. [19], where the Quemada viscosity model [20] is used to account for the non-Newtonian viscosity behavior).

The net rate of momentum influx is

\[
-\frac{\partial (\rho v a)}{\partial x} \, dx = -\rho v \frac{\partial (va)}{\partial x} - \rho va \frac{\partial v}{\partial x}. \tag{8}
\]

The net pressure force at the two ends is given by \(-\partial(pa)/\partial x\) while that on the sloping wall is

\[
2\pi r p \frac{\partial r}{\partial x} = p \frac{\partial a}{\partial x}. \tag{9}
\]

The sum of all pressure forces \( P \) is given by

\[
P = -a \frac{\partial p}{\partial x}. \tag{10}
\]

Balancing the momentum by equating \( M \) given by (7) to the sum of (8) and \( P \) given by (10) we get, after making use of mass conservation (6),

\[
\frac{\partial v}{\partial t} + v \frac{\partial v}{\partial x} = -\frac{1}{\rho} \frac{\partial p}{\partial x}. \tag{11}
\]

Arterial pulse propagation varies along the circulatory system as a result of the
complex geometry and nonuniform structure of the arteries. In order to learn
the basic facts of arterial pulse characteristics, we assume an idealized case of an
infinitely long circular elastic tube that contains a slightly compressible blood,
which is a suspension of particles in what’s basically water. As such, it’s com-
pressibility will be mainly due to the RBC, as explained above. We can think of it
as a two-phase homogeneous nonviscous fluid flow of water and gas bubbles. If
we apply pressure to the water/gas mixture the overall density will decrease as
the gas compresses, leading to the mixture continuity equation that, under the
assumption of zero relative velocity, reduces the equivalent single phase flow of
density \( \rho \) [8]:

\[
\frac{\partial p}{\partial t} + \frac{\partial}{\partial x}(\rho v) = 0.
\]  

In addition, empirical constitutive laws are needed to relate pressure and den-
sity such as equations of state

\[
p = p(\rho, S),
\]  

where \( S \) denotes the entropy. Since in our modeling no temperature gradient is
imposed externally and the gradient of the flow is not too large, we ignore ther-
dal diffusion. The fluid motion is the adiabatic; entropy \( S = S_0 \) is constant. As
a result, \( p = p(\rho, S_0) \) depends only on density. As we have from thermody-
namics,

\[
\left( \frac{\partial p}{\partial \rho} \right)_S = \gamma \left( \frac{\partial p}{\partial \rho} \right)_T,
\]  

where \( T \) is the temperature and \( \gamma \) is the ratio of specific heats at constant
pressure and constant volume. Furthermore, since pressure is a function of den-
sity only, we can write \( p = p(\rho) \). Expanding this function in a Taylor series
about the equilibrium density \( \rho_0 \), we have

\[
p = p_0 + p'(\rho_0) (\rho - \rho_0) + \frac{p''(\rho_0)}{2!} (\rho - \rho_0)^2 + \cdots
\]  

where \( p_0 \) is the equilibrium pressure at which \( \rho = \rho_0 \). Since \( \rho - \rho_0 \) is small,
we can neglect the second- and higher-order terms and write

\[
p = p_0 + \lambda (\rho - \rho_0)
\]  

where \( \lambda \) is a constant. From this equation it follows that

\[
\frac{\partial p}{\partial \rho} = \lambda \frac{\partial \rho}{\partial x} \quad \text{and} \quad \frac{\partial p}{\partial t} = \lambda \frac{\partial \rho}{\partial t}
\]  

Since the force due to gravity is neglected, combining (11), (12) and (17), we
arrive at the governing equations of motion for unknowns velocity \( v(x,t) \),
pressure \( p(x,t) \) and density \( \rho(x,t) \) that are written as follows:

\[
\frac{\partial v}{\partial t} + v \frac{\partial v}{\partial x} = - \frac{1}{\rho} \frac{\partial p}{\partial x},
\]  

\[
\frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho v) = 0,
\]
in which \( t \) is time, \( x \) is a spatial variable and \( \lambda \) is a constant.

We eliminate the pressure from these equations by differentiating the Equation (18) with respect to \( t \) and using the equation of state (20) to get

\[
\frac{\partial^2 v}{\partial t^2} + \frac{\partial v}{\partial x} \frac{\partial v}{\partial t} + v \frac{\partial^2 v}{\partial x \partial t} = -\frac{\lambda}{\rho} \left( \frac{\partial^2 \rho}{\partial x \partial t} - \frac{1}{\rho} \frac{\partial \rho}{\partial t} \frac{\partial \rho}{\partial x} \right).
\]

(21)

Using Equation (19), we can rewrite (21) as

\[
\frac{\partial^2 v}{\partial t^2} + \frac{\partial}{\partial x} \left( v \frac{\partial v}{\partial t} \right) = \frac{\lambda}{\rho} \frac{\partial^2 \rho}{\partial x^2} + \frac{\partial}{\partial x} \left( \frac{\lambda}{\rho} \frac{\partial \rho}{\partial x} \right).
\]

(22)

As follows from Equations (18) and (20), we have

\[
\frac{\lambda v}{\rho} \frac{\partial \rho}{\partial x} = -v \frac{\partial v}{\partial t} - v^2 \frac{\partial v}{\partial x}.
\]

(23)

Substituting this result into (22), we arrive at the following single equation for \( v(x,t) \):

\[
\frac{\partial^2 v}{\partial t^2} + \frac{\partial}{\partial x} \left( 2v \frac{\partial v}{\partial t} + v^2 \frac{\partial v}{\partial x} \right) = \frac{\lambda}{\rho} \frac{\partial^2 v}{\partial x^2}.
\]

(24)

3. First Approach: Approximate Analysis

In order to identify the resonant input in the model, we start with an approximate solution in the form of naive expansion

\[
v(x,t) = v_0 + \sum_{i=1}^{\infty} \varepsilon^i v_i(x,t),
\]

(25)

where \( \varepsilon \) is a small parameter and \( v_0 = \text{const} \) is an exact trivial solution of Equation (24).

3.1. Failure of the Direct Approach

We substitute the expansion (25) into (24) and collect powers of \( \varepsilon \).

Problem 0(\( \varepsilon^1 \)) gives the following equation:

\[
\frac{\partial^2 v_1}{\partial t^2} - \lambda \frac{\partial^2 v_1}{\partial x^2} + \frac{\partial}{\partial x} \left( 2v_0 \frac{\partial v_1}{\partial t} + v_0 \frac{\partial v_1}{\partial x} \right) = 0.
\]

(26)

We look for solution of Equation (26) in the form

\[
v_i(x,t) = A_i \sin \theta_i + B_i \cos \theta_i,
\]

(27)

where \( A_i, B_i \) are constants and \( \theta_i = k_i x - \omega_i t \), in which \( k_i \) and \( \omega_i \) are wave number and frequency of the primary wave. Solution (27) is valid provided the dispersion relation is satisfied:

\[
\omega_i^2 - 2v_0 k_i \omega_i - k_i^2 \left( \lambda - v_0^2 \right) = 0.
\]

(28)

So the 0(\( \varepsilon^1 \)) problems represent a hyperbolic model with two wave modes

\[
\omega_i = k_i \left( v_0 \pm \sqrt{\lambda} \right).
\]

(29)

Problem 0(\( \varepsilon^2 \)) gives the following equation:
\frac{\partial^2 \nu_x}{\partial t^2} - \lambda \frac{\partial^2 \nu_x}{\partial x^2} + \frac{\partial}{\partial x} \left( 2 \nu_0 \frac{\partial \nu_1}{\partial t} + \nu_0^2 \frac{\partial^2 \nu_x}{\partial x^2} \right) = - \frac{\partial}{\partial x} \left( 2 \nu_v \frac{\partial \nu_1}{\partial t} + 2 \nu_v \nu_1 \frac{\partial \nu_1}{\partial x} \right). \quad (30)

In view of presentation (27), the right hand side in Equation (30) is written as

\begin{equation}
- \frac{\partial}{\partial x} \left( 2 \nu_v \frac{\partial \nu_1}{\partial t} + 2 \nu_v \nu_1 \frac{\partial \nu_1}{\partial x} \right) = \delta \left( A_i^2 - B_i^2 \right) \cos (2\theta_i) - 2 \delta \lambda A_i B_i \sin (2\theta_i), \quad (31)
\end{equation}

where we denote

\begin{equation}
\delta = 2k_i (\omega_i - v_i k_i). \quad (32)
\end{equation}

We look for particular solution in the form

\begin{equation}
v_2^{(p)} = H \cos (2\theta_i) + R \sin (2\theta_i). \quad (33)
\end{equation}

Substituting this solution into (31), we obtain the following expressions for \( H \) and \( R \):

\begin{equation}
(H, R) = \frac{\delta \left( B_i^2 - A_i^2, 2A_i B_i \right)}{4 \left[ \omega_i^2 - 2 \nu_i k_i \omega_i - k_i^2 (\lambda - \nu_i^2) \right]}. \quad (34)
\end{equation}

As expected, because of the dispersion relation (28), the right hand side of Equation (30) corresponds to resonance and yields the secular terms.

In particular, if we look for particular solution of the form

\begin{equation}
v_2^{(p)} = t \tilde{H} \cos (2\theta_i) + t \tilde{R} \sin (2\theta_i), \quad (35)
\end{equation}

the resonance input would be removed since, in this case, the constants \( \tilde{H} \) and \( \tilde{R} \) would be

\begin{equation}
(\tilde{H}, \tilde{R}) = -k_i \left( A_i B_i, \frac{A_i^2 - B_i^2}{2} \right) \quad (36)
\end{equation}

and so the particular solution would have the form

\begin{equation}
v_2^{(p)} = -kt A_i B_i \cos (2\theta_i) - \frac{k_i}{2} t \left( A_i^2 - B_i^2 \right) \sin (2\theta_i). \quad (37)
\end{equation}

Since \( v_2^{(p)} \) grows linearly in time, the term \( e^{\nu_1} \) would become comparable to \( e^{\nu_1} \) for large values of time \( t \) (e.g. when time is of order \( \frac{1}{\varepsilon} \)), as shown in Figure 1.

In particular, Figure 1 shows the qualitative behavior of the time series of the second order approximation of solution with secular terms (37) for the values \( k = 1 \) and 2 and \( \lambda = 1 \) and 2. The following values of parameters have been chosen: \( \varepsilon = 0.1, \nu_0 = 1, x = 2, A_i = 0.18, B_i = 0.19 \) and \( \omega_i \) is determined by the dispersion relation (29). Since we are interested only in general solutions, the choice of constants is arbitrary and we are focused on qualitative analysis.

As seen from (37), the first terms of the expansion (25) provide a local (small \( t \)) approximation, at most. The shortcoming of (25) is related to the breakdown of the straightforward approach on nonlinear perturbation analysis of equation (24), but is more transparent to explanations. The nonlinear terms in (24) will slowly, but accumulative, absorb energy and damp the motion. Hence, even
though the term $\varepsilon v_2^{(p)}$ itself is small the long term effects are crucial and the solution cannot be described as being periodic plus a small correction. The consequence for a naive expansion (25) is that the ordering requirement $v_0 > \varepsilon v_1 > \cdots$ is violated. However, it may be instructive to try and fail in order to understand the nature of the resonance phenomena.

### 3.2. Multiple Scale Approach

We introduce the latter scale according to the new variable

$$\tau = \varepsilon t.$$  

(38)

We now consider the fast scale $t$, and the slow scale $\tau$, as independent variables. We rewrite Equation (24) in terms of the new variable (38) and modify the series expansion (25) to the form

$$v(x,t,\tau) = v_0 + \sum_{i=1}^{\infty} \varepsilon^i v_i(x,t,\tau),$$

(39)

which yields the perturbation hierarchy similar to (26) and (30), i.e.

**Problem 0($\varepsilon^1$):**

$$\frac{\partial^2 v_0}{\partial t^2} - \lambda \frac{\partial^2 v_0}{\partial x^2} + \frac{\partial}{\partial x} \left( 2v_0 \frac{\partial v_0}{\partial t} + v_0^2 \frac{\partial v_0}{\partial x} \right) = 0$$

and

**Problem 0($\varepsilon^2$):**

$$\frac{\partial^2 v_0}{\partial t^2} - \lambda \frac{\partial^2 v_0}{\partial x^2} + \frac{\partial}{\partial x} \left( 2v_0 \frac{\partial v_0}{\partial t} + v_0^2 \frac{\partial v_0}{\partial x} \right) = F + Q,$$

(40)

where $F$ represents the right hand side if Equation (30) and

$$Q = -2 \left[ \frac{\partial^2 v_0}{\partial \tau \partial t} - v_0 \frac{\partial^2 v_0}{\partial x \partial t} \right]$$

(41)
appears because of scaling (38). Since the derivative with respect to the fast variable appears only at $0(\varepsilon^2)$, the problem at $0(\varepsilon^1)$ is identical to Equation (26). The slow time variable $\tau$ is implicit in the constant of integration and the most general real-valued solution of the $0(\varepsilon^1)$ problem can be written as

$$v(x,t,\tau) = A_1(\tau)e^{i\theta} + A'_1e^{-i\theta}. \quad (42)$$

With this solution in hand, Equation (40) reads

$$F + Q = 2(\omega_i - v_0k_i)\left[ i\left(\frac{dA_1}{d\tau}e^{i\theta} - \frac{dA'_1}{d\tau}e^{-i\theta}\right) - 2k_i\left(A_1^2e^{2i\theta} + A'_1^2e^{-2i\theta}\right)\right]. \quad (43)$$

To avoid secular terms we must require

$$\left(\frac{dA_1}{d\tau} + \frac{dA'_1}{d\tau}\right)\sin\theta = -2k_i\left(A_1^2 + A'_1^2\right)\cos(2\theta), \quad (44)$$

$$\left(\frac{dA_1}{d\tau} - \frac{dA'_1}{d\tau}\right)\cos\theta = 2k_i\left(A_1^2 - A'_1^2\right)\sin(2\theta). \quad (45)$$

The natural choice is to set $A'_1 = 0$. Then, squaring and adding the resulting equations for $A_1$, we arrive at a single equation

$$\frac{dA_1}{d\tau} = 2k_iA_1^2. \quad (46)$$

Solving Equation (46), we write the solution in the form

$$v(x,t) = v_0 + \varepsilon\cos k_ix - \frac{\omega_1t}{2k_i\varepsilon t + c_i} + 0(\varepsilon^2), \quad (47)$$

where $c_i$ is a constant of integration and $\omega_1$ is related to $k_i$ and $\lambda$ by the dispersion relation (28), i.e.

$$\omega_1 = k_i\left(v_0 \pm \sqrt{\lambda}\right).$$

For the purpose of visualization, Figure 2 is used to compare the qualitative

![Figure 2. Visualization of the approximate solution $v(x,t)$](image-url)
behavior of the time series of the of solution (47) for the values of \( k_1 = 1 \) and \( k_2 = 4 \) when \( c_1 = 1.5 \) and for the same values of parameters \( \epsilon, v_0, x \) and \( A \) as we used above to visualize the time series of \( v_2^{(p)} \) shown in Figure 1.

A question of particular interest is the investigation of asymptotic stability of the trivial solution \( v_0 \). This will be done in the next section.

3.3. Stability of Perturbed Steady Flows

We note that the stationary solution

\[ v = U(x) \]  

solves Equation (24) in the stationary case, i.e.

\[ \frac{\partial}{\partial x} \left( \frac{\partial}{\partial x} v^2 \right) = \lambda \frac{\partial^3 v}{\partial x^3}, \]  

which can be integrated to give the exact solution of the form

\[ U^3(x) - 3 \lambda U(x) = M_1 x + M_2, \]  

where \( M_1 \) and \( M_2 \) are constants. We denote

\[ \sigma = 4M_1 x + 4M_2 + 4\sqrt{(M_1 x)^2 + 2M_1 M_2 x - 4\lambda^3 + M_2^2} \]  

Then the only real branch of the solution for \( U(x) \) of Equation (49) can be written explicitly as

\[ U(x) = \frac{1}{2} \left( \frac{\sigma^3 + 4\lambda}{\sigma} \right)^{\frac{1}{3}}. \]  

Figure 3 is used to visualize the stationary flow \( U(x) \) given by (52) for different values of the constant \( \lambda = \lambda_1, \lambda = \lambda_2 \) and the following fixed values of parameters: \( \lambda_1 = 5, \lambda_2 = 1, M_1 = 5 \) and \( M_2 = 1 \).

Figure 3. Visualization of the stationary solution \( U(x) \).
Since, as Figure 2 shows, \( U(x) \) is growing linearly in \( x \), we classify \( U(x) \) as non-physical solution.

Let us now look for a nonstationary solution of Equation (24) that is close to \( U(x) \) in the form (see e.g. [21] [22])

\[
v(x,t) = U(x) + \varepsilon \tilde{v}(x,t),
\]

where \( \varepsilon \) is a small parameter and \( \tilde{v} \) denotes the perturbation. This procedure is largely formal. Mathematics ideal requires proof that the solution of the complete equations in question for \( \varepsilon \to 0 \) has a solution of the approximate equations at zeroth order of \( \varepsilon \) (at least asymptotically). In fact, this ideal is achieved in very rare cases; researchers are usually limited to the formal construction of an approximate model. The justification is based on physical intuition which opens a wide scope. It is clear that, at the same time, the role of the criterion of practice is greatly increased.

We assume a perturbation of the form of a plane harmonic wave propagating in the positive \( x \) direction,

\[
\tilde{v}(x,t) = Ae^{i(kx-\omega t)},
\]

where \( A \) is a constant amplitude, \( k \) is a wave number and \( \omega \) is the angular frequency of the oscillations. Substituting the presentations (53) and (54) into Equation (24) and collecting the terms of the order \( \varepsilon^0 \), we get the nonlinear equation for the mean flow

\[
\frac{\partial}{\partial x} \left( U^2 \frac{\partial U}{\partial x} \right) - \lambda \frac{\partial^2 U}{\partial x^2} = 0,
\]

which coincides with (49) and thus \( U(x) \) has the form (52).

Similarly, collecting and separating the real and imaginary parts of the terms of the order of \( \varepsilon^0 \), we get the equations

\[
\omega^2 - 2U \omega k + k^2 (U - \lambda) - 2 \frac{\partial}{\partial x} \left( U \frac{\partial U}{\partial x} \right) = 0
\]

and

\[
\frac{\omega}{k} \frac{\partial}{\partial x} \left[ U (1 - U) \right] = 0.
\]

For progressing wave like solution, Equations (56) and (57) implies that there is another particular exact solution of Equation (49) (and, correspondingly, of Equation (24))

\[
v(x,t) = v_0 = \text{const},
\]

provided that \( \omega \) and \( k \) satisfy the dispersion relation (28), i.e.

\[
\omega^2 - 2v_0 \omega k + (v_0 - \lambda) k^2
\]

with two known wave modes given by (29). As one can expect, since the flow is away of frictional boundaries, the dispersion relation (28) confirms asymptotic stability of the mean constant flow (58).

Figure 4 shows snap-shots of the perturbed flow \( v(x,t) = v_0 + \varepsilon \tilde{v}(x,t) \) at
Figure 4. Visualization of the perturbed constant flow $v_0$ for different values of wave number $k$.

initial time $t = 0$ (red line) and later times of $t = 0.5, 1.0, 1.5$ and $1.8$ units (plotted in different colors) for different values of the wave number $k$ and the following fixed values of parameters: $v_0 = 1$ (for $k = 1$), $v_0 = 2$ (for $k = 2$), $v_0 = 3$ (for $k = 10$), and $v_0 = 4$ (for $k = 25$), $a = 1$ and $\varepsilon = 0.1$.

4. Second Approach: Group Theoretical Point of View

Detailed presentations of the theory of symmetries and invariant solutions of differential equations can be found elsewhere [23] [24] [25] [26]. For convenience, we summarize the basic notation from calculus of Lie group analysis in Appendix, which represents a simplified version of the overview of basic concepts of Lie symmetry groups.

A simple inspection shows that Equation (24) admits the one-parameter groups of translations

$$\bar{T} = t + a_t, \quad \bar{x} = x + a_x$$

of $t$ and $x$ and the one-parameter group of uniform scaling transformations in the $(t,x)$-plane:

$$\bar{T} = te^{a_t}, \quad \bar{x} = xe^{a_x}.$$  

The above transformations groups have the following generators (called also infinitesimal symmetries):

$$X_1 = \frac{\partial}{\partial t}, \quad X_2 = \frac{\partial}{\partial x}, \quad X_3 = t\frac{\partial}{\partial t} + x\frac{\partial}{\partial x} \quad (60)$$

Various linear combinations of the generators (60) can serve for constructing group-invariant solutions of Equation (60). A brief description of group invariant solutions and illustrative examples are given in [27].
4.1. Traveling Waves

Traveling waves are invariant solutions with respect to any linear combination of the translation generators $X_1$ and $X_2$. We take the linear combination in the form

$$X_1 + mX_2 = \frac{\partial}{\partial t} + m \frac{\partial}{\partial x}, \quad m = \text{const.} \quad (61)$$

The operator (61) has two independent invariants, $v$ and $z = x - mt$. According to the theory of invariant solution (see, e.g. [27], Section 7.2), the invariant solution has the representation

$$v = f(z). \quad (62)$$

The representation (62) reduces Equation (24) to the ordinary differential equation

$$\left(m^2 - \lambda\right) f'' + f' - (2mf') = 0.$$ 

We integrate it once and obtain

$$\left(m^2 - \lambda\right) f'' + f' - 2mf' = M_1, \quad M_1 = \text{const.}$$

The second integration gives the cubic equation

$$f^3 - 3mf^2 + 3\left(m^2 - \lambda\right)f = M_1\mu + M_2 \quad (63)$$

for determining $f(\mu)$. Thus, the traveling wave solution (62) is determined by the cubic Equation (63) and involves three arbitrary parameters $m, M_1, M_2$. Cardan’s solution for the cubic equation (see, e.g. [27], Section 1.1.1) allows to express the traveling waves in radicals.

Remark: In the special case $m = 0$ in (61), Equations (62) and (63) give the stationary solution $v = U(x)$ given explicitly by the cubic Equation (50).

4.2. Similarity Solution

The invariant solution with respect to the generator

$$X_3 = t \frac{\partial}{\partial t} + x \frac{\partial}{\partial x} \quad (64)$$

of the uniform scaling transformation group is known as a similarity solution. The operator (64) has two independent invariants, $v$ and

$$\xi = \frac{x}{t}. \quad (65)$$

The invariant solution has the representation

$$v = g(\xi). \quad (66)$$

Calculations show that the representation (66) reduces Equation (24) to the second-order ordinary differential equation

$$\left[(g - \mu)^2 - \lambda\right] g'' + 2(g - \mu)g' - 2(g - \mu)g' = 0.$$

Rewriting this equation in the form
and integrating once, we arrive at the first-order equation
\[
\left( (g - \mu)^2 - \lambda \right) g' = C_1, \quad C_1 = \text{const.}
\]  
(68)

We transform Equation (68) into separable form
\[
(w^2 - \lambda) \frac{dw}{d\mu} = -w^2 + \lambda + C_1
\]  
(69)

by introducing the new dependent variable
\[
w = g - \mu.
\]  
(70)

Separating the variables,
\[
\frac{w^2 - \lambda}{\lambda + C_1 - w^2} dw = d\mu,
\]
we write Equation (69) as
\[
-w + \frac{C_1}{\lambda + C_1 - w^2} dw = d\mu,
\]
whence upon integration
\[
C_1 \int \frac{dw}{\lambda + C_1 - w^2} = w + \mu.
\]  
(71)

Evaluating the integral in (71) we obtain the following cases.

**Case 1:** \(\mu + C_1 = 0\). Then the integration in Equation (71) gives
\[
-\frac{\lambda}{w} = w + \mu + C_2,
\]
whence
\[
w = \frac{1}{2} \left[ -\left( \mu + C_1 \right) \pm \sqrt{\left( \mu + C_1 \right)^2 - 4\lambda} \right].
\]  
(72)

**Case 2:** \(\mu + C_1 < 0\). Then we evaluate the integral by writing
\[
C_1 \int \frac{dw}{\lambda + C_1 - w^2} = -C_1 \int \frac{dw}{-\left( \lambda + C_1 \right) + w^2}
\]
and write Equation (71) as
\[
-w + \frac{C_1}{\sqrt{-\left( \lambda + C_1 \right)}} \arctan \frac{w}{\sqrt{-\left( \lambda + C_1 \right)}} = w + \mu + C_2.
\]  
(73)

**Case 3:** \(\mu + C_1 > 0\). In this case Equation (71) becomes
\[
\frac{C_1}{2\sqrt{\lambda + C_1}} \ln \frac{\sqrt{\lambda + C_1} + w}{\sqrt{\lambda + C_1} - w} = w + \mu + C_2
\]  
(74)

when \(w^2 < \mu + C_1\), and
\[
\frac{C_1}{2\sqrt{\lambda + C_1}} \ln \frac{w + \sqrt{\lambda + C_1}}{w - \sqrt{\lambda + C_1}} = w + \mu + C_2
\]  
(75)

when \(w^2 > \mu + C_1\).
Based on the above calculations we conclude that all similarity solutions based on the infinitesimal symmetry (64) are provided by Equations (65), (66), (70) and (72)-(75).

5. Conclusion

We have investigated a nonlinear partial differential equation of second order that could be used to model a simple one-dimensional blood flow inside a tube of varying cross-section. This model can be an approximation for a pulsating elastic artery. We have proposed two different points of view. The first approach represents a singular perturbation theory that formalizes the scale-separation property by explicitly defining multiple scales that exist in the given nonlinear model with the goal of separating derivatives with respect to fast and slow scales into different orders of perturbation theory. The advantage of this approach is that it yields a solvable perturbative hierarchy of equations that provides useful perturbative information already at low orders in \( \epsilon \). However, the disadvantage of this approach is the need to identify the various scales \textit{a priori} and, in the frame of the present modeling, multiple scale approach cannot be brought beyond the leading order. Alternatively, group theoretical approach provides \textit{all possible} exact solutions of the nonlinear model (24) without any perturbations and, consequently, without introducing a small parameter \( \epsilon \), which is a significant advantage. In this article, we have provided the exact solutions that were obtained implicitly by solving the nonlinear ordinary differential equations, which have the deficiency of the latter approach. However, in terms of numerical simulations, the second approach seems more advantageous.

Acknowledgements

This research was supported in part by a Summer Faculty Fellowship Grant provided in Spring, 2017 by College of Natural and Health Sciences, University of Wisconsin-Parkside.

References

[1] Einav, S. and Stolero, D. (1987) Pulsatile Blood Flow in an Arterial Bifurcation. \textit{Medical & Biological Engineering & Computing}, 25, 12-17. \( \text{https://doi.org/10.1007/BF02442814} \)

[2] Elad, D. and Einav, S. (1989) Simulation of Airway Closure during Forced Vital Capacity. \textit{Annals of Biomedical Engineering}, 17, 617-631. \( \text{https://doi.org/10.1007/BF02367466} \)

[3] Rotman, O., Shav, D., Raz, S., Zaretsky, U. and Einav, S. (2013) Biomechanical Aspects of Catheter-Related Thrombophlebitis. \textit{Journal of Biomedical Science and Engineering}, 6, 6-13. \( \text{https://doi.org/10.4236/jbise.2013.612A002} \)

[4] Brand, M., Avrahami, I., Einav, S. and Ryvkin, M. (2014) Numerical Models of Net-Structure Stents Inserted into Arteries. \textit{Computers in Biology and Medicine}, 52, 102-110.

[5] Gummel, E., Milosevic, H., Ragulin, V., Zakharov, Y. and Zimin, Y. (2014) Motion of Viscous in homogeneous Incompressible Fluid of Variable Viscosity. \textit{Zbornik
rabova konferencije MIT 2013, Beograd, 267-274.

[6] Geidarov, N., Zakharov, Y. and Shokin, Y. (2011) Solution of the Problem of Viscous Fluid with a Given Pressure Differential. *Russian Journal of Numerical Analysis and Mathematical Modelling*, 26, 39-48.

[7] Milosevic, H., Gaydarov, N. and Zakharov, Y. (2013) Model of Incompressible Viscous Fluid Flow Driven by Pressure Difference in a Given Channel. *International Journal of Heat and Mass Transfer*, 62, 242-246.

[8] Dolgov, D. and Zakharov, Y. (2015) Modeling of Viscous Inhomogeneous Fluid Flow in Large Blood Vessels. *Vestnik Kemerovo State University*, 2, 30-35. (In Russian)

[9] Kim, J., Ho Yoon, L. and Sehyun, S. (2015) Advances in the Measurement of Red Blood Cell Deformability: A Brief Review. *Journal of Cellular Biotechnology*, 1, 63-79. [https://doi.org/10.3233/JCB-15007](https://doi.org/10.3233/JCB-15007)

[10] Bransky, A., Korin, N., Nemirovski, Y. and Dinnar, U. (2007) Correlation between Erythrocytes Deformability and Size: A Study Using a Microchannel Based Cell Analyzer. *Microvascular Research*, 73, 713.

[11] Rideout, V. and Dick, D. (1967) Difference-Differential Equations for Fluid Flow in Distensible Tubes. *IEEE Transactions on Biomedical Engineering*, 14, 171-177. [https://doi.org/10.1109/TBME.1967.4502495](https://doi.org/10.1109/TBME.1967.4502495)

[12] Quemada, D. (1977) Rheology of Concentrated Disperse Systems and Minimum Energy Dissipation Principle. *Rheologica Acta*, 16, 82-94. [https://doi.org/10.1007/BF01516932](https://doi.org/10.1007/BF01516932)

[13] Formaggia, L., Lamponi, D. and Quarteroni, A. (2003) One Dimensional Models for Blood Flow in Arteries. *Journal of Engineering Mathematics*, 47, 251-276. [https://doi.org/10.1023/B:ENGI.000007980.01347.29](https://doi.org/10.1023/B:ENGI.000007980.01347.29)

[14] Westerhof, N., Bosman, F., Vries, C. and Noordergraaf, A. (1969) Analog Studies of the Human Systemic Arterial Tree. *Journal of Biomechanics*, 2, 121-143.

[15] Formaggia, L., Gerbeau, J., Nobile, F. and Quarteroni, A. (2001) On the Coupling of 3D and 1D Navier-Stokes Equations for Flow Problems in Compliant Vessels. *Computer Methods in Applied Mechanics and Engineering*, 191, 561-582.

[16] Formaggia, L., Nobile, F., Gerbeau, J. and Quarteroni, A. (2002) Numerical Treatment of Defective Boundary Conditions for the Navier-Stokes Equations. *SIAM Journal on Numerical Analysis*, 40, 376-401. [https://doi.org/10.1137/S003614290038296X](https://doi.org/10.1137/S003614290038296X)

[17] Streeter, V., Keitzer, W. and Bohr, D. (1963) Pulsatile Pressure and Flow through Distensible Vessels. *Circulation Research*, 13, 3-20. [https://doi.org/10.1161/01.RES.13.1.3](https://doi.org/10.1161/01.RES.13.1.3)

[18] Phythoud, F., Stergiopulos, N. and Meister, J. (1995) Forward and Backward Waves in the Arterial System: Nonlinear Separation Using Riemann Invariants. *Technology and Health Care*, 3, 201-207.

[19] Wittberg, L., Wyk, S., Fuchs, L., Gutmark, E., Bakeljauw, P. and Little-Gutmark, I. (2016) Effects of Aortic Irregularities on Blood Flow. *Biomechanics and Modeling in Mechanobiology*, 15, 345-360. [https://doi.org/10.1007/s10237-015-0692-y](https://doi.org/10.1007/s10237-015-0692-y)

[20] Quarteroni, A., Tuveri, M. and Veneziani, A. (2000) Computational Vascular Fluid Dynamics Problems, Models and Methods. *Computing and Visualization in Science*, 2, 163-197. [https://doi.org/10.1007/s007910050039](https://doi.org/10.1007/s007910050039)

[21] Ibragimov, R.N. (2007) Oscillatory Nature and Dissipation of the Internal Waves Energy Spectrum in the Deep Ocean. *The European Physical Journal Applied Physics*, 40, 315-334. [https://doi.org/10.1051/epja:2007164](https://doi.org/10.1051/epja:2007164)
[22] Ibragimov, R.N. (2008) Generation of Internal Tides by an Oscillating Background Flow along a Corrugated Slope. *Physica Scripta*, **78**, Article ID: 065801. [https://doi.org/10.1088/0031-8949/78/06/065801](https://doi.org/10.1088/0031-8949/78/06/065801)

[23] Ovsyannikov, L. (2013) Lectures on the Theory of Group Properties of Differential Equations, Novosibirsk University Press, Novosibirsk, 1966. English Trans. Ibragimov, N.H., Higher Education Press, Beijing; World Scientific, Singapore.

[24] Ibragimov, N. (1985) Transformation Groups in Mathematical Physics, Nauka, Moscow, 1983. English Transl.: Transformation Groups Applied to Mathematical Physics, Reidel, Dordrecht.

[25] Olver, P. (1986) Applications of Lie Groups to Differential Equations. Springer-Verlag, New York. [https://doi.org/10.1007/978-1-4684-0274-2](https://doi.org/10.1007/978-1-4684-0274-2)

[26] Bluman, G. and Kumei, S. (1980) Symmetries and Differential Equations. Springer-Verlag, New-York.

[27] Ibragimov, N. (2009) A Practical Course in Differential Equations and Mathematical Modeling, Higher Education Press, Beijing. [https://doi.org/10.1142/7573](https://doi.org/10.1142/7573)

[28] Lie, S. (1891) Vorlesungen uber Differentialgleichungen Mit Bekannten Infinitesimalen Transformationen. Bearbeitet und herausgegeben von Dr. G. Scheffers, B.G. Teubner, Leipzig.

[29] Cohen, A. (1911) An Introduction to the Lie Theory of One-Parameter Groups with Applications to the Solution of Differential Equations. D.C. Heath, New York.

[30] Dickson, L. (1924) Differential Equations from the Group Standpoint. *Annals of Mathematics*, **25**, 287. [https://doi.org/10.2307/1967773](https://doi.org/10.2307/1967773)

[31] Bluman, G. and Anco, S. (2002) Symmetry and Integration Methods for Differential Equations. Springer-Verlag, New York.

[32] Ibragimov, N. (1990) Elementary Lie Group Analysis and Ordinary Differential Equations. John Willey & Sons, Chichester.
Appendix: Outline of Methods from Lie Group Analysis

Basic concepts from Lie group analysis of differential equations that are used in the present paper are assembled here. For further information regarding Lie groups and their applications to the theory of differential equations, the reader should consult the various classical and modern texts in the field, such as [23] [24] [25] [26] [28] [29] [30] [31]. A concise introduction to the calculus of Lie symmetry groups can be found in [27]. Readers interested in applications of Lie groups in fluid dynamics can find a wealth of information in [23] and [32].

**Definition of one-parameter groups.** Let

\[ z_i = f_i^a(z, a), \quad i = 1, \ldots, N, \quad (A1) \]

be a one-parameter family of invertible transformations of points

\[ z = (z^1, \ldots, z^N) \in \mathbb{R}^N \]

into points

\[ z = (z^1, \ldots, z^N) \in \mathbb{R}^N. \]

Here \( a \) is a real parameter from a neighborhood of \( a = 0 \), and we impose the condition that Transformation (A1) is an identity if and only if \( a = 0 \), i.e.,

\[ f_i^a(z, 0) = z^i, \quad i = 1, \ldots, N. \quad (A2) \]

The set \( G \) of transformations (A1) satisfying Condition (A2) is called a (local) one-parameter group of transformations in \( \mathbb{R}^N \) if the successive action of two transformations is identical to the action of a third transformation from \( G \), i.e., if the function

\[ f^i(f(z, a), b) = f^i(z, c), \quad i = 1, \ldots, N, \quad (A3) \]

where

\[ c = \varphi(a, b) \quad (A4) \]

with a smooth function \( \varphi(a, b) \) defined for sufficiently small \( a \) and \( b \). The group parameter \( a \) in the transformation (A1) can be changed so that the function (A4) becomes \( c = a + b \). In other words, the group property (A3) can be written, upon choosing an appropriate parameter \( a \) (called a canonical parameter) in the form

\[ f^i(f(z, a), b) = f^i(z, a + b). \quad (A5) \]

**Group Generator.** Let \( G \) be a group of transformations (A1) satisfying the condition (A2) and the group property (A5). Expanding the functions \( f^i(z, a) \) into Taylor series near \( a = 0 \) and keeping only the linear terms in \( a \), one obtains the infinitesimal transformation of the group \( G \):

\[ z^i \approx z'^i + a \xi_i^i(z), \quad (A6) \]

where

\[ \xi_i^i(z) = \frac{\partial f^i(z, a)}{\partial a} \bigg|_{a=0}, \quad i = 1, \ldots, N. \quad (A7) \]

The first-order linear differential operator

\[ X = \xi_i^i(z) \frac{\partial}{\partial z^i} \quad (A8) \]
is known as the generator of the group $G$.

**Invariants.** A function $J(z)$ is said to be an invariant of the group $G$ if for each point $z = (z^1, \ldots, z^N) \in \mathbb{R}^N$ is is constant along the trajectory determined by the totality of transformed points $\pi: J(\pi) = J(z)$.

The function $J(z)$ is an invariant of the group $G$ with Generator (A8) if and only if

$$X(J) = \xi^\alpha (z) \frac{\partial J}{\partial \xi^\alpha} = 0. \quad (A9)$$

Hence any one-parameter group has exactly $N-1$ functionally independent invariants (basis of invariants). One can take them to be the left-hand sides of $N-1$ first integrals $J_i(z) = C_i, \ldots, J_{N-1}(z) = C_{N-1}$ of the characteristic equations for linear partial differential Equation (A9). Then any other invariant is a function of $J_1(z), \ldots, J_{N-1}(z)$.

**Invariant equations.** We say that a system of equations

$$F_k(z) = 0, \; k = 1, \ldots, s \quad (A10)$$

is invariant with respect to the group $G$ (or admits the group $G$) if the transformations (A1) of the group $G$ map any solution of Equations (A10) into a solution of the same equations, i.e.,

$$F_k(\pi) = 0, \; k = 1, \ldots, s \quad (A11)$$

whenever $z$ solves Equations (A10). The group $G$ with the generator (A8) is admitted by Equations (A10) if and only if

$$X(F_k) \bigg|_{(A10)} = 0, \; k = 1, \ldots, s \quad (A12)$$

where the symbol $|_{(A10)}$ means evaluated on the solutions of Equations (A10).

If $z$ is a collection of independent variables $x = (x^1, \ldots, x^n)$, dependent variables $u = (u^1, \ldots, u^n)$ and partial derivatives $u_{(i)} = \{u^a_i\}, u_{(z)} = \{u^a_z\}, \ldots$, of $u$ with respect to $x$ up to certain order, where

$$u^a_i = \frac{\partial u^a}{\partial x^i}, \; u^a_z = \frac{\partial^2 u^a}{\partial x^i \partial x^j}, \ldots$$

then (A10) is a system of partial differential equations

$$F_k(x, u, u_{(i)}), \ldots = 0, \; k = 1, \ldots, s. \quad (A13)$$

Furthermore, if the transformations (A1) are obtained by the transformations of the independent and dependent variables

$$\bar{x} = f(x, u, a), \; \bar{u} = g(x, u, a) \quad (A14)$$

and the extension of (A14) to all derivatives $u_{(i)}$, etc. involved in the differential Equations (A13), then Equations (A11) define a group $G$ of transformations (A14) admitted by the differential Equations (A13). In other words, an admitted group does not change the form of the system of differential Equations (A13).

The generator of the admitted group $G$ is termed an infinitesimal symmetry (or simply symmetry) of the differential Equations (A13). Equations (A12) serve for obtaining the infinitesimal symmetries and are known as the determining
These equations are linear and homogeneous and therefore the set $L$ of its solutions is a vector space. Integration of determining equations often provides several linearly independent infinitesimal symmetries. Moreover, the determining equations have a specific property that guarantees that the set $L$ is closed with respect to the commutator $[X_1, X_2] = X_1 X_2 - X_2 X_1$. Due to this property, $L$ is called a Lie algebra. If the dimension of the vector space $L$ is equal to $r$, the space is denoted by $L_r$ and is called an $r$-dimensional Lie algebra. An $r$-dimensional Lie algebra $L_r$ generates a group depending on $r$ parameters which is called an $r$-parameter group.

**Invariant solutions.** Let the differential Equations (A13) admit a multi-parameter group $G$, and let $H$ be a subgroup of $G$. A solution

$$u^\alpha = h^\alpha(x), \quad \alpha = 1, \cdots, m$$

(A15)

defined by Equations (A13) is called an $H$-invariant solution (termed for brevity an invariant solution) if Equations (A15) are invariant with respect to the subgroup $H$. If $H$ is a one-parameter group and has the generator $X$, then the $H$-invariant solutions are constructed by calculating a basis of invariants $J_1, J_2, \cdots$. 

---

**Submit or recommend next manuscript to SCIRP and we will provide best service for you:**

Accepting pre-submission inquiries through Email, Facebook, LinkedIn, Twitter, etc.
A wide selection of journals (inclusive of 9 subjects, more than 200 journals)
Providing 24-hour high-quality service
User-friendly online submission system
Fair and swift peer-review system
Efficient typesetting and proofreading procedure
Display of the result of downloads and visits, as well as the number of cited articles
Maximum dissemination of your research work

Submit your manuscript at: [http://papersubmission.scirp.org/](http://papersubmission.scirp.org/)
Or contact jamp@scirp.org
Supply Chain Finance Decision Analysis with a Partial Credit Guarantee Contract

Yueliang Su, Baoyu Zhong

School of Business, South China University of Technology, Guangzhou, China
Email: 490943753@qq.com

Abstract

The innovation of supply chain financial services can alleviate the plight of SMEs financing difficulties. In the aspect of supply chain finance model, there is a credit guarantee financing model, which is different from the simple external financing and internal financing mode of supply chain. Based on this, this paper studies the decision-making of supply chain finance under the partial credit guarantee of core enterprises. First of all, the paper constructs a simple supply chain financing model, consisting of a bank, a core enterprise and a retailer. And then, considering the credit guarantee financing model, calculate the expected profit function. Stackelberg game model is used to give the optimal decision of each subject in decentralized system and the optimal decision in centralized system. Finally, in order to make a more specific and detailed study on the profit and decision-making based on the credit guarantee financing model, the important parameters of the model are analyzed. Through the calculation, it is proved that under the credit guarantee of the core enterprise, the retailer has the optimal ordering strategy, and the core enterprise has the best wholesale price. The influences of the partial credit guarantee coefficient and the retailer’s loan coefficient on the supply chain finance decision-making are also studied.

Keywords

Supply Chain Finance, Decision-Making, Partial Credit Guarantee

1. Introduction

Considering the risk of SMEs, financial institutions have been cautious about the financing of SMEs. Supply chain finance, through the integration of information chain, capital chain, logistics chain, and so on, forms an internal circulation ecosystem, which is a breakthrough to solve the financing difficulties of SEMs. In the development of supply chain finance in China, the evaluation of the en-

DOI: 10.4236/jamp.2017.56112 June 29, 2017
terprise’s credit, collateral is basically carried out by the bank or the bank commissioned by third parties. Therefore, the main mode of supply chain finance in China is banking-oriented model. At the same time, as the most powerful leading enterprise in the supply chain, the core enterprise can help to improve the whole supply chain with the advantage of information and credit. The credit guarantee financing model, providing credit guarantee for the downstream retailers, gradually appeared.

About supply chain decision-making and coordination, most scholars study the wholesale price contract, risk sharing, revenue sharing, buy back contract, rebate contract and other contracts based on two echelon supply chain, retailer-manufacturer or supplier-manufacturer.

In the supply chain, only a small number of scholars consider the impact of financial constraints on the supply chain decisions. The research about the decision-making of three echelon supply chain, retailer-manufacturer-bank, is also less. In this paper, the credit guarantee contract is integrated into the supply chain finance model, to study the decision-making of supply chain finance.

2. Literature Review

In recent years, many scholars have used the basic newsboy model and the Stackelberg game model to do quantitative research on the supply chain with capital constraints. Dada & Hu [1] consider the newsboy model of SCF with capital constraints, and point out that the problem of supply chain financing strategy is mainly focused on the optimal inventory management of retailers when they have capital constraints. Srinivasa [2] further considers the problem of single stage financing decision in two echelon supply chain with capital constraints, which shows that the performance of joint financing is greater than separate financing’s. Kouvelis and Zhao [3] consider the financing decision problem of supply chain with a capital constraint in the game framework that manufacturer led. Yan et al. (2014) [4] establish a multi-agent game model in which the bank acts as a leader and the manufacturer acts as the sub-leader, and study the joint financing decision problem of the supply chain financial system. Nina et al. [5] design a partial credit guarantee contract for SCF, incorporating the bank credit financing and manufacturer’s trade credit guarantee, to analyze its equilibrium financing strategies. Yan & Sun [6] study the optimal strategy of the capital-constrained retailer in SCF through warehouse receipt pledge under the uncertain demand environment. Considering the credit limit and the ruin probability of the retailer, the optimal financing interest rate, the retailer’s optimal order decision and the optimal wholesale price of the manufacturer are analyzed. Lu et al. [7] establish a multi-stage supply chain decision model including a supplier, a downstream manufacturer and a financial institution according to accounts receivable financing model of supply chain, and study the decision-making problem of the firms with and without financing. Yi & Zhou [8] consider a two echelon supply chain consisting of a single supplier and a single retailer. The paper analyses the bank’s loan to value ratios when the retailer
pledging his order contract in a newsboy supply chain. Li & Lou [9] establish a one-to-one supply chain dynamic game model with a capital-constrained supplier, and analyze the effectiveness to the improvement of the efficiency of supply chain by retailer’s advance payment. Lin Chen et al. [10] study the pricing and effort decisions of a supply chain with single manufacturer and single retailer. The results imply that the uncertainty degree of sales effort elasticity has an outstanding influence on the pricing and effort decisions, whereas the uncertainty degree of price elasticity has a modest impact on these decisions. Z. Liu et al. [11] find that the retailer’s optimal order quantity is determined by the inverse distribution of the external demand and the confidence level considering a contract-design problem for two competing heterogeneous suppliers working with a common retailer. H. Chen et al. [12] consider the optimal selling problem of a supplier who sells the same product to two competing retailers under two types of uncertainty—the selling costs of retailers and external demand and the results demonstrate that higher risk levels correlate with lower belief-degree costs of the two retailers and higher belief-degree sizes of the market.

On the basis of the scholars’ research, the credit guarantee contract is integrated into the supply chain finance model. This paper innovatively set up the credit guarantee coefficient and retailer’s loan coefficient and studied the influence of coefficients on decision-making of supply chain finance.

3. Model Description

Supply chain financing is generally short, this paper establishes a single order cycle of supply chain finance model including a capital-constrained retailer, a core manufacturers and a bank. Then we formulate a Stackelberg game model in which the bank acts as a leader.

3.1. Variable Definitions and Parameters

In order to describe the analysis model easily, the relevant variables and parameters are defined as follows.

(1) \( p \): Unit retail price
(2) \( c \): Unit manufacturing cost
(3) \( w \): The manufacturer’s wholesale price
(4) \( a \): The retailer’s loan coefficient, \( 0 \leq a \leq w \)
(5) \( B \): The retailer’s loan amount, \( B = aq \)
(6) \( R_r \): The bank’s endogenous interest rate
(7) \( R_f \): The risk-free interest rate
(8) \( \lambda \): The credit guarantee coefficient, \( 0 < \lambda < 1 \)
(9) \( q \): The retailer’s order quantity
(10) \( x \): Random demand
(11) \( F(x) \): Distribution function of random demand
(12) \( f(x) \): density function of \( F(x) \)

Let \( q, w, R_r \) respectively be the decision variables and \( q^*, w^*, R_r^* \) respectively be the retailer’s optimal order quantity, manufacturer’s optimal whole-
3.2. Model Assumptions

In order to describe the analysis model easily, make the following assumptions.
1) The supply chain finance system considers only one order cycle, consisting of a retailer, a manufacturer and a bank.
2) It is assumed that the distribution function $F(x)$ is a continuous function which is consistent with the increasing rate of failure distribution (IGFR).
3) It is assumed that retailer only sells a single product and the manufacturer can fully meet the demand of the retailer and the bank can meet the retailer’s loan.
4) Assuming that the retailer’s financing volume is proportional to the order quantity and let the loan coefficient be parameter $a$, where $B = aq$.
5) Assume that the product has no residual value and $p \geq a(1 + R_i)$.
6) In the supply chain finance system, the participants are risk neutral, and the goal is to maximize the expected profit.

3.3. The Model Framework

A single cycle supply chain financial system consisting of a retailer, a manufacturer and a bank is established. The capital-constrained retailer needs to apply for a loan. The manufacturer provides a certain credit guarantee for the retailer’s loan, who will pay a certain proportion of the remaining loan when the retailer can’t repay. Bank is the leader of supply chain finance system.

Supply chain finance process are assumed by Figure 1.

1) First of all, as the leader of the supply chain financial system, the bank gives an appropriate loan interest rate $r_R$.
2) Secondly, the sub-leader decides a wholesale price $w$ when the bank has given the interest rate.
3) Then, the retailer makes a decision on the order quantity $q$ after the bank gave interest rate $R_i$ and the manufacturer decided wholesale price $w$. At the same time, the retailer apply to the bank for a loan $B = aq$ because of capital constraints.
4) At the end of the period, the retailer repay the loan according to sales. If the retailer has the ability to repay the loan, the retailer bears all the principal

![Figure 1. Supply chain finance process.](image-url)
and interest. If the retailer’s total sales are insufficient to repay all of the loan, the bank and the manufacturer take up the remaining loan together. The manufacturer bears $\lambda$ of the retailer’s remaining loan.

We formulate a multilevel Stackelberg game model to characterize the interactions among the three SCF participants, which can be expressed as follows.

\[
\begin{align*}
(L) \max_R \left[ \pi_B (R, q, w, x) \right] \\
(SL) \max_w \left[ \pi_M (w, q, R, x) \right] \\
(F) \max_q \left[ \pi_R (q; R, w, x) \right]
\end{align*}
\]

4. The Supply Chain Finance System’s Profits

**Proposition 1**

When the random demand $x$ is greater than the minimum realized demand $x_i$, the retailer can pay the principal and interest by itself at the end of the period, where $x_i = aq(1 + R)/p$. And $x_i \leq q$ is also obtained.

**Proof:** At the end of the period, the retailer need to repay the principal and interest i.e. $aq(1 + R)/p$.

If $x < q$, the total sales of the retailer are $px$. Then, $px \geq px_i = aq(1 + R)/p$ holds for $x \geq x_i = aq(1 + R)/p$.

If $x \geq q$, the total sales of the retailer are $pq$. Then, $pq \geq px_i = aq(1 + R)$ is obtained for $p \geq aq(1 + R)$.

Hence, $x_i = aq(1 + R)/p \leq q$ holds for $px_i = aq(1 + R) = pq$.

4.1. The Supply Chain Finance System’s Profits in Decentralized System

At the beginning of the period, the manufacturer gives the wholesale price $w$, and the retailer orders quantity $q$. Due to financial constraints, the retailer applies for a loan $B = aq$ from bank and its own funds are $wq = aq$. The bank gives a loan interest rate $R$ and signs a credit guarantee contract with the manufacturer who bears $\lambda$ of the retailer’s remaining loan.

At the end of the period, the retailer’s sales revenue is $\min \{px, pq\}$, who need to repay the principal and interest $aq(1 + R)$. If the retailer can’t repay the principal and interest, i.e. $x < x_i$, the remaining loan is $aq(1 + R) - px$ that the retailer can’t repay. The bank and the manufacturer take up the remaining loan together. The manufacturer bears $\lambda$ of the retailer’s remaining loan and the bank bears $1 - \lambda$ of the retailer’s remaining loan.

1) The retailer’s profit function can be expressed as in Equation (1)

\[
R = \begin{cases} 
-(wq - aq), & x \leq x_i \\
px - aq(1 + R) - (wq - aq), & x_i < x < q \\
 pq - aq(1 + R) - (wq - aq), & q \leq x
\end{cases}
\]  

(1)

The retailer’s expected profit function can be expressed as in Equation (2)
\[
\pi_R = E(R) \\
= \int_{0}^{\infty} (wq - aq) f(x) dx + \int_{0}^{\infty} (px - aq(1 + R)) - (wq - aq)) f(x) dx + \int_{0}^{\infty} (px - aq(1 + R)) - (wq - aq)) f(x) dx
\] (2)

2) The manufacturer’s profit function can be expressed as in Equation (3)
\[
M = \begin{cases} 
(w - c)q - \lambda[aq(1 + R) - px], & x < x_i \\
(w - c)q, & x \geq x_i 
\end{cases}
\]

(3)

The manufacturer’s expected profit function can be expressed as in Equation (4)
\[
\pi_M = E(M) \\
= \int_{0}^{\infty} ((w - c)q - \lambda[aq(1 + R) - px]) f(x) dx + \int_{0}^{\infty} (w - c)qf(x) dx
\] (4)

3) The bank’s profit function can be expressed as in Equation (5)
\[
B = \begin{cases} 
aq(R_f - R_f) - (1 - \lambda)[aq(1 + R) - px], & x < x_i \\
aq(R_f - R_f), & x \geq x_i 
\end{cases}
\]

(5)

The bank’s expected profit function can be expressed as in Equation (6)
\[
\pi_B = E(B) = \int_{0}^{\infty} aq(R_f - R_f) - (1 - \lambda)[aq(1 + R) - px]) f(x) dx + \int_{0}^{\infty} aq(R_f - R_f) f(x) dx
\] (6)

4.2. The Supply Chain Finance System’s Profit in Centralized System

The supply chain finance system’s profit function can be expressed as in Equation (7)
\[
S = \begin{cases} 
px - cq - aqR_f, & x < q \\
pq - cq - aqR_f, & x \geq q 
\end{cases}
\]

(7)

The supply chain finance system’s expected profit function can be expressed as in Equation (8)
\[
\pi_S = E(S) = \int_{0}^{q} (px - cq - aqR_f) f(x) dx + \int_{q}^{\infty} (pq - cq - aqR_f) f(x) dx
\] (8)

5. The Supply Chain Finance System’s Optimal Decisions

We solve the model via backward induction to determine the optimal decisions of the supply chain finance system.

5.1. The Supply Chain Finance System’s Decisions in Decentralized System

Proposition 2

Given the unit retail price \( p \), the manufacturer’s wholesale price \( w \), unit manufacturing cost \( c \), retailer’s loan coefficient \( a \), the bank’s endogenous interest rate \( R_f \) and the risk-free interest rate \( R_f \), for IGFR distributions of demand, the
capital-constrained retailer’s unique and optimal order quantity that satisfies

\[ q^* = F^{-1}\left( \frac{p-(w-a)-a(1+R_c)F(x^*_i)}{p} \right), \text{ where } x^*_i = aq^*(1+R_c)/p. \]

**Proof:**

From Equation (2), taking the first-order and second-order derivative of \( \pi_r \) with respect to \( q \), it follows that

\[
\frac{d\pi_r}{dq} = -(w-a) - a(1+R_c)\frac{F(x_i)}{p} + pF(q)
\]

\[
\frac{d^2\pi_r}{dq^2} = \frac{a^2(1+R_c)^2}{p}f(x_i) - pf(q)
\]

If the distribution of demand is IGFR, \( f(x_i) \leq f(q) \) holds for \( x_i \leq q \). Besides, \( \frac{a^2(1+R_c)^2}{p} \leq a(1+R_c) \leq p \) holds for \( a(1+R_c) \leq p \). Hence, we have \( \frac{d^2\pi_r}{dq^2} \leq 0 \). From the first-order condition of \( \frac{d\pi_r}{dq} = 0 \), we have

\[ q^* = F^{-1}\left( \frac{p-(w-a)-a(1+R_c)F(x_i)}{p} \right) \]

**Lemma 1**

\[ \frac{dq^*}{dw} < 0, \frac{dx^*_i}{dw} < 0, \frac{dq^*}{dR_c} < 0 \]

**Proof:**

From Proposition 2, we have

\[ F(q^*) = \frac{p-(w-a)-a(1+R_c)F(x_i(q^*))}{p} \]

Taking the first-order derivative of \( F(q^*) \) with respect to \( w \), we have

\[ \frac{dq^*}{dw} f(q^*) = \frac{-p + a^2(1+R_c)^2 f(x_i(q^*))}{p^2} \frac{dq^*}{dw}. \]

Hence,

\[ \frac{dq^*}{dw} = \frac{p}{a^2(1+R_c)^2 f(x_i(q^*)) - p^2 f(q^*)}. \]

\( \frac{dq^*}{dw} < 0 \) holds for \( a^2(1+R_c)^2 f(x_i(q^*)) - p^2 f(q^*) < 0 \) and \( p > 0 \). It is obvious that

\[ \frac{dx^*_i}{dw} = \frac{a(1+R_c)}{p} \frac{dq^*}{dw} < 0. \]
\[ \frac{dq^*}{dR_r} < 0 \] can be obtained too.

**Proposition 3**

Given the unit retail price \( p \), unit manufacturing cost \( c \), retailer’s loan coefficient \( a \), the bank’s endogenous interest rate \( R_r \), the risk-free interest rate \( R_f \), the credit guarantee coefficient \( \lambda \) and the optimal order quantity

\[ q^* = F^{-1}\left(\frac{p - (w - a - a(1 + R_r)F(x_i))}{p}\right), \]

for IGFR distributions of demand, the manufacturer’s unique and optimal wholesale price that satisfies

\[ w^* = \frac{\lambda a(1 + R_r)F(x_i^*)\frac{dq^*}{dw^*} - q^*}{\frac{dq^*}{dw^*}} + c. \]

**Proof:**

The wholesale price has a definite closed interval that \( w \in [c, p] \), and the manufacturer’s profit function is continuous on the closed interval, there must be a maximum expected profit.

Considering \( q^* \) and \( x_i^* = \frac{aq^*(1 + R_r)}{p} \) from Proposition 2 and Equation (4),

\[ \pi_M(q^*) = E(M) = \int_0^\infty \left( (w-c)q^* - \lambda [aq^*(1 + R_r) - px_i] \right) f(x) dx + \int_{q^*}^\infty (w-c)q^* f(x) dx \]

Taking the first-order derivative of \( \pi_M(q^*) \) with respect to \( w \),

\[ \frac{d\pi_M}{dw} = q^* + (w-c)\frac{dq^*}{dw} - \lambda a(1 + R_r)F(x_i^*)\frac{dq^*}{dw} \]

\[ = q^* + (w-c - \lambda a(1 + R_r)F(x_i^*))\frac{dq^*}{dw}, \]

\( q^* \) and \( (w-c - \lambda a(1 + R_r)F(x_i^*))\frac{dq^*}{dw} \) monotonically decrease for \( \frac{dq^*}{dw} < 0 \) and hence, \( \frac{d^2\pi_M}{dw^2} < 0 \). From the first-order condition of \( \frac{d\pi_M}{dw} = 0 \), we have

\[ w^* = \frac{\lambda a(1 + R_r)F(x_i^*)\frac{dq^*}{dw^*} - q^*}{\frac{dq^*}{dw^*}} + c \]

**Proposition 4**

Given the unit retail price \( p \), unit manufacturing cost \( c \), retailer’s loan coefficient \( a \), the risk-free interest rate \( R_f \), the credit guarantee coefficient \( \lambda \), the optimal order quantity \( q^* = F^{-1}\left(\frac{p - (w - a - a(1 + R_r)F(x_i))}{p}\right) \) and the optim-
al wholesale price \( w^* = \frac{\lambda a (1 + R_f) F(x_i^*) \frac{dq^*}{dw^*} - q^*}{\frac{dq^*}{dw^*}} + c \), for IGFR distributions of demand, the expected profit of the bank increases with the loan interest rate when \( R_f < (1 - \lambda) a (1 + R_f) F(x_i^*) + R_f \).

**Proof:**

\[
\frac{d\pi_s}{dR_f} = \frac{d\pi_s}{dq^*} \frac{dq^*}{dR_f}
\]

From Lemma 1, \( \frac{dq^*}{dR_f} < 0 \). Hence, \( \frac{d\pi_s}{dR_f} \) can be transformed into \( \frac{d\pi_s}{dq^*} \). Taking the first-order derivative of \( d\pi_s \) with respect to \( q^* \)

\[
\frac{d\pi_s}{dq^*} = a(R_f - R_f) - (1 - \lambda) a (1 + R_f) F(x_i^*) \]

\( \frac{d\pi_s}{dq^*} < 0 \) for \( R_f < (1 - \lambda) a (1 + R_f) F(x_i^*) + R_f \). And then \( \frac{d\pi_s}{dR_f} > 0 \) for \( \frac{d\pi_s}{dq^*} < 0 \) and \( \frac{dq^*}{dR_f} < 0 \).

**5.2. The Supply Chain Finance System’s Decision in Centralized System**

**Proposition 5**

In centralized system, given the unit retail price \( p \), unit manufacturing cost \( c \), retailer’s loan coefficient \( a \) and the risk-free interest rate \( R_f \), for IGFR distributions of demand, the capital-constrained retailer’s unique and optimal order quantity that satisfies \( q^* = F^{-1}\left(\frac{p - c - aR_f}{p}\right) \).

**Proof:**

From Equation (8), taking the first-order and second-order derivative of \( \pi_s \) with respect to \( q \), it follows that

\[
\frac{d\pi_s}{dq} = pF(q) - c - aR_f
\]

\[
\frac{d^2\pi_s}{dq^2} = -pf(q)
\]

It is obvious that \( \frac{d^2\pi_s}{dq^2} = -pf(q) \leq 0 \).

From the first-order condition of \( \frac{d\pi_s}{dq} = 0 \), we have \( q^*_s = F^{-1}\left(\frac{p - c - aR_f}{p}\right) \).

**6. Numerical Example**

Assuming that distribution function of random demand obeys uniform distribution with a mean of 200, the unit retail price \( p = 8 \), manufacturer’s wholesale
price \( c = 6 \), the risk-free interest rate \( R_f = 0.03 \).

Given the retailer’s loan coefficient \( a = 6 \) and credit guarantee coefficient \( \lambda = 0.9 \), Figure 2 describes the change of retailer’s optimal order quantity with the loan interest rate \( R_f \) in decentralized and centralized system. Figure 3 describes the change of SCF participants and system’s optimal profit with the loan interest rate \( R_f \) in decentralized and centralized system.

(1) From Proposition 5, the retailer’s optimal order quantity is only related to \( p, c, a, R_f \) in centralized system. Therefore, the retailer’s optimal order quantity and the optimal profit of the supply chain finance system do not change with the loan interest rate \( R_f \).

(2) After the game analysis, the retailer’s optimal order quantity decreases with the increase of the loan interest rate \( R_f \) in the decentralized system. On the other hand, the bank’s optimal profit increases with the increase of loan interest rate \( R_f \), and the growth rate slows down with the increase of loan interest rate \( R_f \). The retailer, manufacturer and SCF system’s optimal profit decrease with the increase of the loan interest rate. In order to encourage retailer and manufacturer, the bank should choose the appropriate interest rate \( R_f \).

![Figure 2. Retailer’s optimal order quantity under different loan interest rate.](image)

![Figure 3. SCF participants and system’s optimal profit under different loan interest rate.](image)
(3) The bank’s optimal profit is higher than the retailer’s one after the bank’s interest rate reaches about 0.07, and the bank’s optimal profit is higher than the manufacturer’s one after the bank’s interest rate reaches about 0.14.

(4) According to Figure 3, the optimal profit of the supply chain financial system in centralized system is always higher than the sum of the three SCF participants’ optimal profit in the decentralized system. The higher $R_c$ is, the larger optimal profit gap between the decentralized and centralized system is.

Given the bank’s endogenous interest rate $R_b$ and credit guarantee coefficient $\lambda = 0.9$, Figure 4 describes the change of retailer’s optimal order quantity with the retailer’s loan coefficient $a$ in decentralized and centralized system. Figure 5 describes the change of SCF participants and system’s optimal profit with the retailer’s loan coefficient $a$ in decentralized and centralized system.

![Figure 4](image4.png)

**Figure 4.** Retailer’s optimal order quantity under different retailer’s loan coefficient.

![Figure 5](image5.png)

**Figure 5.** SCF participants and system’s optimal profit under different retailer’s loan coefficient.
1) In centralized system, the retailer’s optimal order quantity and the optimal profit of the supply chain finance system decrease with the increase of the loan coefficient $a$.

2) In decentralized system, the retailer’s optimal order quantity decreases with the increase of $a$, and then increases at a faster rate. The bank’s optimal profit increases with the increase of $a$. And with the increase of $a$, the retailer’s optimal profit firstly declines at a faster rate, then decreases slowly, and finally decreases sharply. The manufacturer’s optimal profit decreases with the increase of $a$, and then increases at a faster rate. The variation tendency of the three SCF participants’ total optimal profit is approximately the same as that of the manufacturer.

3) The bank’s optimal profit and the retailer’s optimal profit are equal at about 6.5 of $a$.

4) From Figure 5, the optimal profit of the supply chain financial system in centralized system is always higher than the sum of the three SCF participants’ optimal profit in the decentralized system too. When the the loan amount is the same as the order cost ($aq = wq$), that is, the retailer’s own capital is zero, the optimal profit of the SCF system in centralized system is equal to the sum of the three SCF participants’ optimal profit in the decentralized system. Appropriately increasing retailer’s loan coefficient can narrow the optimal profit gap between the decentralized and centralized system.

Given the bank’s endogenous interest rate $R_r = 0.06$ and the retailer’s loan coefficient $a = 6$, Figure 6 describes the change of retailer’s optimal order quantity with the credit guarantee coefficient $\lambda$ in decentralized and centralized system. Figure 7 describes the change of SCF participants and system’s optimal profit with the credit guarantee coefficient $\lambda$ in decentralized and centralized system.

1) In centralized system, the retailer’s optimal order quantity and the optimal profit of the supply chain finance system do not change with the credit guarantee coefficient $\lambda$.

![Figure 6. Retailer’s optimal order quantity under different credit guarantee coefficient.](image_url)
2) In decentralized system, the optimal order quantity of the retailer decreases with the increase of $\lambda$. The bank’s optimal profit increases with the increase of $\lambda$ at a faster rate. The optimal profit of retailer and manufacturer decreases with the increase of $\lambda$, whose variation tendency is approximately the same. The sum of the three SCF participants’ optimal profit in the decentralized system also decreases with the increase of $\lambda$.

3) When $\lambda$ is below about 0.6, the bank’s best profit is negative. The bank’s optimal profit is positive only when $\lambda$ is large enough.

4) According to Figure 7, the optimal profit of the supply chain financial system in centralized system is always higher than the sum of the three SCF participants’ optimal profit in the decentralized system. The higher $\lambda$ is, the larger optimal profit gap between the decentralized and centralized system is.

7. Conclusions

This paper studies the decision-making problem of three echelon supply chain, retailer-manufacturer-bank, in centralized and decentralized systems. Through the calculation, it is proved that under the credit guarantee of the core enterprise, the retailer has the optimal ordering strategy, and the core enterprise has the optimal wholesale price. Besides, the optimal profit of the SCF system in centralized system is always higher than the sum of the three SCF participants’ optimal profit in the decentralized system.

The retailer’s loan coefficient and the credit guarantee coefficient can narrow the optimal profit gap between the decentralized and centralized system to a certain extent. It is embodied in two aspects. On the one hand, when the retailer’s loan coefficient is consistent with the wholesale price, that the retailer’s loan amount is 0, the optimal profit of the SCF system in centralized system is the same as the sum of the three SCF participants’ optimal profit in the decentralized system. Considering that the principal and interest of the retailer are the...
Y. L. Su, B. Y. Zhong

retailer’s total sales when the supply chain financial system is balanced, i.e. the retailer’s profit is zero. Hence, the retailer’s loan coefficient can narrow the optimal profit gap between the decentralized and centralized system to a certain extent. On the other hand, the smaller $\lambda$ is, the narrower optimal profit gap between the decentralized and centralized system is.

Considering that the bank’s expect profit is negative when the credit guarantee coefficient is too small, the credit guarantee coefficient can narrow the optimal profit gap between the decentralized and centralized system to a certain extent.

There are also shortcomings in this paper. In this paper, the decisions of parameter $a$ and $\lambda$ are not taken into account in the game model, but the influences on the game result are analyzed. On the other hand, this paper considers a simpler supply chain model. The actual situation is that supply chain finance financing is facing a more complex supply chain system, and involves dynamic evolution.

Acknowledgements

This research was supported by “the Fundamental Research Funds for the Central Universities” (2015ZKYJZX02).

References

[1] Dada, M. and Hu, Q. (2008) Financing Newsvendor Inventory. Operations Research Letters, 36, 569-573. https://doi.org/10.1016/j.orl.2008.06.004

[2] Srinivasa Raghavan, N.R. and Mishra, V.K. (2011) Short-Term Financing in a Cash-Constrained Supply Chain. International Journal of Production Economics, 134, 407-412. https://doi.org/10.1016/j.ijpe.2009.11.014

[3] Kouvelis, P. and Zhao, W. (2012) Financing the Newsvendor: Suppliers, Bank, and the Structure of Optimal Trade Credit Contracts. Operations Research, 60, 566-580. https://doi.org/10.1287/opre.1120.1040

[4] Yan, N., Dai, H. and Sun, B. (2014) Optimal Bi-Level Stackelberg Strategies for Supply Chain Financing with Both Capital-Constrained Buyers and Sellers. Applied Stochastic Models in Business and Industry, 30, 783-796. https://doi.org/10.1002/asmb.2021

[5] Yan, N.N., Sun, B.W., Zhang, H. and Liu, C.Q. (2016) A Partial Credit Guarantee Contract in a Capital-Constrained Supply Chain: Financing Equilibrium and Coordinating Strategy. International Journal of Production Economics, 173, 122-133. https://doi.org/10.1016/j.ijpe.2015.12.005

[6] Yan, N.N. and Sun, B.W. (2011) Optimal Strategies for Supply Chain Financing System Based on Warehouse Receipts Financing with Credit Line. System Engineering-Theory & Practice, No. 9, 1674-1679.

[7] Lu, Q.H. and Zeng, L.F. (2012) Research on Decision-Making and Value of Supply Chain Financing with Accounts Receivables. Journal of Management Sciences in China, 5, 10-18.

[8] Yi, Z. (2012) Study on Loan-to-Value Ratios of Bank in the Supply Chain Finance. Chinese Journal of Management Sciences, No. 1, 102-108.

[9] Li, C. and Luo, J.W. (2016) Revenue-Sharing Coordination Mechanism for a Finan-
cial Constrained Supply Chain Based on Advance Payment. *Chinese Journal of Management*, No. 5, 763-771.

[10] Chen, L., Peng, J., Liu, Z. and Zhao, R. (2016) Pricing and Effort Decisions for a Supply Chain with Uncertain Information. *International Journal of Production Research*, **55**, 264-284. https://doi.org/10.1080/00207543.2016.1204475

[11] Liu, Z., Zhao, R., Liu, X. and Chen, L. (2016) Contract Designing for a Supply Chain with Uncertain Information Based on Confidence Level. *Applied Soft Computing*, **56**, 617-631. https://doi.org/10.1016/j.asoc.2016.05.054

[12] Chen, H., Wang, X., Liu, Z. and Zhao, R. (2017) Impact of Risk Levels on Optimal Selling to Heterogeneous Retailers under Dual Uncertainties. *Journal of Ambient Intelligence and Humanized Computing*, 1-19. https://doi.org/10.1007/s12652-017-0481-9

---

Submit or recommend next manuscript to SCIRP and we will provide best service for you:

Accepting pre-submission inquiries through Email, Facebook, LinkedIn, Twitter, etc.
A wide selection of journals (inclusive of 9 subjects, more than 200 journals)
Providing 24-hour high-quality service
User-friendly online submission system
Fair and swift peer-review system
Efficient typesetting and proofreading procedure
Display of the result of downloads and visits, as well as the number of cited articles
Maximum dissemination of your research work

Submit your manuscript at: [http://papersubmission.scirp.org/](http://papersubmission.scirp.org/)
Or contact jamp@scirp.org
N-Rotating Loop-Soliton Solution of the Coupled Integrable Dispersionless Equation

Souleymanou Abbagari\(^1\), Saliou Youssoufa\(^2\), Hermann T. Tchokouansi\(^2\), Victor K. Kuetche\(^2\), Thomas B. Bouetou\(^2\), Timoleon C. Kofane\(^2\)

\(^1\)Department of Basic Science, Law and Humanities, Institute of Mines and Petroleum Industries, University of Maroua, Maroua, Cameroon
\(^2\)Laboratory of Mechanics, Materials and Structures, Department of Physics, University of Yaoundé I, Yaoundé, Cameroon
\(^3\)National Advanced School of Engineering, University of Yaoundé I, Yaoundé, Cameroon
\(^4\)Centre d’Excellence en Technologies de l’Information et de la Communication (CETIC), University of Yaoundé I, Yaoundé, Cameroon

Email: abbagaris@yahoo.fr

Abstract

In this paper, we investigate the Rotating N Loop-Soliton solution of the coupled integrable dispersionless equation (CIDE) that describes a current-fed string within an external magnetic field in 2D-space. Through a set of independent variable transformation, we derive the bilinear form of the CIDE equation. Based on the Hirota’s method, Perturbation technique and Symbolic computation, we present the analytic N-rotating loop soliton solution and proceed to some illustrations by presenting the cases of three- and four-soliton solutions.

Keywords

Coupled Integrable Dispersionless Equation, Bilinear Method, Soliton Solutions, Perturbation Technique, Symbolic Computation

1. Introduction

During the past several years, the study of coupled nonlinear evolution Equations has played an important role in explaining many interesting phenomena, like electromagnetic wave propagation in impurity media, water waves, pulse in biological chains and so on [1] [2] [3]. At the same time, the coupled integrable dispersionless system (CIDE) has attracted much interest in view of its wide range of application in various fields of mathematics, physics, applied mathematics, theory of quantum and theory of conformal maps on the complex plane [4] [5] [6] [7]. The CIDE, has first been presented by Konno and Oono in Ref.
based on a Lie-group $\mathcal{G} = SU(2)$, and its generalization based on the Lie-group $\mathcal{G} = SL(2,\mathbb{R})$ are examples of such system [7] [9] [10], which have attracted great deal of interest because of its nice integrability structure and soliton solution. Based on this standpoint, the solitons show loop shapes in the three-dimensional Euclidean space. The angular momentum conservation law can be derived from the Equations of motion of the string such that we can expect rotating loop solitons.

So far, several successful methods have been developed to obtain explicit solutions for soliton Equations, such as the Inverse Scattering Transformation (IST) [1] [11] [12], Bäcklund and Darboux Transformations [13] [14], the Hirota’s method [15] [16], the Wronskian and Cassoratian techniques [17] [18], the Algebra geometric method [19] and so on. Among these methods, the Hirota’s bilinear method has been proven to be an efficient and direct approach to construct soliton solutions to nonlinear evolution Equations via the bilinear forms from the dependent variables transformation.

In Ref [8], Konno and Oono have presented the well known CIDE

$$q_{xt} + \frac{1}{2}(rs)_t = 0,$$

$$r_{xt} - q_r r = 0,$$

$$s_{xt} - q_s s = 0,$$

where Equation (1) describes the current-fed within an external magnetic field [20]. In Equation (1), $q$, $r$, and $s$ are all functions of $x$ and $t$, the subscripts denote partial derivatives with respect to the space-like and time-like variables respectively.

The aim of this work is to verify if the congestion, due to the displacement of a great number of soliton will modify the conservation properties observed for the case of two solitons. Indeed, we provide the explicit expression of the $N$-Rotating loop soliton solution to the CIDE for the general positive integer $N \geq 2$ and to illustrate our general result, we discuss particular cases of $N$. Thus the following paper is organized as follows. In section 2, we summarize the transformation of the CIDE Equation (1) into an Equation in bilinear form. In section 3, we give the full expression of the $N$-Rotating loop soliton solution and we illustrate our results by considering in detail the cases of $N=1,2,3,4$ and we end this work with a brief summary.

2. Hirota’s Bilinearization of the CIDE

Let us consider the following setting [20] [21] [22]

$$r = X + iY, \quad s = X - iY,$$

$$q = Z, \quad \sigma = x + t, \quad \tau = x - t,$$

which inserted into Equation (1) gives

$$r_{\tau\tau} - r_{\sigma\sigma} = (r_r + r_s) \times (J \times r),$$

where $r = (X, Y, Z)$ stands to be the vector position of the string, $J = (1, 0, 1)$
is the constant electric current \[23\]. In Equation (3) the factor \( r_s + r_o \) can be interpreted as the Lorentz force acting on effective internal current, \( r_e \) can be considered as an internal electric current and \( r_s \) is a correction term induced by the motion of string to \( r_e \). Equation (3) can therefore represent a current-fed string interacting with the external magnetic field \( B = J \times r \) which satisfies the two Maxwell’s Equations \( \text{rot} B = 2J \) and \( \text{div} B = 0 \). Using the boundary condition \( r \rightarrow (0,0,\sigma) \) for \(|\sigma| \rightarrow \infty\), we bilinearize Equation (3) as

\[
D_s D_r Q \cdot F = Q \cdot F, \quad D_r^2 F \cdot F = \frac{1}{2} Q \cdot Q'^*, \quad (4)
\]

using the transformation

\[
r = \frac{Q}{F}, \quad q = x - 2\partial_t \ln F, \quad (5)
\]

where \( D \) denotes the Hirota’s derivative \[15\] \[16\]. Now, expanding \( Q \) and \( F \) as series

\[
F = 1 + \epsilon^2 F_2 + \epsilon^4 F_4 + \cdots + \epsilon^{2i} F_{2i} + \cdots, \\
Q = \epsilon Q_1 + \epsilon^3 Q_3 + \cdots + \epsilon^{2i+1} Q_{2i+1} + \cdots. \quad (6)
\]

Substituting the expansion into the above bilinear Equations, we find that there are only even order terms of \( \epsilon \) in the first Equation while odd order terms in the second one. Arranging the coefficients at each order of \( \epsilon \), we have

\[
\epsilon : D_s D_r (1 \cdot Q_1) = Q_1, \\
\epsilon^2 : D_r^2 (1 \cdot F_2 + F_2 \cdot 1) = \frac{1}{2} Q_1 Q_1', \\
\epsilon^3 : D_s D_r (1 \cdot Q_3 + Q_1 \cdot F_2) = Q_3 + Q_1 F_2, \\
\epsilon^4 : D_r^2 (1 \cdot F_4 + F_2 \cdot F_2 + F_4 \cdot 1) = \frac{1}{2} (Q_3 Q_1' + Q_1 Q_3'), \\
\epsilon^5 : D_s D_r (1 \cdot Q_5 + Q_3 \cdot F_2 + Q_1 \cdot F_4) = Q_5 F_2 + Q_3 F_4 + Q_1 F_2, \\
\epsilon^{2i} : D_r^2 \left( \sum_{m=0}^{i-1} F_{2m} \cdot F_{2i-2m} \right) = \frac{1}{2} \left( \sum_{k=0}^{i-1} Q_{2i+1} Q_{2i-2k-1} \right), \\
\epsilon^{2i+1} : D_s D_r \left( \sum_{m=0}^{i-1} Q_{2i+1} \cdot F_{2i-2m} \right) = \left( \sum_{k=0}^{i-1} Q_{2i+1} Q_{2i-2k} \right). \quad (7)
\]

It is then possible to obtain at the required order the required number of soliton solutions by determining the full expansion of \( F \) and \( Q \).

**3. Rotating one and Two-Loop Soliton Solution**

In this section, we derive the rotating solitons \( i.e., \) solutions that the \( Z \) component of the angular momentum is a conserved quantity. In order to construct one-rotating soliton solution, we take

\[
Q_1 = \exp(\eta_1), \quad (8)
\]

where \( \eta_1 = k_s x + \omega_1 t + \gamma_1 \). Substituting it into Equation (7), limiting our interest to the terms of \( \epsilon^i, \quad i \leq 2 \), we obtain
\( \omega k_i = 1, \quad F_z = A_i^\prime \exp(\eta_i + \eta_i^\prime), \quad i_i = j_i = 1, \quad (9) \)

the first part of Equation (9) standing for the dispersion relation and the coefficient \( A_i^\prime \) is giving by \( A_i^\prime = \frac{1}{4(\omega_i + \omega_i^\prime)^2} \). This show that the expansion can be truncated as the finite sum

\[
F = 1 + \frac{\epsilon \exp(\eta_i + \eta_i^\prime)}{4(\omega_i + \omega_i^\prime)^2}, \quad Q = \epsilon \exp(\eta_i).
\]  

(10)

Absorbing the parameter \( \epsilon \) into the phase constant \( \gamma_i \) gives the one-rotating soliton solution of the CIDE as it is depicted in Figure 1.

Next, we choose the solution of Equation (7) while limiting our interest to the terms of \( \epsilon^i, \ i \leq 4 \) to be

\[
Q_i = A_i^\prime \exp(\eta_i) + A_i^\prime \exp(\eta_i^\prime),
\]

(11)

where the phase \( \eta_i = k_i x + \omega_i t + \gamma_i \) and the dispersion relation \( k_i \omega_i = 1 \) with \( i = 1, 2 \). From Equation (7) we have

\[
F_z = A_i^\prime \exp(\eta_i + \eta_i^\prime) + A_i^\prime \exp(\eta_i + \eta_i^\prime) \\
+ A_i^\prime \exp(\eta_i + \eta_i^\prime) + A_i^\prime \exp(\eta_i + \eta_i^\prime),
\]

\[
Q_z = A_i^{12} \exp(\eta_i + \eta_i + \eta_i^\prime) + A_i^{12} \exp(\eta_i + \eta_i + \eta_i^\prime) \\
F_z = A_i^{12} \exp(\eta_i + \eta_i + \eta_i^\prime + \eta_i^\prime),
\]

(12)

where

Figure 1. From left to right panels rotating one-loop soliton solution to the CIDE Equation (1): For left we depict at times \( t = -30 \) (blue color), \( t = 0 \) (red color) and \( t = 30 \) (black color) corresponding to three moving states, with \( \nu = 0.66 \) and the computed angular velocities of such wave is \( \Omega = 0.40 \), respectively.
According to the above analysis, the two-rotating soliton solution is obtained when we substitute Equations (11)-(13) into Equation (5) as it is depicted in Figure 2.

Generally we can conjecture the N-rotating soliton solution as

\[
F = 1 + \sum_{m=1}^{[\frac{N}{2}]} \sum_{\alpha=\beta}^{\gamma_{\alpha} \gamma_{\beta}} A_{\gamma_{\alpha} \gamma_{\beta}}^{v_{\alpha} v_{\beta}} \exp \left( \eta_{\alpha} + \cdots + \eta_{\alpha}^{*} + \eta_{\beta} + \cdots + \eta_{\beta}^{*} \right),
\]

\[
Q = \sum_{m=0}^{[\frac{N-1}{2}]} \sum_{\alpha=\beta}^{\gamma_{\alpha} \gamma_{\beta}} A_{\gamma_{\alpha} \gamma_{\beta}}^{v_{\alpha} v_{\beta}} \exp \left( \eta_{\alpha} + \cdots + \eta_{\alpha}^{*} + \eta_{\beta} + \cdots + \eta_{\beta}^{*} \right),
\]

where the phase \( \eta_{p} = k_{p} x + \omega_{\gamma} t + \gamma_{p} \) and the dispersion relation \( k_{p} \omega_{p} = 1 \) with \( p = 1, \ldots, N \).

\[
A^{\delta} = 1, \quad A_{\gamma_{\alpha}}^{v_{\alpha}} = \frac{1}{4 \left( \omega_{\alpha}^{*} + \omega_{\gamma_{\alpha}}^{*} \right)^{2}}, \quad A_{\gamma_{\alpha} \gamma_{\beta}}^{v_{\alpha} v_{\beta}} = 4 A_{\gamma_{\alpha}}^{v_{\alpha}} A_{\gamma_{\beta}}^{v_{\beta}} \left( \omega_{\alpha} - \omega_{\gamma_{\alpha}} \right)^{2},
\]

\[
A_{\gamma_{\alpha} \gamma_{\beta}}^{v_{\alpha} v_{\beta}} = 4^{2} A_{\gamma_{\alpha} \gamma_{\beta}}^{v_{\alpha} v_{\beta}} \left( \omega_{\alpha}^{*} - \omega_{\alpha}^{*} \right)^{2} \left( \omega_{\beta}^{*} - \omega_{\beta}^{*} \right)^{2}, \quad \left( i_{\alpha} = 1; i_{\beta} = 2 \right)
\]

**Figure 2.** From left to right panels rotating two-loop soliton solution to the CIDE Equation (1): For right we depict at times \( t = -30 \) (blue color), \( t = 0 \) (red color) and \( t = 30 \) (black color) corresponding to three moving states, with \( v_{1} = 2, \ v_{2} = 3.33 \) and the computed angular velocities of such wave is \( \Omega_{1} = 0.10 \) and \( \Omega_{2} = 0.15 \), respectively.
where \([N]\) denotes the maximum integer which does not exceed \(N\), \(\mathcal{C}_m\) indicate the summation over all possible combinations of \(m\) elements from \(N\) and \((m)\) indicates the product of all possible combinations of \(m\) elements with \((\alpha < \beta)\). Using the real parameters, we write the phase into two parts as

\[
\eta_n = (k_{n,rc} + \omega_{n,rc} r + \gamma_{n,rc}) + i(k_{n,im} + \omega_{n,im} r + \gamma_{n,im}), \quad n = 1, \ldots, N, \tag{17}
\]

where the real parts and imaginary parts of the parameters \(k_n\) and \(\omega_n\) are obtained using the dispersion relation as

\[
k_{n,rc} = \sqrt{1 - v_n^2 \Omega_n^2} v_n, \quad \omega_{n,rc} = \sqrt{1 - v_n^2 \Omega_n^2},
\]

\[
\omega_{n,im} = \Omega_n, \quad k_{n,im} = -v_n \Omega_n,
\]

here, \(v_n\) and \(\Omega_n\) are the phase velocity and the angular velocity of the soliton, which respect the following condition

\[
v_n > 0, \quad \Omega_n \in \left[-1/\sqrt{v_n^2}, 1/\sqrt{v_n^2}\right]. \tag{19}
\]

Now, let us consider two simple cases: \(N = 3\) and \(N = 4\).

- Case \(N = 3\)

We then write the following expressions of \(F\) and \(Q\) with all coefficients, where \(\exp(\eta_n^*) = 1\). This leads to the three-rotating soliton solution depicted in Figure 3.

**Figure 3.** From left to right panels rotating three-loop soliton solution to the CIDE Equation (1): For left we depict at times \(t = -30\) (blue color), \(t = 0\) (red color) and \(t = 30\) (black color) corresponding to three moving states, with \(v_1 = 2\), \(v_2 = 0.30\), \(v_3 = 0.55\) and the computed angular velocities of such wave is \(\Omega_1 = 0.25\), \(\Omega_2 = 1.00\), \(\Omega_3 = 0.50\), respectively.
\[ F = 1 + \sum_{i=1}^{4} A^i \exp \left( \eta_i + \eta_i^* \right) + \sum_{i=1}^{4} A^{i,j} \exp \left( \eta_i + \eta_i^* + \eta_j + \eta_j^* \right) \]
\[ + \sum_{i=1}^{4} A^{i,j,k} \exp \left( \eta_i + \eta_k + \eta_i^* + \eta_k^* + \eta_j + \eta_j^* \right), \]
\[ Q = \sum_{i=1}^{4} A^i \exp \left( \eta_i + \eta_i^* \right) + \sum_{i=1}^{4} A^{i,j} \exp \left( \eta_i + \eta_j + \eta_i^* + \eta_j^* \right) \]
\[ + \sum_{i=1}^{4} A^{i,j,k} \exp \left( \eta_i + \eta_k + \eta_i^* + \eta_k^* + \eta_j + \eta_j^* \right), \]
\[ A^i, = A^i = 1, \quad (i = 1, 2, 3), \]
\[ A^{i,j} = \frac{1}{4 (\omega_i + \omega_i^*)^2}, \quad \left( i = 1, 2, 3 \right), \]
\[ A^{i,j,k} = 4 A^i A^j \left( \omega_i - \omega_k \right)^2, \quad \left( i = 1, 2; j = i + 1 \right) \]
\[ A^{i,j,k,l} = 4^2 A^i A^j A^k A^l \left( \omega_i - \omega_l \right)^2 \left( \omega_j - \omega_l \right)^2, \quad \left( i = 1, 2; j = i + 1 \right) \]
\[ + \sum_{j=1}^{4} A^{i,j} \exp \left( \eta_i + \eta_j + \eta_i^* + \eta_j^* + \eta_i + \eta_j \right) \]
\[ + \sum_{i=1}^{4} A^{i,j} \exp \left( \eta_i + \eta_j + \eta_i^* + \eta_j^* \right), \quad (i = 1, 2; j = i + 1) \]
\[ A^{i,j,k} = 4^4 A^i A^j A^k A^l \left( \omega_i - \omega_l \right)^2 \left( \omega_j - \omega_l \right)^2 \left( \omega_k - \omega_l \right)^2 \left( \omega_i^* - \omega_l^* \right)^2 \left( \omega_j^* - \omega_l^* \right)^2 \left( \omega_k^* - \omega_l^* \right)^2, \quad \left( i = 1, 2; j = i + 1 \right) \]

\[ (20) \]

• Case \( N = 4 \)
In this case the four-rotating soliton solution is obtain by
\[ F = 1 + \sum_{i=1}^{4} A^i \exp \left( \eta_i + \eta_i^* \right) + \sum_{i=1}^{4} A^{i,j} \exp \left( \eta_i + \eta_j + \eta_i^* + \eta_j^* \right) \]
\[ + \sum_{i=1}^{4} A^{i,j,k} \exp \left( \eta_i + \eta_k + \eta_i^* + \eta_k^* + \eta_j + \eta_j^* \right), \]
\[ Q = \sum_{i=1}^{4} A^i \exp \left( \eta_i + \eta_i^* \right) + \sum_{i=1}^{4} A^{i,j} \exp \left( \eta_i + \eta_j + \eta_i^* + \eta_j^* \right) \]
\[ + \sum_{i=1}^{4} A^{i,j,k} \exp \left( \eta_i + \eta_k + \eta_i^* + \eta_k^* + \eta_j + \eta_j^* \right), \]
\[ A^i, = A^i = 1, \quad (i = 1, 2, 3, 4), \]
\[ A^{i,j} = \frac{1}{4 (\omega_i + \omega_i^*)^2}, \quad \left( i = 1, 2, 3, 4 \right), \]
\[ (21) \]

\[ (22) \]
\[ A_{h,j}^{i_1,i_2} = 4A_h^1A_h^2(\omega_h - \omega_{j_1})(\omega_h - \omega_{j_2})^2, \quad \left \{ i_1 = 1,2,3; \quad i_2 = i_1 + 1 \right \}, \]

\[ A_{h,j_2}^{i_1,i_2} = 4^2A_h^1A_h^2A_h^3A_h^4(\omega_h - \omega_{j_1})(\omega_h - \omega_{j_2})^2, \quad \left \{ i_1 = 1,2,3; \quad i_2 = i_1 + 1 \right \}, \]

\[ A_{h,j_2}^{i_1,i_2} = 4^3A_h^1A_h^2A_h^3A_h^4A_h^5A_h^6(\omega_h - \omega_{j_1})(\omega_h - \omega_{j_2})^2 \times (\omega_h - \omega_{j_1})(\omega_h - \omega_{j_2})^2, \quad \left \{ i_1 = 1,2,3; \quad i_2 = i_1 + 1; i_3 = i_2 + 1 \right \}, \]

\[ A_{h,j_2}^{i_1,i_2} = 4^4A_h^1A_h^2A_h^3A_h^4A_h^5A_h^6A_h^7A_h^8A_h^9(\omega_h - \omega_{j_1})(\omega_h - \omega_{j_2})^2 \times (\omega_h - \omega_{j_1})(\omega_h - \omega_{j_2})^2, \quad \left \{ i_1 = 1,2,3; \quad i_2 = i_1 + 1; i_3 = i_2 + 1; i_4 = i_3 + 1 \right \}, \]

\[ A_{h,j_2}^{i_1,i_2} = 4^{12}A_h^1A_h^2A_h^3A_h^4A_h^5A_h^6A_h^7A_h^8A_h^9A_h^{10}A_h^{11}A_h^{12} \times (\omega_h - \omega_{j_1})(\omega_h - \omega_{j_2})^2 \times (\omega_h - \omega_{j_1})(\omega_h - \omega_{j_2})^2 \times (\omega_h - \omega_{j_1})(\omega_h - \omega_{j_2})^2, \quad \left \{ i_1 = 1; i_2 = 2; i_3 = 3; i_4 = 4 \right \}. \]

**Figure 4** gives the depiction of the four-rotating soliton solutions to the CIDE.

### 4. Summary and Discussion

In this work, we have investigated the CIDE under the viewpoint of Hirota’s bilinearization. Investigating its one- and two-soliton solution, we have come to propose a generalization of such solution to explicit N-soliton solution of the same system. As a matter of illustration, we have provided explicit expressions of 3- and 4-soliton solutions to the CIDE, and have provided figures to enforce our results. In this figures it has appeared clearly that the solution exhibit particle character, since they interact elastically. Since the CIDE is of many physical implications, the N-soliton solution we have obtained is helpful in understanding the propagation of waves in some media such as the propagation electric field in optical fibers, since in Ref. [22] has provided the relation that link the CIDE and the short pulse system. In this work, we have not gone deeply in studying the interaction process between solitons. Such a study will help understand better the interaction process that occurs during the propagation of such waves in some media including optical fibers.
From left to right panels rotating four-loop soliton solution to the CIDE Equation (1): For right we depict at times $t = -90$ (blue color), $t = 0$ (red color) and $t = 90$ (black color) corresponding to three moving states, with $v_1 = 0.91$, $v_2 = 1.82$, $v_3 = 1.33$, $v_4 = 1.11$ and the computed angular velocities of such wave is $\Omega_1 = 0.015$ and $\Omega_2 = 0.050$, $\Omega_3 = 0.090$, $\Omega_4 = 0.990$ respectively.

Acknowledgements

The authors would like to express their sincere thanks to the anonymous referees for their critical comments and appropriate suggestions which have made this paper more precise and readable.

References

[1] Ablowitz, M.J. and Clarkson, P.A. (1991) Solitons, Nonlinear Evolution Equations and Inverse Scattering. Cambridge University Press, New York. https://doi.org/10.1017/CBO9780511623998

[2] Malomet, B.A. (1992) Bound Solitons in Coupled Nonlinear Schrödinger Equations. Physical Review A, 45, 8321-8323. https://doi.org/10.1103/PhysRevA.45.R8321

[3] Nakkeeran, D.J. and Porsesian, K.J. (1995) Solitons in an Erbium-Doped Nonlinear Fibre Medium with Stimulated Inelastic Scattering. Journal of Physics A: Mathematical and General, 28, 3817-3823. https://doi.org/10.1088/0305-4470/28/13/025

[4] Konopelchenko, B.G. and Magri, F. (2007) Coisotropic Deformations of Associative Algebras and Dispersionless Integrable Hierarchies. Communications in Mathematical Physics, 274, 627-658. https://doi.org/10.1007/s00220-007-0295-2

[5] Kakuhta, H. and Konno, K. (1997) Canonical Formulation of a Generalized Coupled Dispersionless System. Journal of Physics A: Mathematical and General, 30L, 401-407. https://doi.org/10.1088/0305-4470/30/12/002

[6] Aoyama, S. and Kodama, Y. (1994) Topological Conformal Field Theory with a Rational Potential and the Dispersionless KP Hierarchy. Modern Physics Letters A, 9, 2481-2492. https://doi.org/10.1142/S0217732394002355

[7] Hassan, M.J. (2009) Darboux Transformation of the Generalized Coupled Dispersionless Integrable System. Journal of Physics A: Mathematical and Theoretical, 42, 1-11.
[8] Konno, K. and Oono, H. (1994) New Coupled Integrable Dispersionless Equations. *Journal of the Physical Society of Japan, 63*, 377-378. https://doi.org/10.1143/JPSJ.63.377

[9] Kakuhata, H. and Konno, K. (1996) A Generalization of Coupled Integrable, Dispersionless System. *Journal of the Physical Society of Japan, 65*, 340-341. https://doi.org/10.1143/JPSJ.65.340

[10] Kotlyarov, V.P. (1994) On Equations Gauge Equation Uivalent to the Sine-Gordon and Pohlmeyer-Lund-Regge Equations. *Journal of the Physical Society of Japan, 63*, 3535-3537. https://doi.org/10.1143/JPSJ.63.3535

[11] Novikov, S.P., Manakov, S.V., Pitaevskii, L.P. and Zakkarov, V.E. (1984) Theory of Solitons, The Inverse Scattering Method. Consultant Bureau, New York.

[12] Konno, K. and Kakuhata, H. (1996) Novel Solitonic Evolutions in a Coupled Integrable, Dispersionless System. *Journal of the Physical Society of Japan, 65*, 713-721. https://doi.org/10.1143/JPSJ.65.713

[13] Levi, D. (1988) On a New Darboux Transformation for the Construction of Exact Solutions of the Schrodinger Equation. *Inverse Problem, 4*, 165-172. https://doi.org/10.1088/0266-5611/4/1/014

[14] Matveev, V.B. and Salle, M.A. (1991) Darboux Transformations and Solitons. Springer, Berlin. https://doi.org/10.1007/978-3-662-00922-2

[15] Hirota, R. (2004) The Direct Method in Soliton Theory. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511543043

[16] Hirota, R. and Ohtam, Y. (1991) Hierarchies of Coupled Soliton Equations. I. *Journal of the Physical Society of Japan, 60*, 798-809. https://doi.org/10.1143/JPSJ.60.798

[17] Ma, W.X. and You, Y.C. (2005) Solving the Korteweg-de Vries Equation by Its Bilinear Form: Wronskian Solutions. *Transactions of the American Mathematical Society, 357*, 1753-1778. https://doi.org/10.1090/S0002-9947-04-03726-2

[18] Ma, W.X. and Maruno, K.I. (2004) Complexiton Solutions of the Toda Lattice Equation. *Physica A, 343*, 219-237. https://doi.org/10.1016/j.physa.2004.06.072

[19] Geng, X.G. (2003) Algebraic-Geometrical Solutions of Some Multidimensional Nonlinear Evolution Equations. *Journal of Physics A—Mathematical and General, 36*, 2289-2303. https://doi.org/10.1088/0305-4470/36/9/307

[20] Kakuhata, H. and Konno, K. (1999) Loop Soliton Solutions of String Interacting with External Field. *Journal of the Physical Society of Japan, 68*, 757-762. https://doi.org/10.1143/JPSJ.68.757

[21] Kakuhata, H. and Konno, K. (2002) Rotating Loop Soliton of the Coupled Dispersionless Equations. *Theoretical and Mathematical Physics, 133*, 1675-1683. https://doi.org/10.1023/A:1021366309313

[22] Kuetche, V.K., Bouetou, T.B. and Kofane, T.C. (2008) On Exact N-Loop Soliton Solution to Nonlinear Coupled Dispersionless Evolution Equations. *Physics Letters A, 372*, 665-669. https://doi.org/10.1016/j.physleta.2007.08.023

[23] Kuetche, V.K., Bouetou, T.B. and Kofane, T.C. (2007) Comment on Loop Soliton Solutions of String Interacting with External Field? *Journal of the Physical Society of Japan, 76*, Article ID: 126001. https://doi.org/10.1143/JPSJ.76.126001
Influence of the Foundation on the Threshold of Stability for Rotating Machines with Roller Bearings—A Theoretical Analysis

Ulrich Werner

Faculty of Electrical Engineering, Precision Engineering, Information Technology, Georg Simon Ohm University of Applied Sciences, Nuremberg, Germany
Email: ulrich.werner@th-nuernberg.de

Abstract

The paper presents a mathematical model for analyzing the threshold of stability for rotating machines, where the rotor is linked to the stator by roller bearings, bearing housings and end-shields and where the stator feet are mounted on a soft foundation. The internal (rotating) damping of the rotor is the only source of instability, which is considered in the paper. After the mathematical coherences of the multibody model are described, a procedure is presented for deriving the threshold of stability. Additionally, a numerical example is shown, where the threshold of stability is calculated for different boundary conditions. It could be demonstrated, that the stiffness of the foundation—even if the foundation stiffness is isotropic—can help stabilizing this kind of vibration system in the same way as orthotropic bearing stiffness or orthotropic bearing housing and end-shield stiffness for a rigid foundation.

Keywords

Rotordynamics, Instability, Roller Bearings, Rotating Damping

1. Introduction

When designing rotating machines, it is important to calculate the vibration behavior and to consider the influence of the foundation [1]-[9]. Beside the forced vibrations—due to e.g. unbalance—also self-excited vibrations have to be considered. There are many effects, which cause self-excited vibrations, e.g. not equal cross-coupling stiffness coefficient in the oil film of sleeve bearings, steam excitations in steam turbines, electromagnetic field damping effects in induction motors, and internal (rotating) damping of the rotor shaft, referring to [3] [4] [5] and [8]. When designing rotating machines, it is important to know, at
which rotor speed the threshold of rotor stability is reached. If this rotor speed is exceeded, self-excited vibrations are caused, occurring with a natural frequency of the system. This threshold of stability can be pushed to higher rotor speeds, if external damping is added to the rotor, e.g. by squeeze film dampers. But also orthotropic bearing and/or orthotropic support stiffness help to increase the threshold of stability, referring to [1] [2] [3] and [5]. The aim of the paper is now to derive a vibration model for a special kind of rotating machine, where the rotor is linked to the stator by roller bearings, bearing housings and end-shields and where the stator feet are mounted on a soft foundation, so that the centre of gravity of the stator is displaced by the height $h$ from the foundation (Figure 1). A soft foundation may be realized by e.g. rubber elements, where the machine is mounted, or by a steel frame foundation, because steel frame foundations are often very flexible, because of economically reasons. Therefore, in the model not only the rotor, the bearings and the support of the bearings are considered, but also the mass and inertia of the stator at its centre of gravity, and the foundation under the machine feet.

2. Vibration Model

The vibration model is a simplified model, which describes the movement in the $yz$-plane (Figure 2). The model is generally based on the model in [9], but modified especially for rotating machines with roller bearings instead of sleeve bearings. The model covers a wide range of rotating machines, and not only electrical machines. Therefore no electromagnetism is here considered, contrarily to [9], where electromagnetic field damping is in the focus. However, the most important difference to [9] is that in this paper here not forced vibrations are analyzed but self-exciting vibrations due to instability, caused by internal (rotating) damping of the rotor shaft.

The vibrations system consists of two main masses, the rotor mass $m_w$, which is concentrated as a lumped mass in the middle between the two bearings, and the stator mass $m_s$, which is concentrated in the centre of gravity $S$ of the stator with the mass inertia $\theta_s$.
Figure 2. Vibration model.

Beside these two main masses, two additionally masses are considered, the mass of the shaft journal $m_v$ and the mass of the bearing housing $m_b$, mostly to avoid zeros at the main diagonal of the mass matrix. The rotor has the rotor stiffness $c$ and the internal damping $d_i$ and rotates with the rotary angular frequency $\Omega$. The rotor is connected to the end-shields by bearing housings and roller bearings, which suppose to be equal for each machine side. Many methods and strategies are described in literature to derive the stiffness of roller bearings, e.g. [10]-[20]. In this paper, a simplified bearing model is used, where the stiffness of the roller bearings is described by the roller bearing stiffness matrix $C_r$, with the vertical bearing stiffness $c_{rz}$ and horizontal bearing stiffness $c_{ry}$. Cross coupling coefficients of the roller bearings are neglected as well as damping of the roller bearings. The stiffness and damping of the bearing housing and end-shields is described by the bearing housing and end-shield stiffness and damping matrix $C_b$ and $D_b$, which also suppose to be equal for each machine side. The stator structure is here assumed to be very stiff, compared to the foundation stiffness, so the stator structure can be modeled rigid. The stator feet - $F_L$ (left side) and $F_R$ (right side) - are connected to the ground by the foundation stiffness and damping matrix $C_f$ and $D_f$, which are also assumed to be equal for the right side and left side of the machine. When deriving the damping coefficients, it has to be considered, that the natural vibration of the critical mode occurs with the angular natural frequency $\omega_{stab}$ at the threshold of stability, which is the rotary angular frequency $\Omega_{stab}$. Therefore, the whirling angular frequency $\omega_w$ of the rotor becomes $\omega_{stab}$, at the rotary angular frequency of $\Omega = \Omega_{stab}$:

$$\omega_w = \omega_{stab}$$ (1)
The internal material damping of the rotor $d_i$ can be described by the stiffness of the rotor $c$ and mechanical loss factor $\tan \delta_i$ of the rotor, depending on the whirling angular frequency $\omega_F$, referring to [3]:

$$d_i(\omega_F) = \frac{c \cdot \tan \delta_i}{\omega_F}$$  \hspace{1cm} (2)

The same approach is deduced for the damping coefficients of the bearing housing end end-shield and of the foundation:

$$d_{bh}(\omega_F) = \frac{c_{bh} \cdot \tan \delta_h}{\omega_F} ; \quad d_{by}(\omega_F) = \frac{c_{by} \cdot \tan \delta_y}{\omega_F}$$  \hspace{1cm} (3)

$$d_{bf}(\omega_F) = \frac{c_{bf} \cdot \tan \delta_f}{\omega_F} ; \quad d_{fy}(\omega_F) = \frac{c_{fy} \cdot \tan \delta_y}{\omega_F}$$  \hspace{1cm} (4)

With the stiffness of the bearing housing end end-shield $c_{bh}$ and $c_{by}$ and the stiffness of the foundation at each machine side (left and right side) $c_{bf}$ and $c_{fy}$ and the loss factor of the bearing housing and end-shield $\tan \delta_h$ and of the foundation $\tan \delta_f$.

### 3. Mathematical Model

To get the threshold of stability, it is necessary to derive the homogenous differential equation by separating the vibration system into four single systems: a) rotor mass system, b) journal system, c) bearing house system and d) stator mass system (Figure 3).

**a) Rotor mass system**

**b) Journal system**

**c) Bearing house system**

**d) Stator mass system**

Note: The negative vertical displacement in $z$, relating to the coordinate $x_L$, is considered by the direction of the vertical forces in $F_L$, so $z_L$ has to be described in the differential equation by: $z_L = -z_L + \varphi_L \cdot \hat{h}$

**Figure 3.** Vibration system cut free into subsystems.
The displacements of the stator mass \((z_s, y_s, \varphi_s)\) is small, compared to the dimensions of the machine \((h, b, \Psi_s)\), therefore following linearization is possible:

\[
\begin{align*}
    z_{zf} &= z_s - \varphi_s \cdot h; \\
    z_{fr} &= z_s + \varphi_s \cdot h; \\
    y_{fr} &= y_s = y_s - \varphi_s \cdot h
\end{align*}
\]  

\[
(5)
\]

The homogenous differential equation system can be derived by analyzing the equilibrium of at each single system:

\[
M \cdot \ddot{q} + D \cdot \dot{q} + C \cdot q = 0
\]

\[
(6)
\]

with the coordinate vector \(q\):

\[
q = [z_s; z_u; y_u; \varphi_s; z_b; z_y; y_b]^T
\]

\[
(7)
\]

with the mass matrix \(M\):

\[
M = \begin{bmatrix}
    m_s & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
    0 & m_u & 0 & 0 & 0 & 0 & 0 & 0 \\
    0 & 0 & m_s & 0 & 0 & 0 & 0 & 0 \\
    0 & 0 & 0 & m_u & 0 & 0 & 0 & 0 \\
    0 & 0 & 0 & 0 & \Theta_{tt} & 0 & 0 & 0 \\
    0 & 0 & 0 & 0 & 0 & 2m_y & 0 & 0 \\
    0 & 0 & 0 & 0 & 0 & 0 & 2m_y & 0 \\
    0 & 0 & 0 & 0 & 0 & 0 & 0 & 2m_y
\end{bmatrix}
\]

\[
(8)
\]

with the damping matrix \(D\):

\[
D = \begin{bmatrix}
    2(d_{zf} + d_{bc}) & 0 & 0 & 0 & 0 & 0 & -2d_{bc} & 0 & 0 \\
    0 & d_i & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
    0 & 0 & 2(d_{fy} + d_{by}) & 0 & -2d_{fy} \cdot h & 0 & 0 & 0 & -2d_{by} \\
    0 & 0 & 0 & d_i & 0 & 0 & 0 & 0 & 0 \\
    0 & 0 & -2d_{fy} \cdot h & 0 & 2(d_{fy} h^2 + d_{fy} b^2) & 0 & 0 & 0 & 0 \\
    0 & -d_i & 0 & 0 & 0 & d_i & 0 & 0 & 0 \\
    -2d_{bc} & 0 & 0 & 0 & 0 & 0 & 2d_{bc} & 0 & 0 \\
    0 & 0 & 0 & -d_i & 0 & 0 & 0 & d_i & 0 \\
    0 & 0 & -2d_{by} & 0 & 0 & 0 & 0 & 0 & 2d_{by}
\end{bmatrix}
\]

\[
(9)
\]

with the stiffness matrix \(C\):

\[
C = \begin{bmatrix}
    2(c_{zf} + c_{bc}) & 0 & 0 & 0 & 0 & 0 & -2c_{bc} & 0 & 0 \\
    0 & c & 0 & \Omega d_i & 0 & -c & 0 & -\Omega d_i & 0 \\
    0 & 0 & 2(c_{fy} + c_{by}) & 0 & -2c_{fy} h & 0 & 0 & 0 & -2c_{by} \\
    0 & -\Omega d_i & 0 & c & 0 & \Omega d_i & 0 & -c & 0 \\
    0 & 0 & -2c_{fy} h & 0 & 2(c_{fy} h^2 + c_{fy} b^2) & 0 & 0 & 0 & 0 \\
    0 & -c & 0 & -\Omega d_i & 0 & 2c_{fy} + c & -2c_{fy} & \Omega d_i & 0 \\
    -2c_{bc} & 0 & 0 & 0 & 0 & -2c_{bc} & 2(c_{bc} + c_{be}) & 0 & 0 \\
    0 & \Omega d_i & 0 & -c & 0 & -\Omega d_i & 0 & 2c_{by} + c & -2c_{by} \\
    0 & 0 & -2c_{by} & 0 & 0 & 0 & 0 & -2c_{by} & 2(c_{by} + c_{bc})
\end{bmatrix}
\]

\[
(10)
\]
The internal (rotating) damping \( d_i \) of the rotor in conjunction with the rotary angular frequency \( \Omega \) leads here to an anti-symmetric stiffness matrix, which causes instability, when the threshold of stability is exceeded (\( \Omega > \Omega_{\text{stab}} \)).

The limit of vibration stability \( \Omega_{\text{stab}} \) can be calculated, when increasing the rotary angular frequency \( \Omega \), and analyzing the eigenvalues. If a real part of one eigenvalue gets zero, the limit of vibration stability is reached. Increasing the rotary angular frequency \( \Omega \) furthermore will cause a positive real part and the vibration system gets instable. Using the state-space formulation

\[
\begin{bmatrix}
\ddot{q}_b \\
q_b
\end{bmatrix}_x = \begin{bmatrix}
0 & I \\
-M^{-1} \cdot C & -M^{-1} \cdot D
\end{bmatrix}
\begin{bmatrix}
\dot{q}_b \\
q_b
\end{bmatrix}_x
\]

the eigenvalues can be derived. With the formulation \( x = \dot{x} \cdot e^{i \xi} \), the eigenvalues are calculated by:

\[
\det[A - \lambda \cdot I] = 0
\]

At the threshold of stability the eigenvalue \( \lambda \) of the critical mode gets:

\[
\lambda = \lambda_{\text{stab}} = \pm j \cdot \omega_{\text{stab}}
\]

The real part of the critical eigenvalue \( \lambda_{\text{stab}} \) is zero and the whirling angular frequency \( \omega_F \) is then identical to \( \omega_{\text{stab}} \), while the rotor is rotating with \( \Omega_{\text{stab}} \). Considering, that the coefficients \( d_i \), \( d_{iz} \), \( d_{yz} \), \( d_{zy} \), \( d_{yz} \) are depending on the whirling angular frequency \( \omega_F \), an iterative solution has to be deduced, according to Figure 4.

**Figure 4.** Flow diagram to derive the threshold of stability.
First, a start value of the whirling angular frequency \( \omega_p = \omega_{stab,0} \) has to be estimated. This can be done e.g. by following estimation, which is based on a ridged mounted machine, without external damping and with the assumption that \( c_{ry} < c_{re} \) and \( c_{by} < c_{be} \) and that the first natural angular frequency \( \omega_{y,0} \) is here the whirling angular frequency at the threshold of rotor stability:

\[
\omega_{stab,0} = \omega_{y,0} = \sqrt{\frac{c_{total}}{m_w}} \quad \text{with:} \quad c_{total} = \frac{1}{c + \frac{1}{2c_{ry}} + \frac{1}{2c_{by}}} \tag{14}
\]

With this assumption the damping coefficients \( d_1, d_{ke}, d_{by}, d_{fe}, d_{ff} \) can be derived, and therefore also the threshold of stability and the natural angular frequency, leading to \( \Omega_{stab,1} \) and \( \omega_{stab,1} \). With this new angular whirling frequency \( \omega_p = \omega_{stab,1} \) the damping coefficients \( d_1, d_{ke}, d_{by}, d_{fe}, d_{ff} \) are calculated again, leading to a new threshold of stability \( \Omega_{stab,2} \) and a new natural angular frequency \( \omega_{stab,2} \). If the ratio \( \left| \frac{\omega_{stab,2} - \omega_{stab,1}}{\omega_{stab,1}} \right| \) is less than \( \Delta \) - an arbitrarily chosen value - the calculation is finished and \( \Omega_{stab} = \Omega_{stab,2} \) and \( \omega_{stab} = \omega_{stab,2} \). If the ratio is larger as the chosen value \( \Delta \), a loop has to be run through till the ratio is less than \( \Delta \).

4. Numerical Example

Based on the mathematical derivation, a numerical example is shown, where the threshold of stability is analyzed.

4.1. Boundary Conditions

The rotating machine consists of a rotor, roller bearings, bearing housings, end-shields and a stator (Figure 1), which is mounted on a welded steel frame foundation. The data of the rotating machine, roller bearings and foundation is shown in Table 1.

4.2. Analysis of Natural Vibrations and Threshold of Stability

In Figure 5 the real part and the imaginary part of the eigenvalues are presented, depending on the rotor speed.

It can be shown, that at a rotor speed of about 26130 rpm the real part \( \alpha_3 \) becomes zero and therefore the threshold of stability is reached. The corresponding natural angular frequency is \( \omega_3 = 394.3 \text{ rad/s} \), which is equal to the whirling angular frequency \( \omega_p = \omega_{stab} \) at the limit of stability of the critical mode, which is here mode 3. Increasing the rotor speed above 26130 rpm, leads to instability of the vibration system.

Figure 6 shows the different mode shapes at the threshold of stability \( n_{stab} = 26130 \text{ rpm} \). Because of the clarity, only the orbits of the rotor mass, stator mass and machine feet are shown, and not the orbits of the shaft journal points and the bearing housing points. As it can be seen, all eigenvalues of the mode shapes have negative real parts, except mode 3, where the real part \( \alpha_3 \) is
Table 1. Data of rotating machine, roller bearings and foundation.

| Machine data                | Description                              | Value         |
|-----------------------------|------------------------------------------|---------------|
| Mass of the stator          | $m_s = 3900$ kg                          |               |
| Mass inertia of the stator at x-axis | $\theta_s = 530$ kg·m²                |               |
| Mass of the rotor           | $m_r = 930$ kg                           |               |
| Mass of the rotor shaft journal | $m_i = 5$ kg                      |               |
| Mass of the bearing housing | $m_b = 20$ kg                           |               |
| Height of the centre of gravity | $h = 450$ mm                      |               |
| Distance between motor feet | $2b = 850$ mm                          |               |
| Stiffness of the rotor      | $c = 1.72 \times 10^7$ kg/s²            |               |
| Horizontal stiffness of bearing housing and end shield | $c_{h} = 7.0 \times 10^7$ kg/s²         |               |
| Vertical stiffness of bearing housing and end shield | $c_{v} = 7.0 \times 10^7$ kg/s²        |               |
| Mechanical loss factor of bearing housing and end shield | $\tan \delta = 0.04$                   |               |
| Mechanical loss factor of the rotor | $\tan \delta = 0.03$                   |               |

| Bearing data                | Description                              | Value         |
|-----------------------------|------------------------------------------|---------------|
| Bearing type                | Ball bearing: Type 6220 C3               |               |
| Horizontal stiffness of the roller bearing | $c_{r} = 2.0 \times 10^7$ kg/s²        |               |
| Vertical stiffness of the roller bearing | $c_{r} = 2.0 \times 10^7$ kg/s²       |               |

| Foundation data             | Description                              | Value         |
|-----------------------------|------------------------------------------|---------------|
| Type of foundation          | Welded steel frame foundation             |               |
| Vertical stiffness of the foundation at each motor side | $c_{f} = 1.5 \times 10^7$ kg/s²       |               |
| Horizontal stiffness of the foundation at each motor side | $c_{f} = 1.0 \times 10^7$ kg/s²      |               |
| Mechanical loss factor of the foundation | $\tan \delta = 0.04$                   |               |

zero. When increasing the rotor speed furthermore, this real part $\alpha_3$ gets positive. Therefore mode 3 is the critical mode shape.

4.3. Variation of Single Parameters

Now different cases are investigated, and the threshold of stability $n_{\text{stab}}$ is calculated as well as the natural angular frequency $\omega_{\text{stab}}$ at the threshold of stability (Table 2).

Table 2 shows, that neglecting the damping of the bearing housings and end shields (case b) only decreases here the threshold of stability $n_{\text{stab}}$ marginal ($-1.07\%$). Without foundation damping (case c) a clearly reduction of $n_{\text{stab}}$ is obvious ($-4.82\%$). A strong reduction occurs, if the foundation would be rigid (cases d). Here the threshold of stability occurs already at a rotor speed of 3840 rpm, which means a reduction of $-85.3\%$. 
Figure 5. Eigenvalues, depending on the rotor speed and threshold of stability.

If then the bearing stiffness would be changed from isotropic \( c_{rs} = c_{rs} = 2.0 \times 10^4 \text{ kg/s}^2 \) to orthotropic \( c_{rs} \neq c_{rs} ; \ c_{rs} = 1.5 \times 10^8 \text{ kg/s}^2 \) and \( c_{rs} = 2.5 \times 10^8 \text{ kg/s}^2 \), the threshold of stability can be increased again up to 13020 rpm (case e).

4.4. Arbitrarily Variation of Foundation Stiffness

In this section, the influence of the foundation stiffness on the threshold of stability \( n_{stab} \) and on the whirling angular frequency \( \omega_{stab} \) is analyzed.

Therefore, the foundation stiffness is varied from the rated values in Table 1 with factors between 0.2 and 5, which means, that the foundation stiffness is varied in a range between \( 2 \times 10^7 \text{ kg/s}^2 \) and \( 7.5 \times 10^7 \text{ kg/s}^2 \) (Figure 7).
Figure 6. Mode shapes at the threshold of stability with $n_{stab} = 26130$ rpm.

Table 2. Threshold of stability for different cases.

| Case Description | $\omega_{stab}$ [rad/s] | $n_{stab}$ [rpm] | $\Delta n_{stab}$ to $a$ [%] |
|------------------|-------------------------|------------------|-----------------------------|
| a) Basic Data (Data Table 2) | 394.26 | 26130 | 0 |
| b) Data Table 2 with $d_{in} = d_{ou} = 0$ (No damping of the bearing housings and end shields) | 393.75 | 25850 | −1.07 |
| c) Data Table 2 with $d_{in} = d_{ou} = 0$ (No damping of the foundation) | 391.20 | 24870 | −4.82 |
| d) Data Table 2 with $c_v = c_e \rightarrow \infty$ (Infinitely stiff foundation) | 344.85 | 3840 | −85.3 |
| e) Data Table 2 with $c_v = c_e \rightarrow \infty$ and $c_v = 1.5 \times 10^7 \text{ kg/s}^2$; $c_e = 2.5 \times 10^7 \text{ kg/s}^2$ (Infinitely stiff foundation and orthotropic bearing stiffness) | 340.69 | 13020 | −50.2 |
4.5. Arbitrarily Variation of Bearing Stiffness for the Soft Foundation

Now, the influence of bearing stiffness is analyzed for the rated soft foundation (Table 1). Therefore, the bearing stiffness is varied from the rated values in Table 1 by ±50%, which means that the bearing stiffness is varied in a range between $1 \times 10^8$ kg/s$^2$ and $3 \times 10^8$ kg/s$^2$, also considering orthotropic bearing stiffness ($c_{p} \neq c_{q}$) (Figure 8).

Figure 7. Influence of the foundation stiffness on (a) the limit of stability $n_{stab}$ and on (b) the whirling angular frequency $\omega_{stab}$.
Figure 8. Influence of the bearing stiffness on (a) the limit of stability $n_{stab}$ and on (b) the whirling angular frequency $\omega_{stab}$ for the rated soft foundation (Table 1).

4.6. Arbitrarily Variation of Bearing Housing and End-Shield Stiffness for the Soft Foundation

In this section, the influence of bearing housing and end-shield stiffness is analyzed, for the rated soft foundation (Table 1). Therefore, the bearing housing and end-shield stiffness is varied from the rated values in Table 1 also by ±50%, which means that the bearing housing and end-shields stiffness is varied in a range between $3.5 \times 10^8$ kg/s$^2$ and $1.05 \times 10^9$ kg/s$^2$, also considering orthotropic bearing housing and end-shield stiffness ($c_{bx} \neq c_{by}$) (Figure 9).
Figure 9. Influence of the bearing housing and end-shield stiffness on (a) the limit of stability $n_{\text{stab}}$ and on (b) the whirling angular frequency $\omega_{\text{stab}}$ for the rated soft foundation (Table 1).

4.7. Arbitrarily Variation of Bearing Stiffness for a Rigid Foundation

Here, the influence of the bearing stiffness is analyzed again, but now for a rigid foundation \( (c_\beta = c_\gamma \rightarrow \infty) \). Therefore, the bearing stiffness is again variated in a range between \( 1.0 \times 10^8 \text{ kg/s}^2 \) and \( 3.0 \times 10^8 \text{ kg/s}^2 \) (Figure 10).

4.8. Arbitrarily Variation of Bearing Housing and End-Shield Stiffness for a Rigid Foundation

In this section, the influence of bearing housing and end-shield stiffness is analyzed again, but for a rigid soft foundation \( (c_\beta = c_\gamma \rightarrow \infty) \). Therefore, the
Figure 10. Influence of the bearing stiffness on (a) the limit of stability \( n_{\text{stab}} \) and on (b) the whirling angular frequency \( \omega_{\text{stab}} \) for a rigid foundation.

Figure 10 shows, that for a rigid foundation, the threshold of stability \( n_{\text{stab}} \) and on the whirling angular frequency \( \omega_{\text{stab}} \) is increased clearly, if orthotropic bearing stiffness \( c_{rz} \neq c_{r\theta} \) exists, which is also
Figure 11. Influence of the bearing housing and end-shield stiffness on (a) the limit of stability $n_{stab}$ and on (b) the whirling angular frequency $\omega_{stab}$ for a rigid foundation.

described in literature ([3] [4] [5] and [8]). In this paper the bearing stiffness is variated in the range of ±50%, leading to a maximum threshold of stability of about 29,000 rpm. The same effect is caused, if the bearing housing and end-shield stiffness gets orthotropic ($c_{tx} \neq c_{ty}$), which can be seen in Figure 11. Here the stiffness is also variated in the range of ±50%, but only leading to a maximum threshold of stability of about 9900 rpm. The reason is, that both stiffness, bearing stiffness and bearing housing and end-shield stiffness are connected in series, and the rated bearing stiffness is much lower than the rated bearing housing and end-shield stiffness ($c_{tx} = c_{ty} = 2.0 \times 10^8 \text{ kg/s}^2 < c_{tx} = c_{ty} = 7.0 \times 10^8 \text{ kg/s}^2$).

The innovation of the paper is now, that it can be demonstrated (Figure 7), that the threshold of stability can also be increased by a soft foundation, even if
the foundation stiffness is isotropic \( c_{f_z} = c_{f_y} \). The reason is the kind of rotating machine, with a stator, mounted with its feet on a soft foundation, so that the centre of gravity of the stator is displaced by the height \( h \) from the foundation (Figure 1). This leads to different mode shapes (Figure 6), which cause a similar effect on the threshold of stability as orthotropic bearing stiffness or orthotropic bearing housing and end-shield stiffness, for a rigid foundation.

In this example, the threshold of stability could be increased even to maximum of about 143000 rpm, at a foundation stiffness of \( c_{f_z} = 5.54 \times 10^8 \text{ kg/s}^2 \) and \( c_{f_y} = 7.5 \times 10^8 \text{ kg/s}^2 \) (Figure 7). Increasing the foundation stiffness furthermore in the considered range, leads to a decrease of the threshold of stability, which can be seen in Figure 7. If the foundation stiffness would be increased to infinite, the threshold of stability would drop to 3840 rpm (Table 2; case d). But it has to be considered here, that a boundary condition of the model is, that the stiffness of the stator structure is much higher than the foundation stiffness, so that the stator structure is assumed to be rigid. As a rough estimation: Up to a foundation stiffness of about \( c_{f_z} \leq 5.0 \times 10^8 \text{ kg/s}^2 \) and \( c_{f_y} \leq 5.0 \times 10^8 \text{ kg/s}^2 \), this boundary condition is acceptable for this example, above this values the elasticity of the stator structure has to be considered. The influence of bearing stiffness and bearing housing and end-shield stiffness on the threshold of stability for the rated soft foundation is also demonstrated in Figure 8 and Figure 9. In Figure 8 the maximum threshold of stability of about 65200 rpm is reached at a bearing stiffness of \( c_{r_z} = 3.0 \times 10^8 \text{ kg/s}^2 \) and \( c_{r_y} = 1.0 \times 10^8 \text{ kg/s}^2 \). In Figure 9 the maximum threshold of stability of about 37800 rpm is reached at a bearing housing and end-shield stiffness of \( c_{b_z} = 1.05 \times 10^9 \text{ kg/s}^2 \) and \( c_{b_y} = 3.5 \times 10^9 \text{ kg/s}^2 \).

Most of the calculated thresholds of stability are fare above the limit of the roller bearings and fare above the limit, what the rotor structure would stand. Additionally it has to be noticed, that with increasing rotor speed, higher bending modes of the rotor become more and more important and therefore also the gyroscopic effect. But of course, this analysis helps to estimate, whether within the rotor speed limits of the roller bearing and of the rotor structure an instability would occur or not, if higher bending modes of the rotor and gyroscopic effects can be neglected.

5. Conclusion

The paper presents a mathematical model especially for analyzing the threshold of stability for a special kind of rotating machines, consisting of a rotor, stator, end-shields, bearing housings and roller bearings, mounted on a soft foundation, so that the centre of gravity of the stator is displaced by the height \( h \) from the foundation (Figure 1). After the mathematical coherences of the model have been described, a procedure was presented for deriving the threshold of stability. Additionally, a numerical example was shown, where the threshold of stability was calculated for different boundary conditions. The influence of the stiffness of the foundation, of the bearings and of the bearing housings and end-shields was demonstrated, as well as the influence of the damping of the foundation and
the damping of the bearing housings and end-shields on the threshold of stability. The main task and the innovation of the paper are to demonstrate that for this kind of rotating machines, the stiffness of the soft foundation—even if the foundation stiffness is isotropic—can help stabilizing the vibration system and therefore leading to a similar effect as orthotropic bearing stiffness or orthotropic bearing housing and end-shield stiffness for a rigid foundation. Of course, the presented model is a simplified model of the system, but the conclusions and the procedure for deriving the threshold of stability can also be applied in a finite element analysis. As a future work, experimental validation of the presented theory may be deduced, based e.g. on a small induction motor, to demonstrate the stabilization influence of the foundation.

References

[1] Kirk, R.G., De Chowdhury, P. and Gunter, E.J. (1974) The Effect of Support Flexibility on the Stability of Rotors Mounted in Plain Cylindrical Bearings. Proceedings of IUTAM Symposium Dynamics of Rotors, Lyngby, Denmark, 244-298.

[2] Gasch, R., Maurer, J. and Sarfeld, W. (1984) The Influence of the Elastic Half Space on Stability and Unbalance of a Simple Rotor-Bearing Foundation System. Proceedings of Conference Vibration in Rotating Machinery, C300/84, Edinburg, UK, 1-12.

[3] Gasch, R., Nordmann, R. and Pfützner, H. (2002) Rotordynamik. Springer-Verlag, Berlin-Heidelberg.

[4] Childs, D.W. (1993) Turbomachinery Rotordynamics: Phenomena, Modeling, and Analysis. John Wiley & Sons Inc., New York.

[5] Genta, G. (2005) Dynamics of Rotating Systems. Springer Science & Business Media, New York. https://doi.org/10.1007/0-387-28687-X

[6] Kuemmlee, H., Siegl, G. and Woywode, P. (2008) Influence of Elastic Foundation Structures on the Rotor Dynamics of Drive Trains. Proceedings of Petrol and Chemical Industry Conference EUROPE-Electrical and Instrumentation Applications, Weimar, Germany, 1-8.

[7] Ehehalt, U., Luneburg, B., Staubach, R., Daniel, C., Stackeljan, J. and Woschke, E. (2009) Methods to Incorporate Foundation Elasticities in Rotordynamic Calculations. Proceedings of SIRM 2009 8th International Conference on Vibrations in Rotating Machines, Vienna, Austria, 10 p.

[8] Krämer, E. (1993) Dynamics of Rotors and Foundations. Springer-Verlag, Berlin-Heidelberg. https://doi.org/10.1007/978-3-662-02798-1

[9] Werner, U. (2017) Mathematical Multibody Model of a Soft Mounted Induction Motor Regarding Forced Vibrations due to Dynamic Rotor Eccentricities Considering Electromagnetic Field Damping. Journal of Applied Mathematics and Physics, 5, 346-364. https://doi.org/10.4236/jamp.2017.52032

[10] Hamrock, B.J. and Anderson, W.J. (1973) Analysis of an Arched Outer-Race Ball Bearing Considering Centrifugal Forces. ASME Journal of Lubrication Technology, 95, 265-276.

[11] Hernot, X., Sartor, M. and Guillot, J. (2000) Calculation of the Stiffness Matrix of Angular Contact Ball Bearings by Using the Analytical Approach. Journal of Mechanical Design, 122, 83-90. https://doi.org/10.1115/1.533548

[12] Harris, T.A. (2001) Rolling Bearing Analysis. 4th Edition, John Wiley & Sons, New York.
York.

[13] Dougdag, M., Titouche, N.E., Djaoui, M. and Quali, M. (2008) The Calculation of Ball Bearing Nonlinear Stiffness Theoretical and Experimental Study with Comparisons. *Journal of Engineering and Applied Science*, 3, 872-883.

[14] Guo, Y. and Parker, R.G. (2012) Stiffness Matrix Calculation of Rolling Element Bearings Using a Finite Element/Contact Mechanics Model. *Mechanism and Machine Theory*, 51, 32-45.

[15] Bauer, C. and Werner, U. (2016) Method to Detect the Critical Bending Speed of an Induction Motor by Using Static Rotor Eccentricity for Deriving the Roller Bearing Stiffness. *Proceedings of 11th International Conference on Vibrations in Rotating Machinery*, Manchester, 177-185.

[16] Guay, P. and Frikha, A. (2015) Ball Bearing Stiffness. A New Approach Offering Analytical Expressions. *Proceedings of 16th European Space Mechanisms and Tribology Symposium*, Bilbao, 23-25 September 2015.

[17] Wu, W., Hong, J., Li, Y. and Li, X. (2017) Investigation of Non-Uniform Preload Effect on Stiffness Behavior of Angular Contact Ball Bearings. *Advances in Mechanical Engineering*, 9, 1-19. https://doi.org/10.1177/1687814017694118

[18] Nonato, F. and Cavalca, K.L. (2014) An Approach for Including the Stiffness and Damping of Elastohydrodynamics Point Contacts in Deep Groove Ball Bearing Equilibrium Models. *Journal of Sound and Vibration*, 333, 6960-6978.

[19] Dougdag, M., Ouali, M., Boucherit, H., Titouche, N.E. and Djaoui, M. (2012) An Experimental Testing of a Simplified Model of a Ball Bearing: Stiffness Calculation and Defect Simulation, *Meccanica*, 47, 335-354. https://doi.org/10.1007/s11012-011-9434-0

[20] Li, Y., Cao, K. and Zhu, Y. (2013) Study on Nonlinear Stiffness of Rolling Ball Bearing under Varied Operating Conditions. *Proceedings of IEEE International Symposium on Assembly and Manufacturing*, Xi’an, 30 July-2 August 2013, 8-11.
**PP and \( P\bar{P} \) Multi-Particles Production Investigation Based on CCNN Black-Box Approach**

El-Sayed A. El-Dahshan\(^1,2\)

\(^1\)Department of Physics, Faculty of Sciences, Ain Shams University, Cairo, Egypt  
\(^2\)Egyptian E-Learning University (EELU), Giza, Egypt  
Email: sedahshan@eelu.edu.eg

**Abstract**

The multiplicity distribution \( \langle P(n_{ch}) \rangle \) of charged particles produced in a high energy collision is a key quantity to understand the mechanism of multi-particle production. This paper describes the novel application of an artificial neural network (ANN) black-box modeling approach based on the cascade correlation (CC) algorithm formulated to calculate and predict multiplicity distribution of proton-proton (antiproton) \( PP \) and \( P\bar{P} \) inelastic interactions full phase space at a wide range of center-mass of energy \( \sqrt{s} \). In addition, the formulated cascade correlation neural network (CCNN) model is used to empirically calculate the average multiplicity distribution \( \langle n_{ch} \rangle \) as a function of \( \sqrt{s} \). The CCNN model was designed based on available experimental data for \( \sqrt{s} = 30.4 \text{ GeV}, 44.5 \text{ GeV}, 52.6 \text{ GeV}, 62.2 \text{ GeV}, 200 \text{ GeV}, 300 \text{ GeV}, 540 \text{ GeV}, 900 \text{ GeV}, 1000 \text{ GeV}, 1800 \text{ GeV}, \text{and} 7 \text{ TeV} \). Our obtained empirical results for \( P(n_{ch}) \), as well as \( \langle n_{ch} \rangle \) for \( PP \) and \( P\bar{P} \) collisions are compared with the corresponding theoretical ones which obtained from other models. This comparison shows a good agreement with the available experimental data (up to 7 TeV) and other theoretical ones. At full large hadron collider (LHC) energy \( \sqrt{s} = 14 \text{ TeV} \) we have predicted \( P(n_{ch}) \) and \( \langle n_{ch} \rangle \) which also, show a good agreement with different theoretical models.

**Keywords**

Proton-Proton and Proton-Antiproton Collisions, Multiparticle Production, Multiplicity Distributions, Intelligent Computational Techniques, CCNN-Neural Networks, Black-Box Modeling Approach

**1. Introduction**

Multiparticle production is an essential entity in high-energy proton-proton col-
E.-S. A. El-Dahshan

Collisions. The hadron-hadron (hh) observables like charged particle multiplicity “n_{ch}” and pseudorapidity density “d\n_{ch}/d\eta” are essential key to characterize the properties of matter created in proton-proton (pp) collisions [1] [2], where, \eta (the pseudorapidity) = -\ln\left[\tan\frac{\theta}{2}\right] and \theta are the polar angle with the beam axis. The dependence of these observables on collision energy (center-of-mass energy “\sqrt{s}”) and the collision geometry are a key tool to understand the underlying particle production mechanism [3] [4] [5]. The investigation of these observables has been used to improve, or reject, models of particle production which are often available as Monte Carlo event generators [6] [7] [8] [9]. The charged particles multiplicity is the simplest observables to understanding of multi-particle production in collisions of hadrons at high-energy [6] [7] [8] [9].

The charged particle multiplicity distributions \(P(n_{ch})\) provide an indispensable tool in the investigation of the dynamics of multi-particle production processes. Their measurements form an important part of the “hh” collision experimental activity. Some new experimental information on the multi-particle production has been reported in the recent past [6]. Consequently, a lot of efforts have also been put forward to analyze and/or organize the experimental data by using various theoretical as well as phenomenological schemes [5]-[12].

There are several models (empirical models and deterministic models) attempt to describe the multiplicity distributions [5] [6] [7]. The first step towards a successful understanding of the multiplicity distributions was done by Feynman in 1969 [3]. Moving to higher energies, deviations from those first models were observed, and the \(P(n_{ch})\) data were described using single Negative Binomial Distributions (NBD) [4], which successfully describe \(P(n_{ch})\) in full phase up to \(\sqrt{s} = 540\text{ GeV}\) as in different \(\eta\)-intervals [4]. But there is a deviation of \(P(n_{ch})\) from NBD for large \(\eta\)-intervals for \(\sqrt{s} = 900\text{ GeV}\), but Giovannini and Ugoccioni [1] [2] describe the measured \(P(n_{ch})\) for \(\sqrt{s} = 900\text{ GeV}\) by the combination of the two weighted NBD [4]. However, this issue is still an open question of interest from the point of view of theoretical and experimental physicists.

To alleviate this problem, we have developed a black-box modeling methodology based on applying the artificial neural network (ANN) approach [13]. ANN Black-box models are powerful and promising tools for complex system modeling. Utilization of an ANN model is, in general, highly suitable for simulating the nonlinear behavior of charged particles multiplicity distributions in proton-proton interactions. This is due to the formulations of neural network models being based on nonlinear functions and having a flexible mathematical structure. In recent years, there has been an increasing amount of applications of ANN models in the field of high energy physics (HEP) [14]-[19].

The most commonly used Neural Network model is the Back-propagation (BP) Network [13], which is a multi-layer feed-forward network trained according to error back propagation algorithm. BP network can be used to learn and discover a mathematical equation that mapping the relation of input-output
model [13]. The disadvantage with multi-layer feed-forward networks using error back propagation is that the best number of hidden layers and units varies from task to task and so must be determined manually through trials and errors. One approach to automatically determine a good size for a network is to start with a minimal network and then add hidden units and connections as required like the Cascade Correlation Neural Networks (CCNN) [20]-[25]. CCNN have several advantages over the ANN, such as they are self organized (i.e. built automatically), less computation cost and complexity (can be obtained with little adjusting parameters) and the training is very fast.

The objective of this paper is to develop a mathematical model based on CCNN approach to calculate and predict the charged particle multiplicity distributions \( P(n_a) \) and the energy dependence of the average multiplicity for \( PP \) and \( P\bar{P} \) inelastic scattering. The CCNN approach learns based on experimental data for full phase space collected from several collaborations [26]-[34], to discover \( P(n_a) \) as a nonlinear response function represented by the network parameters. The \( P(n_a) \) is calculated and predicted by the discovered nonlinear function that representing the CCNN-model, as well as, the energy dependence of the average multiplicity \( \langle n_a \rangle \) for a wide range of energies is calculated and predicted. The obtained results are compared with the ones from different theoretical models such as Dynamical Gluon Mass (DGM) model [35] [36].

The paper is organized as follows: Details of the CCNN black-box model for \( PP \) and \( P\bar{P} \) multiplicity distribution are described In Section 2. The results obtained are presented in Section 3. Finally, the main conclusions of this study are formulated in Section 4.

2. CCNN Black-Box Model for \( P(n_{ch}) \) and \( \langle n_{ch} \rangle \)

Developing a mathematical model that can accurately describe the physical behavior of the complex physical problem is a challenging task. Meanwhile, neural networks are a very promising tool for empirical “black-box” modeling of complex systems without going into mathematical details. An artificial neural network (ANN) is a non linear empirical model that inspired on the biological neural networks [13]. ANN Black-box models do not need detailed prior knowledge of the structure and different interactions that exist between important variables of the nonlinear system that under investigation. Therefore, ANN is a powerful and promising tool for complex system modeling. ANN can be trained with the Cascade-Correlation (CC) learning method to “learn” complex dynamic behaviors of physical systems. A CCNN acts as a black box and learns to predict the value of specific output variables given sufficient input information. The cascade correlation neural network is capable of global function approximation, i.e. it represents a function in a whole data set [20]-[25].

In this paper, we explore the use of CCNN for developing mathematical black-box modeling from experimental data \( PP \) and \( P\bar{P} \) collisions. In the following subsection, we will give a brief introduction to the CCNN approach.
2.1. Overview on CCNN Approach

Artificial neural networks (ANNs) are classified as intelligent computing systems because of their ability to learn. All Artificial neural networks were inspired by the human brain. ANNs consist of artificial neurons connected with each other, and they are termed as nodes. Each neuron has group of inputs, outputs and a transfer function. The mathematical model of a neuron can be described by the equation

\[ y = f \left( \sum_{k=1}^{n} w_k \cdot x_k \right) \]

where \( y \) is the output value, \( x_k \) is the \( k \)th input, \( w_k \) is the weight of the connection related to the \( k \)th input and \( f \) is the transfer function which is usually the radial basis function or the sigmoid function [24] [25].

It's known that a feed forward neural network (FFANN) with one hidden layer is an universal function approximator, so it can approximate any nonlinear function with arbitrary precision. Furthermore, any FFANN can be trained (in the supervised way) by the BP algorithm. The BP algorithm calculates the gradient of the network according to the synaptic weights [13].

The main problem in ANN is the designing of the network with the appropriate number of hidden layers and their units to learn a given concept. If a network has too few hidden units, it will not have the computational power to learn the concept well. Given too many hidden units it will over-fit the training dataset and generalize poorly to new examples that not included in the training data. The CC approach which constructs neural network from bottom to top was proposed by “Fahlman and Lebiere, 1990” [24] in order to solve the problem of low convergence speed of traditional BP, the local minima problem, the step-size problem, the moving target program on and to avoid having to define the number of hidden nodes in advance.

The cascade-correlation architecture supports a variety of learning algorithms, One of the most robust back-propagation variant, called “Quick prop”, was published by Fahlman (1998) [25].

At first the learning algorithm begins with a minimal network (input/output units without hidden unit). The output layer weight was adjusted by the gradient descent algorithm. The error of the network was measure, if the network’s performance was not satisfactory, generate and train a candidate unit.

This candidate neuron is trained by maximizing the magnitude of the correlation between the candidate’s output and the error term to be minimized. Gradient descent is used to minimize the network’s output error, while a gradient ascent is employed to maximize the correlation.

By maximizing the correlation C between the candidate’s output and the network output. Once a neuron is finally added to the network (activated), its input connections become frozen and do not change anymore. Train the network (input/output/hidden unites) until the residual error of the network is minimized (minimize the overall error of the net). This process of optimizing the output weights, creating a hidden neuron, optimizing the hidden neuron weights, connecting it to the output neurons, and adjusting the output neuron weights is
repeated until an acceptably small error is produced or a maximum number of nodes are reached. The following lines summarize the main steps of the CCNN algorithm.

The Cascade Correlation algorithm cycles through two phases: an output phase in which weights entering units are trained in order to reduce network error, and an input phase in which weights entering candidate recruits are trained in order to correlate with network error [23] [24] [25]. The connection weights should be adjusted in the two phases to maximize the correlation and minimize the network error:

**In the first phase:**
- Initialize the CCNN network (2 layers)
- Calculate the actual output \[ y = f \left( \sum_{k} w_k \cdot x_k \right) \]
- The output weights are adjusted until no further progress is made using quick propagation (QuickProp).

\[ E = \sum_{o,p} (y_{op} - t_{op})^2 \]
- Minimize the error (-ve gradient descent of the gradient \( \frac{\partial E}{\partial W} \)) where \( y_{op} \) is the observed value of the output for training pattern output and \( t_{op} \) is the desired output value.

**In the second phase:**
- Add candidate.
- Initialize (weights and learning constant).
- Calculate its output.
- Train candidate to maximize C (by gradient method QuickProp) by “+ve gradient ascent”
- Calculate the correlation between the candidate unite and the residual error of the network.

\[ C_{ij} = \sum_p \left( \sum (O_p - \bar{O})(E_{op} - \bar{E}_o) \right) \]
- Where \( \bar{O} \) and \( \bar{E}_o \) are the average value of candidate hidden units output \( O_p \) and the original network’s output units residual error \( E_{op} \) overall the training samples.

When \( C \) reach Max \( \frac{\partial C}{\partial W} \) it weights freeze
- Add to the main net

We use the QuickProb algorithm to compute and update the network “w” the iteration \( t \)

\[ \Delta w_{n,ij}(t) = \frac{\nabla S_{n,ij}(t)}{\nabla S_{n,ij}(t-1) - S_{n,ij}(t)} \Delta S_{n,ij}(t-1) \quad (2) \]

\[ \nabla S = \frac{\partial C}{\partial W} \text{ or } \frac{\partial E}{\partial W} \quad (3) \]

where, \( S \) is the derivative of the function being optimized (\( E \) in the case of the output phase should be minimized, \( C \) for the input phase should be maximized)

Weight change computed by:
The first phase is started again to train the main net output. These two phases are repeated until either the training pattern has been learned to a predefined level of acceptance or a preset maximum number of hidden units have been added, whichever occurs first. For more details see Ref. [22] [23] [24] [25]. The following subsections discuss the development of CCNN model based on the collected experimental data (which collected from many hadron collider experiments [26]-[34]).

2.2. $P(n_{ch})$ and $\langle n_{ch} \rangle$ Model Development

The objective of this paper is to develop a mathematical model based on CCNN approach to calculate and predict for $PP$ and $P\bar{P}$ scattering. The mathematical model is based on numerous experiments conducted on different Labs [26]-[34], and a neural network approximate method which is employed to predict or extrapolate the experimental results.

To train and test the proposed model, the CCNN program code was developed by using MATLAB language [The Math works Inc. USA]. CCNN has the disadvantage of over-fitting the training data. Due to this, the accuracy values are quite high in case of training data, but low in testing data. So, for preventing the over-fitting of the training data the CCNN model is validated as it grows using the 3-fold cross validation.

To compute the performance of CCNN model, we have examined the performance indices ($R^2$ and RMSE) until no significant improvement occurred. Once the training is complete, the CCNN model would have learned to approximate $P(n_{ch})$ and $\langle n \rangle$ i.e. reproduce, interpolate and extrapolate the data that are not included in the training data.

3. Results and Discussion

In this section, we have applied the CCNN model to calculate and predict $P(n_{ch})$ and $\langle n_{ch} \rangle$ using the available experimental data [26]-[34].

In the present CCNN, we have obtained $R^2 = 0.998$ and RMSE = 0.000137. The following network training parameters are used: Minimum neurons in hidden layer: 2; Maximum neurons in hidden layer: 200; Hidden neuron kernel function: Gaussian, Output neuron kernel function: Linear and Over-fitting protection control = 3-fold cross-validation.

In this regard, we have modeled the $P(n_{ch})$ at a wide range of available experimental data for center-of-mass energy: $\sqrt{s} = 30.4$ GeV, 44.5 GeV, 52.6 GeV, 62.2 GeV, 200 GeV, 300 GeV, 540 GeV, 900 GeV, 1000 GeV, 1800 GeV, and 7 TeV (From ISR energies in the 1970’s to the highest LHC for $PP$ and $P\bar{P}$ scattering). We have compared the obtained results with the recently published experimental, empirical and/or phenomenological results. We also, provide predictions of the $P(n_{ch})$ in $pp$ collisions at the full LHC energies (14 TeV).

Figures 1(a)-1(c) shows our calculated and predicted results of $P(n_{ch})$ as a
E. - S. A. El-Dahshan

\[ \sqrt{s} = 30.4 \text{ GeV} \] (pp)

\[ \sqrt{s} = 44.5 \text{ GeV} \] (pp)

\[ \sqrt{s} = 52.6 \text{ GeV} \] (pp)

\[ \sqrt{s} = 62.2 \text{ GeV} \] (pp)

\[ \sqrt{s} = 200 \text{ GeV} \] (p-anti(p))

\[ \sqrt{s} = 300 \text{ GeV} \] (p-anti(p))

\[ \sqrt{s} = 546 \text{ GeV} \] (p-anti(p))

\[ \sqrt{s} = 900 \text{ GeV} \] (p-anti(p))
Figure 1. (a)-(c). Multiplicity distribution $P(n_{ch})$ of charged particles in $pp$ and $p\bar{p}$ inelastic collisions at $30.4 \text{ GeV} \leq \sqrt{s} \leq 7 \text{ TeV}$. Figure 1(c) (Bottom-Right panel) prediction of $P(n_{ch})$ for $pp$ at $\sqrt{s} = 14 \text{ TeV}$. —DGM model, * CCNN model (our model) and o experimental data of the five Collaborations [26]-[34].

Function of $n_{ch}$ and $\sqrt{s}$. Also, this figure shows the comparison between our calculated and predicted $P(n_{ch})$ values and the other theoretical and experimental values [25]-[34]. In this comparison, our model results show closer agreement with the experimental data and the theoretical ones. Figures 1(a)-1(c) demonstrates that, the predicted $P(n_{ch})$ spectra values are very close to the actual values (experimental data) which indicates that CCNN can be used as an effective tool for modeling the $P(n_{ch})$ based on the $n_{ch}$ and $\sqrt{s}$.

Figure 1(c) (Bottom-Right panel) manifests the prediction of multiplicity distribution of the produced particles at LHC energy ($\sqrt{s} = 14 \text{ TeV}$) which is compared with those distributions obtained by other models [26]-[34]. According to our CCNN model, the prediction of $P(n_{ch})$ at 14 TeV having the same trend as the theoretical one [35] [36]. In addition, the $\langle n_{ch} \rangle$ energy dependence was modeled (calculated and predicted at a wide range of $\sqrt{s}$) (from 30 GeV to 7 TeV) and as well as predicted at the highest LHC energy (14 TeV).

The probability of production of particles decreases with the increase of $\sqrt{s}$ as well as shifted towards the increase of $n_{ch}$. Also, we notice that the width of the distribution is broadened with the increase of $\sqrt{s}$ as shown in Figure 2. This figure shows the multiplicity distributions $P_{n_{ch}}(n_{ch})$ of charged particles in $pp$ and $p\bar{p}$ collisions “full phase space” from 30 GeV to 14 TeV.
Fig. 2. Multiplicity distributions \( P_{ch}(n_{ch}) \) of charged particles in \( pp \) and \( p\bar{p} \) collisions from 30 GeV to 14 TeV “full phase space”.

Based on the proposed CCNN model, the values of energy dependence of the average charged multiplicity in \( pp \) collision are calculated \( \left< n_{ch} \right> = \sum_{n_{ch}} n_{ch} P(n_{ch}) \) and compared with corresponding experimental and theoretical results. **Figure 3** shows the energy dependence of the average charged multiplicity in \( pp \) collision for \( \sqrt{s} \) ranging from 30 Gev to 14 TeV. The calculated values are compared with the corresponding experimental and theoretical results. Also, from **Figure 3** we notice that \( \left< n_{ch} \right> \) increases with the increase of \( \sqrt{s} \) which shows the same trend as the experiment [26]-[34]. The results of the present open the route into applying modern soft-computing procedures such as neural network into the modeling of HEP.

### 4. Conclusions

The charged-particle multiplicity belongs to the simplest observable that provides important insights into the mechanisms of particle production. In the present work we have used CCNN network for modeling the multiplicity distribution of charged particles produced in \( pp \) and \( p\bar{p} \) interactions at Center-of-mass energy \( \sqrt{s} = 30.4 \text{ GeV}, 44.5 \text{ GeV}, 52.6 \text{ GeV}, 62.2 \text{ GeV}, 200 \text{ GeV}, 300 \text{ GeV}, 540 \text{ GeV}, 900 \text{ GeV}, 1000 \text{ GeV}, 1800 \text{ GeV}, \) and 7 TeV. In this regard, we have developed the CCNN mathematical black-box models to calculate and predict the multiplicity distribution of charged particles \( P(n_{ch}) \) produced in proton-proton collisions as a function of \( n_{ch} \) and \( \sqrt{s} \), as well as the energy dependence the energy dependence of average multiplicity \( \left< n_{ch} \right> \). The results indicate that the proposed CCNN model shows a good correspondence between
Figure 3. Shows the comparison between our calculated and predicted values for the energy dependence of the average charged multiplicity and the corresponding experimental and theoretical data [25]-[34].

the experimental data and our calculated results according to the statistical performance. We have also compared our results for $P(n_{ch})$ and $\langle n_{ch} \rangle$ with the models that based on Monte Carlo model, which successfully explains multiplicity distribution. In addition, the predictions for $P(n_{ch})$ and $\langle n_{ch} \rangle$ of charged particles in pp interactions at $\sqrt{s} = 14$ TeV are found to be in agreement with Dynamical Gluon Mass model [35] [36]. The obtained results confirm the reliability of our model and will encourage physicists to apply other ANN techniques to calculate and predict other problems in multiparticle production investigation.

Acknowledgements

The author would like to thank Prof. Dr. M Y. El-Bakry, Faculty of Education, and Prof. Dr. S.Y. El-Bakry, Faculty of Science, Ain Shams University for the useful discussion on the problems considered in this article.

References

[1] Giovannini, A. and Ugoccioni, R. (1999) Soft and Semi-Hard Components Structure in Multiparticle Production in High Energy Collisions. Nuclear Physics B—Proceedings Supplements, 71, 201-210.

[2] Giovannini, A. and Ugoccioni, R. (2003) Signals of New Physics in Global Event Properties in pp Collisions in the TeV Energy Domain. Physical Review D, 68, Article ID: 034009, https://doi.org/10.1103/PhysRevD.68.034009

[3] Beggio, P.C. (2017) Inelastic Cross Sections, Overlap Functions and Moments from ISR to LHC Energies in Proton Interactions. Journal of Physics G. Nuclear and Particle Physics, 44, Article ID: 025002. https://doi.org/10.1088/1361-6471/aa51f5
[4] Feynman, R.P. (1969) Very High-Energy Collisions of Hadrons. *Physical Review Letters*, 23, 1415. [https://doi.org/10.1103/physrevlett.23.1415](https://doi.org/10.1103/physrevlett.23.1415)

[5] ALICE Collaboration (2015) Charged-Particle Multiplicities in Proton-Proton Collisions at $\sqrt{s} = 0.9$ to 8 TeV. *The European Physical Journal C*, 77, 33.

[6] Mukherjee, M. (2016) Multiplicity Distributions and Fluctuations in Proton-Proton and Heavy-Ion Collisions. In: *EPJ Web of Conferences*, Vol. 112, EDP Sciences, Article No. 04004. [https://doi.org/10.1051/epjconf/201611204004](https://doi.org/10.1051/epjconf/201611204004)

[7] Ghosh, P. and Muhuri, S. (2013) Multiplicity Distributions in the Forward Rapidity Region in Proton-Proton Collisions at the Large Hadron Collider. *Physical Review D*, 87, Article ID: 094020. [https://doi.org/10.1103/physrevd.87.094020](https://doi.org/10.1103/physrevd.87.094020)

[8] Kumar, A., Srivastava, P.K., Singh, B.K. and Singh, C.P. (2013) Charged Hadron Multiplicity Distribution at Relativistic Heavy-Ion Colliders. *Advances in High Energy Physics*, 2013, Article ID: 352180. [https://doi.org/10.1155/2013/352180](https://doi.org/10.1155/2013/352180)

[9] Grosse-Oetringhaus, J.F. and Reygers, K. (2010) Charged-Particle Multiplicity in Proton-Proton Collisions. *Journal of Physics G: Nuclear and Particle Physics*, 37, Article ID: 083001. [https://doi.org/10.1088/0954-3899/37/8/083001](https://doi.org/10.1088/0954-3899/37/8/083001)

[10] Toia, A. (2016) First Run II Results from ALICE. In: *EPJ Web of Conferences*, Vol. 120, EDP Sciences, Article No. 09003. [https://doi.org/10.1051/epjconf/201612009003](https://doi.org/10.1051/epjconf/201612009003)

[11] Kamoldt, K., Abel, N., Abeysekara, U., Quintana, A.A., Abramyan, A., Adamova, D. and Ahammed, Z. (2010) Charged-Particle Multiplicity Measurement in Proton-Proton Collisions at $\sqrt{s} = 0.9$ and 2.36 TeV with ALICE at LHC. *The European Physical Journal C*, 68, 89-108. [https://doi.org/10.1140/epjc/s10052-010-1339-x](https://doi.org/10.1140/epjc/s10052-010-1339-x)

[12] Monteno, M. and ALICE Collaboration (2005) Particle Multiplicity in Proton-Proton Collisions with ALICE. *Journal of Physics: Conference Series*, 5, 209.

[13] Haykin, S. (2008) Neural Networks and Learning Machines. 3rd Edition, Prentice Hall, Upper Saddle River.

[14] Schmidt, M. and Lipson, H. (2009) Distilling Free-Form Natural Laws from Experimental Data. *Science*, 324, 81-85.

[15] Link, J.M., Yager, P.M., Anjos, J.C., Bediaga, I., Castromonte, C., Göbel, C. and Pepe, I.M. (2005) Application of Genetic Programming to High Energy Physics Event Selection. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 551, 504-527.

[16] Teodorescu, L. and Sherwood, D. (2008) High Energy Physics Event Selection with Gene Expression Programming. *Computer Physics Communications*, 178, 409-419.

[17] El-Dahshan, E.S. (2011) Application of Genetic Programming for Proton-Proton Interactions. *Open Physics*, 9, 874-883. [https://doi.org/10.2478/s11534-010-0088-7](https://doi.org/10.2478/s11534-010-0088-7)

[18] Denby, B. (1988) Neural Networks and Cellular Automata in Experimental High Energy Physics. *Computer Physics Communications*, 49, 429-448.

[19] David, C., Freslter, M. and Aichelin, J. (1995) Impact Parameter Determination for Heavy-Ion Collisions by Use of a Neural Network. *Physical Review C*, 51, 1453. [https://doi.org/10.1103/physrevc.51.1453](https://doi.org/10.1103/physrevc.51.1453)

[20] McCulloch, W.S. and Pitts, W. (1943) A Logical Calculus of the Ideas Immanent in Nervous Activity. *The Bulletin of Mathematical Biophysics*, 5, 115-133. [https://doi.org/10.1007/BF02478259](https://doi.org/10.1007/BF02478259)

[21] Rosenblatt, F. (1958) The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychological Review*, 65, 386. [https://doi.org/10.1037/h0042519](https://doi.org/10.1037/h0042519)
[22] Hoefeld, M. and Fahlman, S.E. (1991) Technical Report CMUCS-91-130. School of Computer Science, Carnegie Mellon University, Pittsburgh.

[23] Fahlman, S.E. (1991) The Recurrent Cascade-Correlation Architecture. Carnegie-Mellon University, Department of Computer Science, Pittsburgh.

[24] Fahlman, S.E. and Liebiere, C. (1990) The Cascade—Correlation Learning Architecture. In: Touretzky, D.S., Ed., Advances in Neural Information Processing Systems, Vol. 2, Morgan Kaufmann, San Mateo, 524-532.

[25] Fahlman, S.E. (1988) An Empirical Study of Learning Speed in Back-Propagation Networks. Technical Report CMU-CS-88-162, Carnegie-Mellon University, Pittsburgh.

[26] Ansorge, R.E., et al. (1989) Charged Particle Multiplicity Distributions at 200-GeV and 900-GeV Center-Of-Mass Energy. Zeitschrift für Physik C, 43, 357.

[27] Breakstone, A., Campanini, R., Crawley, H.B., Dallavalle, G.M., Deninno, M.M., Doroba, K. and Fischer, H.G. (1984) Charged Multiplicity Distribution in pp Interactions at CERN ISR Energies. Physical Review D, 30, 528. https://doi.org/10.1103/PhysRevD.30.528

[28] Alexopoulos, T., Anderson, E.W., Biswas, N.N., Bujak, A., Carmony, D.D., Erwin, A.R. and Lindsey, C.S. (1998) The Role of Double Parton Collisions in Soft Hadron Interactions. Physics Letters B, 435, 453-457.

[29] Alner, G.J., Alpga, K., Anderer, P., Ansorge, R.E., Berglund, S., Berkelman, K., Carlson, P., et al. (1987) UA5: A General Study of Proton-Antiproton Physics at √s = 546 GeV. Physics Reports, 154, 247-383.

[30] Alexopoulos, T., Allen, C., Anderson, E.W., Balamurali, V., Banerjee, S., Beery, P.D. and Carmony, D.D. (1993) Mass-Identified Particle Production in Proton-Antiproton Collisions at √s = 300,540,1000 and 1800 GeV. Physical Review D, 48, 984. https://doi.org/10.1103/PhysRevD.48.984

[31] Chatrchyan, S., Khachatryan, V., Sirunyan, A.M., Tumasyan, A., Adam, W., Bergauer, T. and Fruehwirth, R. (2011) Measurement of the Inclusive W and Z Production Cross Sections in pp Collisions at √s = 7 TeV with the CMS Experiment. Journal of High Energy Physics, 2011, 132. https://doi.org/10.1007/JHEP10(2011)132

[32] Mitsou, V., FerrerSoria, A., VallsFerrer, J.A., Castillo Giménez, M.V., Cabrera Urbán, S., Fuster Verdu, J.A. and Higón Rodríguez, E. (2011) Charged-Particle Multiplicities in pp Interactions Measured with the ATLAS Detector at the LHC. New Journal of Physics, 13, Article ID: 053033.

[33] Aad, G., Abat, E., Abbott, B., Abdallah, J., Abdelalim, A.A., Abdesselam, A. and Abreu, H. (2010) Charged-Particle Multiplicities in pp Interactions at Measured with the ATLAS Detector at the LHC. Physics Letters B, 688, 21-42.

[34] Kamoit, K., Abel, N., Abeysekara, U., Quintana, A.A., Acero, A., Adamova, D. and Ahammed, Z. (2010) First Proton-Proton Collisions at the LHC as Observed with the ALICE Detector: Measurement of the Charged-Particle Pseudorapidity Density at √s = 900 GeV. The European Physical Journal C, 65, 111-125. https://doi.org/10.1140/epjc/s10052-009-1227-4

[35] Beggio, P.C. and Luna, E.G.S. (2014) Cross Sections, Multiplicity and Moment Distributions at the LHC. Nuclear Physics A, 929, 230-245.

[36] Beggio, P.C. (2013) Oscillations of Factorial Cumulates to Factorial Moments Ratio from an Eikonal Approach. Nuclear Physics A, 913, 264-275.
Journal of Applied Mathematics and Physics is an international journal dedicated to the latest advancement of applied mathematics and physics. The goal of this journal is to provide a platform for researchers and scientists all over the world to promote, share, and discuss various new issues and developments in different areas of applied mathematics and physics. We aim to publish high quality research articles in terms of originality, depth and relevance of content, and particularly welcome contributions of interdisciplinary research on applied mathematics, physics and engineering.

Subject Coverage
The journal publishes original papers including but not limited to the following fields:

- Applications of Systems
- Applied Mathematics
- Applied Non-Linear Physics
- Applied Optics
- Applied Solid State Physics
- Biophysics
- Computational Physics
- Condensed Matter Physics
- Control Theory
- Cryptography
- Differentiable Dynamical Systems
- Engineering and Industrial Physics
- Experimental Mathematics
- Fluid Mechanics
- Fuzzy Optimization
- Geophysics
- Integrable Systems
- Laser Physics
- Methodological Advances
- Multi-Objective Optimization
- Nanoscale Physics
- Nonlinear Partial Differential Equations
- Non-Linear Physics
- Nuclear Physics
- Numerical Computation
- Optical Physics
- Portfolio Selection
- Riemannian Manifolds
- Scientific Computing
- Set-Valued Analysis
- Soliton Theory
- Space Physics
- Symbolic Computation
- Topological Dynamic Systems
- Variational Inequality
- Vector Optimization

We are also interested in: 1) Short reports—2-5 page papers where an author can either present an idea with theoretical background but has not yet completed the research needed for a complete paper or preliminary data; 2) Book reviews—Comments and critiques.

Notes for Intending Authors
The journal publishes the highest quality original full articles, communications, notes, reviews, special issues and books, covering both the experimental and theoretical aspects including but not limited to the above materials, techniques and studies. Papers are acceptable provided they report important findings, novel insights or useful techniques within the scope of the journal. All manuscript must be prepared in English, and are subjected to a rigorous and fair peer-review process. Accepted papers will immediately appear online followed by prints in hard copy.

Website and E-Mail
http://www.scirp.org/journal/jamp   E-mail: jmp@scirp.org
What is SCIRP?
Scientific Research Publishing (SCIRP) is one of the largest Open Access journal publishers. It is currently publishing more than 200 open access, online, peer-reviewed journals covering a wide range of academic disciplines. SCIRP serves the worldwide academic communities and contributes to the progress and application of science with its publication.

What is Open Access?
All original research papers published by SCIRP are made freely and permanently accessible online immediately upon publication. To be able to provide open access journals, SCIRP defrays operation costs from authors and subscription charges only for its printed version. Open access publishing allows an immediate, worldwide, barrier-free, open access to the full text of research papers, which is in the best interests of the scientific community.

- High visibility for maximum global exposure with open access publishing model
- Rigorous peer review of research papers
- Prompt faster publication with less cost
- Guaranteed targeted, multidisciplinary audience

Website: http://www.scirp.org
Subscription: sub@scirp.org
Advertisement: service@scirp.org