JOURNAL EDITORIAL BOARD

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

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

Editorial Board
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. Wanyang Dai
Nanjing University, China

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

Prof. D.S. Hooda
GJ University of Science & Tech, Hisar, India

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 Lv
Beijing Jiaotong University, China

Dr. Jafar Fawzi Mansi AlOmari
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
Atilim 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. Yaning Tang
Northwestern Polytechnical University, Xi’an, P.R. China, China

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 8   Number 6  June 2020**

<table>
<thead>
<tr>
<th>Title</th>
<th>Authors</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parameter Sensitivity and Qualitative Analysis of Dynamics of Ovarian Tumor Growth Model with Treatment Strategy</td>
<td>M. S. Alam, M. Kamrujjaman, M. S. Islam</td>
<td>941</td>
</tr>
<tr>
<td>Some Chaotic Properties of a Kind of Coupled Map Lattices</td>
<td>X. F. Yang, T. X. Lu, G. Liu</td>
<td>968</td>
</tr>
<tr>
<td>Criterion for the Emergence of Meta-Stable States in Traffic Systems</td>
<td>L. H. Zhu</td>
<td>976</td>
</tr>
<tr>
<td>An Efficient Projected Gradient Method for Convex Constrained Monotone Equations with Applications in Compressive Sensing</td>
<td>Y. P. Hu, Y. J. Wang</td>
<td>983</td>
</tr>
<tr>
<td>The Confirmation of Hypothesis of the Absolute Reference System</td>
<td>K. Patrinos</td>
<td>999</td>
</tr>
<tr>
<td>A Complex Algorithm for Solving a Kind of Stochastic Programming</td>
<td>Y. P. Luo, X. S. Ma</td>
<td>1016</td>
</tr>
<tr>
<td>Comparison of Ising Model and Potts Model on Non-Local Directed Small-World Networks</td>
<td>M. A. Sumour, M. Kh. SrouR, S. M. Baraka, M. A. Radwan, R. J. Khozondar, M. M. Shabat</td>
<td>1031</td>
</tr>
<tr>
<td>On Two Double Inequalities (Optimal Bounds and Sharps Bounds) for Centroidal Mean in Terms of Contraharmonic and Arithmetic Means</td>
<td>M. El Mokhtar Ould El Mokhtar, H. Alharbi</td>
<td>1039</td>
</tr>
<tr>
<td>Ballistic Principle of the Property Balance in Space and Its Application to Modeling of Fluid Dynamics Problems</td>
<td>N. Kislov</td>
<td>1081</td>
</tr>
</tbody>
</table>
Collocation Method for Solving the Generalized KdV Equation
T. Geyikli

Precursive Time, the Hidden Variable
R. Burri

Steady State Gas Flow in Pipeline Networks: Existence and Uniqueness of Solution
A. Atena, W. Tekalign, T. Muche

On the Uphill Domination Polynomial of Graphs
T. Alsalomy, A. Saleh, N. Muthana, W. Al Shammakh

Application of Stability Theory in Study of Local Dynamics of Nonlinear Systems
T. Azizi, G. Kerr
Parameter Sensitivity and Qualitative Analysis of Dynamics of Ovarian Tumor Growth Model with Treatment Strategy

Md. Shah Alam¹, Md. Kamrujjaman²,³, Md. Shafiqul Islam⁴

¹Department of Mathematics, University of Louisiana at Lafayette, Lafayette, LA, USA
²Department of Mathematics, University of Dhaka, Dhaka, Bangladesh
³Department of Mathematics and Statistics, University of Calgary, Calgary, AB, Canada
⁴School of Mathematical and Computational Sciences, University of Prince Edward Island, PE, Canada

Email: mdshah.alam1@louisiana.edu, kamrujjaman@du.ac.bd, sislam@upei.ca


Abstract
In this paper, we are interested to find the most sensitive parameter, local and global stability of ovarian tumor growth model. For sensitivity analysis, we use Latin Hypercube Sampling (LHS) method to generate sample points and Partial Rank Correlation Coefficient (PRCC) method, uses those sample points to find out which parameters are important for the model. Based on our findings, we suggest some treatment strategies. We investigate the sensitivity of the parameters for tumor volume, y, cell nutrient density, Q and maximum tumor size, ymax. We also use Scatter Plot method using LHS samples to show the consistency of the results obtained by using PRCC. Moreover, we discuss the qualitative analysis of ovarian tumor growth model investigating the local and global stability.

Keywords
Parameter Sensitivity, Latin Hypercube Sampling, Partial Rank Correlation Coefficient, Scatter Plot, Monotonicity, Stability Analysis

1. Introduction
Ovarian cancer is the fifth leading cause of death from non-skin cancers among women around the globe. “Silent killer” is another name of ovarian cancer, causes more deaths than any other gynecological malignancies. The American Cancer Society estimates 22,240 new cases of ovarian cancer and 14,070 deaths due to ovarian cancer only in United States in 2018. Almost 300,000 new patients have been diagnosed with ovarian cancer in 2018.
Ovarian cancers were previously believed to begin only in the ovaries, but recent evidence suggests that many ovarian cancers may actually start in the cells in the far (distal) end of the fallopian tubes [1]. Only 20 percent ovarian cancers are detected at early age. Despite the advancement of last few decades, ovarian cancer still remains a major problem [2]. A mathematical expression called Droop’s cell quota model [3] governs the tumor growth, where cell quota represents the intracellular concentration of necessary nutrients provided through blood supply [4] [5].

For every mathematical model, input factors such as parameters are not always known with a sufficient degree of certainty because of natural variation, error in measurements or even simply a lack of current techniques to measure them. Our goal is to analyze the uncertainty of parameters of our model. Because uncertainty in parameter values chosen, introduces variability to the model’s prediction of resulting dynamics. So, the more uncertain parameters there are, the more significant the variability introduced. LHS-PRCC sensitivity analysis is an efficient tool often employed in uncertainty analysis to explore the entire parameter space of a model [6]. Scatter plot is an alternate of PRCC which works in a different way but able to give a meaningful understanding of the parameters behavior for a particular mathematical model [7].

The mathematical methods used in modeling biological systems vary according to different steps of the process. We focus on the mathematical representation of the system. However, other important steps in the modeling processes are parameters fitting and model selection [8]. Mathematical models of complex biological systems are central to systems biology [9] [10]. They can be used as an exploratory tool to complement and guide experimental work. Model simulations can be used to predict the system-wide effects of molecular targets, such as, determine the effects of molecular target(s) inhibition in specific populations [11] [12]. They can also serve as an important clinical tool, for example, classify benign and malignant tumors, predict disease prognosis for individual patients, and predict outcomes of treatments [13] [14] [15] [16]. All these studies employed to fit both on-treatment and off-treatment preclinical data using the same biologically relevant parameters. Using mathematical tools, similar study was considered in a proposed model that consisted of healthy cells, tumor cells, and mature vascular endothelial cells in the tumor [17]. The growth of the cancerous cells can also be limited by the lack of blood vessels, which carry important nutrients and supplies.

Scientists have been using ordinary and partial differential equations to model biological systems for a long time. As these models are utilized as a part of an attempt to better understanding of more and more complex phenomena, it is becoming obvious that the simple models cannot capture the complexity of dynamics observed in natural systems [18]. Different types of approaches can be taken to deal with these complexities. One way of dealing with this issue is constructing large system of ODE, which can be quite good at approximating observed behavior [19] [20]. Such models have the benefit of merging a simple, in-
tuitive derivation with an extensive variety of possible behavior regimes for a single system. Also, these models hide lots of detailed workings of complex biological systems, where sometimes precise details are important for this system.

Ordinary differential equations (ODE) and delay differential equations (DDE) are useful in framing many biological phenomena [20]. Though delay differential equations and ordinary differential equations have many similarities, DDE have several features which make their analysis more difficult. ODE hold derivative which depend on the solution at current value of the independent variable (time), while DDE additionally contain derivatives, which are dependent on solutions at earlier times. One of the most significant difference between the ODE and DDE is the initial data. The solution of an ODE is resolved by its value at the initial point \( t = a \). However, for the interval \( a \leq t \leq b \), the term like \( y(t - \tau) \) requires initial point to solve the system [20].

Basically, ovarian tumor growth model is DDE which has two phases namely on-treatment and off-treatment. Here we are investigating for only on-treatment case and for this reason, the model turns into ODE. We have introduced the Runge-Kutta method of order 5 to solve the system of non-linear differential Equation (1), as prescribed in the next Section 2.

The rest of the paper is organized as follows. The mathematical model is described in Section 2 with parameter estimations. The solution methodology is prescribed in Section 3. Also the Scatter plotting idea is articulated in this section. The results of numerical illustrations are presented in Section 4. LHS performance, Monotonicity plots analysis and PRCC studies are also investigated in this section. The contents of Section 5 are analyzing the treatment strategy to reduce the Ovarian cancer. Theoretical results such as local and global stability analysis are presented in Section 6. Finally, Section 7 concludes summary and discussion of the results.

The dynamics of mathematical model are integrated in the following section.

2. Mathematical Model of Ovarian Cancer Dynamics

Ovarian Tumor Growth Model is a simple vascularized model; a type of tumor that forms from cells that make blood vessels or lymph vessels. Vascular tumors may form on the skin, in the tissues below the skin, and/or in an organ. There are many types of vascular tumors. The most common type of vascular tumor is hemangioma, which is a benign tumor that usually occurs in infants and goes away on its own. In this model, the idea of nutrient limited induced angiogenesis has been used [21]. As already mentioned, the model is a DDE, but for on-treatment case it becomes an ODE where the delay part \( \tau = y(t_0 - \tau) \) has established values ranging 200 - 10,300 [22]. We are investigating the case from the beginning of the on-treatment case for the time delay, \( \tau = 10\) and initial time \( t_0 = 29\) which tends to give approximately \( \tau = y(t_0 - \tau) = y(29 - 10) = y(19) = 375 \) (see Figure 7 in [22]).

Following is the ovarian tumor growth model [22],
\[
\begin{align*}
\frac{dy}{dt} &= \mu_m \left(1 - \frac{q}{Q}\right) y - yd, \\
\frac{dQ}{dt} &= \alpha_i p_i \left(\frac{y(t_0 - \tau)}{y} - \mu_m (Q - q)\right).
\end{align*}
\]

For convenience and parameter estimations, the variables and parameters of system (1) are described in Table 1.

We simulate the model for 100 days of on-treatment case to get both Tumor volume, \(y\) and Cell Quota \(Q\), see Figure 1.

Now it’s time to describe the solution methods.

3. Methodology

3.1. LHS-PRCC

To explore the uncertainty of parameters, one of the most useful sensitivity analysis method is Latin Hypercube Sampling-Partial Rank Correlation Coefficient (LHS-PRCC). It determines the full parameter space of a model with an optimal number of computer simulations [6]. This method uses the combination of two statistical procedures, Latin Hypercube Sampling (LHS), which was first presented by McKay [23] in 1979 and Partial Rank Correlation Coefficient (PRCC) analysis.

![Figure 1](image1.png)

**Figure 1.** Solution curves for Tumor Volume (\(y\)) and Cell Quota (\(Q\)).

Table 1. Variable and parameter list for ovarian tumor growth model.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition(Unit)</th>
<th>Value</th>
<th>Ref</th>
</tr>
</thead>
<tbody>
<tr>
<td>(y)</td>
<td>Tumor volume (vol)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>(Q)</td>
<td>Cell nutrient density (mol/vol)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>(q)</td>
<td>Minimum cell nutrient density (mol/vol)</td>
<td>0.0021 - 0.0099</td>
<td>assume</td>
</tr>
<tr>
<td>(\mu_m)</td>
<td>Maximum growth rate (per day)</td>
<td>0.41 - 1.58</td>
<td>[18]</td>
</tr>
<tr>
<td>(d)</td>
<td>Death rate (per day)</td>
<td>0.28 - 1.43</td>
<td>[18]</td>
</tr>
<tr>
<td>(\alpha_i)</td>
<td>Nutrient uptake coefficient (mol/(vol day))</td>
<td>0.0084 - 0.70</td>
<td>assume</td>
</tr>
<tr>
<td>(p_i)</td>
<td>Reduction in nutrient uptake rate (-)</td>
<td>0.17 - 0.47</td>
<td>assume</td>
</tr>
<tr>
<td>(\tau)</td>
<td>Time delay (day)</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>(\tau)</td>
<td>Tumor size 10 days before treatment (vol)</td>
<td>375</td>
<td>[22]</td>
</tr>
</tbody>
</table>
Within a given range of parameters value, LHS samples them to generate different values at each simulation and PRCC uses those value to describe the relation of parameters with the output of a particular mathematical model [24]. PRCC is a sample based method to examine the correlation between a model output variable and parameters using sample points generated by LHS.

The goal of LHS-PRCC sensitivity analysis is to identify significant parameters which have great impact for model prediction and to rank these parameters depending on their contribution for a precise model prediction [25]. This section presents details on the steps of LHS. The method uses the following procedure:

1) Make a list of the parameters for the model with their consistent values.

2) We have to predict the uncertain parameters from the parameter lists. For some of these, it might not be difficult to find the possible range where the exact values might fall.

3) Next step is to decide the sample size and to do this we need to determine the number of simulations we intend to run. Assume we decide to run \( N \) model simulations for analysis and we have \( K \) uncertain parameters, \( \nu_i, 1 \leq i \leq K \). Then the parameter space for the uncertain parameters would be defined by \( K \) dimensions.

4) \( K \) dimensions will correspond to uncertain parameters and \( N \) determines the length of dimensions. For every uncertain parameter, each of the \( N \) input values would be selected/determined by the LHS sampling scheme.

5) We need to specify a probability density or distribution function (pdf) for each uncertain parameter to implement this LHS sampling scheme. In this way, the variability in the pdf becomes a direct measure of the variability of the uncertain parameter.

6) Each probability density function is divided into \( N \) non-overlapping equiprobable intervals for sampling the values of each parameter.

7) Each equiprobable interval of each parameter is then randomly sampled once. The parameters are uncorrelated because each parameter is sampled independently.

8) Once step 7 is complete, each of the \( K \) uncertain parameters, \( \nu_i, 1 \leq i \leq K \), will have \( N \) values. Hence, we store the sampled values in an \( N \times K \) table/matrix.

### 3.2. Scatter Plot

Scatter plots are occasionally used to examine the correlation between a model output variable and parameters visually [26]. It's a variance based method to find a trend or pattern of the data obtained from LHS sampling. An output variable that is sensitive to the selected parameter will yield an obvious correlated pattern in the scatter plot. Generally, a Monte Carlo algorithm is used to sample the parameter space, and multiple scatter plots are drawn illustrating the relationship between each parameter and each output variable of interest. But in this paper, we use LHS instead of Monte Carlo algorithm for sampling the parameter.
space. Visual recognition of the correlation between parameters and model output values can be contingent on the choice of axis scales.

For the sake of comprehension and clarity, we state and discuss our illustrated key results at this point.

4. Results

4.1. Performing LHS on the Ovarian Tumor Growth Model

There are five (since $\bar{y}$ and $\tau$ have a fixed values) uncertain or Latin Hypercube Sampling (LHS) parameters in Table 1. To identify their various roles in the model predictions, we start performing LHS. In our analysis, 100 model simulations were performed with 1000 samples per run. Thus, the parameter space (LHS matrix) for the LHS parameters has dimension of length five with each dimension specifying an uncertain parameter vector of length 100. We assume the maximum and minimum values for each of the five LHS parameters (see Table 1). The baseline value for each LHS parameter has been set to a value at or near the middle of the range for a particular parameter. For every LHS parameter, each of the 100 input values are obtained by sampling a uniform probability density distribution. The 100 input values are then used to populate the LHS matrix from which we produced monotonic plots for each variable. We calculate tumor volume, $y$, cell quota or cell nutrient density, $Q$ and maximum tumor size, $y_{max}$ for each run after 100 days.

4.2. Analyzing the Monotonicity Plots

Analysis of Monotonicity plot is a precondition to apply PRCC on LHS generated samples. Three outcome measures mentioned in the last section are presented in the following figures. The subplots in Figure 2 ($x$-axis represents the parameter values and $y$-axis represents the respective outcome measure) have monotonic relation with all the outcome measures except reduction in nutrient uptake rate ($p_i$) with maximum tumor size, $y_{max}$.

To solve this issue one approach is to breakdown that graph into two monotonic regions. If instead of the small range of outcome measures observed for $p_i$ in $y_{max}$, the range had been several hundred or thousand units, we would have considered truncating the range and looking at each truncated half separately. However, the current effect of $p_i$ in $y_{max}$ is minimal since the range is very small (i.e. $1000 ~ 1010$) for number of orders of $10^3$. So, no action is needed. Hence, all LHS parameters of our model have a monotonic relationship with the outcome measures tumor Volume, $y$, cell quota, $Q$ and maximum tumor size, $y_{max}$.

4.3. Analyzing the PRCC

In PRCC analysis, we consider the parameters with PRCC values $>0.5$ (for direct relation) or $<-0.5$ (for inverse relation) and corresponding small P-values ($<0.05$) as the most influential parameters for the model.
In each PRCC plot (Figures 3-5), x-axis contains the parameters and the y-axis contains PRCC values. Also, for each PRCC plot there is a corresponding P-value plot where x-axis represents the parameters and the y-axis represents the corresponding P-values.

We observe that each PRCC and P-value plot show strong correlation of death rate, \( d \) and maximum growth rate, \( \mu_m \) with all three outputs.

4.4. Analyzing the Scatter Plots

In Scatter plot analysis we try to find a pattern (relation or trend) for each output corresponding to each parameter of our model.

To get the result, we use 1000 sample in each run. The subplots of Figure 6 (x-axis represents outcome measures \( y \), \( Q \) and \( ymax \) respectively in each subplot and y-axis represents range of parameters (Table 1)) show that the most important parameters are the death rate, \( d \) and the maximum growth rate, \( \mu_m \) as we can clearly get a trend for these two parameters. Additionally, from the first and third subplot of Figure 6, we observe that nutrient uptake coefficient (\( \alpha \)) shows a good trend for both tumor volume, \( y \) and maximum tumor size, \( ymax \) but fails to be considered as important as \( d \) and \( \mu_m \) due to more spreading shape at the beginning. Moreover, it’s trending nature is far away from the trend of both \( d \) and \( \mu_m \) for the output measure cell quota, \( Q \) (second subplot of Figure 6).

Hence death rate, \( d \) and maximum growth rate, \( \mu_m \) are the two most sensitive parameters of our model.

Next, let us proceed to test the treatment strategies for sensitive parameters.

5. Treatment Strategy

In this section, we will suggest few treatment strategies depending on the results of model simulations for different values of the two most sensitive parameters (death rate, \( d \) and maximum growth rate, \( \mu_m \)).

- Figure 1 shows the primary solution curves of the model with base values (see Table 1) for \( d \) and \( \mu_m \).
- Figure 7 and Figure 8 show the change in tumor volume and cell quota respectively due to the change in \( d \) while \( \mu_m \) is fixed at 0.995.
Figure 9 and Figure 10 show change in tumor volume and cell quota respectively due to the change in $\mu_m$ while $d$ is fixed at 0.855.

Now we want to understand the relation of the two most sensitive parameters death rate, $d$ and maximum growth rate, $\mu_m$ with tumor volume, $y$ and cell quota, $Q$.

Figure 11 shows that the tumor volume, $y$ decreases with the increase of death rate, $d$ and increases as the maximum growth rate, $\mu_m$ increases. We also observe that for the value (approximately) of $d > 0.95$ and $\mu_m < 0.98$, tumor volume, $y$ approaches zero.

Figure 12 shows that the cell quota $Q$ remains zero until death rate, $d$ approaches to 1.4 (approximately) and then increases rapidly for a small increase in $d$. It also shows that $Q$ decreases rapidly for a small increase in maximum growth rate, $\mu_m$ until it reaches approximately 0.45. After that, $Q$ remains zero for any increasing value of $\mu_m$.

The following section shows the stability of equilibrium solutions.

Figure 3. PRCC and P-values Plot for Tumor Volume ($y$).

Figure 4. PRCC and P-values Plot for Cell Quota ($Q$).

Figure 5. PRCC and P-values Plot for Maximum Tumor Size ($y_{max}$).
Figure 6. Scatter plots for Tumor Volume, Cell Quota and Maximum Tumor Size, respectively.

Figure 7. Model Simulations for $\mu_m = 0.995$ and $d = 0.5, 0.9936, 1.2$, respectively.

Figure 8. Model Simulations for $d = 0.855$ and $\mu_m = 0.5, 0.85604, 1.2$, respectively.

Figure 9. Model Simulations for $d = 0.855$ and $\mu_m = 0.5, 0.85604, 1.2$, respectively.
6. Qualitative Analysis

In this section we are going to discuss the stability of our model. Since $d$ and $\mu_m$ are the two most sensitive parameters then they should play vital roles in the stability analysis of this model. Stability are of two types namely local stability and global stability.

6.1. Local Stability

Local stability of a system of ODE occurs only surrounding by a small neighborhood of equilibrium points. If we move far away from the neighborhood, local stability can be altered. Using the system (1), we consider

$$f(y, Q) = \mu_m \left[1 - \frac{Q}{y}\right] y - y d$$  \hspace{1cm} (2)

$$g(y, Q) = \frac{375 \alpha p_1}{y} - \mu_m (Q - q)$$  \hspace{1cm} (3)

We have used $\bar{r} = 375$ \cite{22} in the second equation of (1). Now for Equili-
In this mathematical model, \( y \) being the tumor volume, cannot be zero (also, \( y = 0 \) makes \( g(y, Q) \) undefined). So, the only biological meaningful equilibrium point of the system is

\[
E(y, Q) = E\left(\frac{375\alpha_m p_m (\mu_m - d)}{qd\mu_m}, \frac{q\mu_m}{\mu_m - d}\right)
\]

and for the choice of \( \mu_m > d \), \( E \) is always positive.

Now the Jacobian Matrix is defined as follows:

\[
J = \begin{bmatrix}
\mu_m \left(1 - \frac{q}{Q}\right) - q\mu_m y & \frac{\mu_y q y}{Q} \\
-375\alpha_m p_m y^2 & -\mu_m
\end{bmatrix}
\]

at, \( E \)

\[
J = \begin{bmatrix}
0 & \frac{375\alpha_m p_m (\mu_m - d)}{(q\mu_m)^2 d} \\
-(q\mu_m d^2) & -\mu_m
\end{bmatrix}
\]

So, the trace and determinant of the Jacobian Matrix are

\[
Tr(J) = -\mu_m < 0 \quad \text{(for any choice of } \mu_m\text{)}, \quad Det(J) = \frac{d}{\mu_m - d} > 0,
\]

for the choice of \( \mu_m > d \), which concludes that \( E \) is locally asymptotically stable as long as \( \mu_m > d \).

Let us now define the characteristic equation of the Jacobian matrix

\[
\lambda^2 - Tr(J)\lambda + Det(J) = 0
\]

\[
\Rightarrow \lambda^2 + \lambda\mu_m + \frac{d}{\mu_m - d} = 0
\]

\[
\Rightarrow \lambda = \frac{-\mu_m \pm \sqrt{\mu_m^2 - 4d}}{2}
\]

\[
\Rightarrow \lambda = \frac{-\mu_m \pm \sqrt{\mu_m^2 - \mu_m^2 d - 4d}}{2\sqrt{\mu_m - d}}
\]

The non-zero (see Table 1 for the value of \( \mu_m \)) real part of the eigenvalues \( \lambda \) shows that \( E \) is hyperbolic. So, the stability of \( E \) is robust \( i.e. \) for small perturbation on \( E \), the stability may distort but does not change the phase portrait near the equilibrium qualitatively as shown in Figure 13.
Also, the discriminant of the characteristic equation is

\[ D = \mu_m^2 - \frac{4d}{\mu_m - d} \]

Being locally asymptotically stable, \( E \) will be a stable node if \( D > 0 \) and will be a stable spiral if \( D < 0 \) [27].

6.2. Global Stability

If equilibrium points are stable everywhere (beyond of the small neighborhood) then they will be globally stable. We can check the global stability of our model using Lyapunov stability. Finding a Lyapunov function for our model will be cumbersome due to the number of parameters and the non-linearity of the model. So, we will start with the usual Lyapunov function for a two dimensional system of ODE

\[ V(y,Q) = \frac{1}{2}(y^2 + Q^2) \]  (4)

then

\[ \dot{V}(y,Q) = y^2 \mu_m - \frac{y^2 \mu_m q}{Q} - y^2 d + \frac{375 \alpha_1 p_j Q}{y} - \mu_m Q^2 + \mu_m qQ \]  (5)

\[ \Rightarrow \dot{V}(y,Q) = \left( \frac{y^2 \mu_m q}{Q} + y^2 d + \mu_m Q^2 \right) + \left( y^2 \mu_m + \frac{375 \alpha_1 p_j Q}{y} + \mu_m qQ \right) \]  (6)

For global stability,

\[ \dot{V}(y,Q) < 0 \]

\[ \Rightarrow a + b < 0 \]

\[ \Rightarrow y^3 (\mu_m Q - \mu_m q - Qd) + Q^2 (375 \alpha_1 p_j + y \mu_m q) - y \mu_m Q^3 < 0 \]

which is vague in terms of biological meaning. So, we want to adopt a different approach to cope with this situation.

Clearly, \( a \) will always dominate over \( b \) i.e. \( a > b \) which yields \( \dot{V}(y,Q) < 0 \).

To verify our statement numerically, we calculate \( \dot{V}(y,Q) \) for several (starting value, base value, ending value etc.) values of all parameters (using Table 1) and variables \( y \) and \( Q \). Every time, we get \( \dot{V}(y,Q) < 0 \) and by Lyapunov stability
theorem, $E$ is globally asymptotically stable.

We can also proceed with Poincaré-Bendixon theorem to verify that $E$ is globally asymptotically stable. Using Equations (2) and (3), we get

$$\frac{\partial f}{\partial y} + \frac{\partial g}{\partial Q} = -\frac{\mu m q}{Q} - d < 0$$

So, by Bendixon’s negative criterion, we can find a simply connected and positively invariant set containing no closed orbits. Note that, both $y$ and $Q$ will be bounded (for a quick guess see Figure 1) on that connected positively invariant set.

Then by Poincaré-Bendixon theorem, every solution starting in that connected, positively invariant set will approach to $E$. Hence $E$ is globally asymptotically stable.

7. Conclusions

Both PRCC and Scatter plot method techniques are very useful to identify the parameters that have significant impacts on a mathematical model. In our investigation, we used LHS to sample points for both these methods. Since both death rate, $d$ and maximum growth rate, $\mu m$ have significant effects on tumor volume, $y$, cell quota $Q$ and maximum tumor size $y_{max}$ identifying them as the most sensitive parameters can help us to introduce new treatment strategies in this field.

In this study, we observed and listed out the main findings:

1) When there is no treatment therapy for maximum growth rate $\mu_m = 1.58$ (see Table 1), then limiting the supply of nutrient affects the growth of cancer of cells, which results in higher death rate, $d$.

2) On the other hand, when there is a no treatment therapy for death rate, $d = 0.28$ (see Table 1), then reducing the maximum growth rate, $\mu_m$ will reduce the number of cancer cells.

3) Controlling nutrient supply for cancer cells with some level of treatment can have remarkable effects on cancer treatment strategies.

4) It is concluded about the treatment by referring Figure 11 that, if we can increase the death rate ($d > 0.95$) (by restricting nutrient supply or by using medicine) or control the maximum growth rate ($\mu_m < 0.85$) of tumor cells then the tumor volume will eventually get smaller and approach to zero (die out).

5) Both $\mu_m$ and $d$ controlled the stability of the model. Based on their relation ($\mu_m > d$), the system gets local asymptotic stability around the equilibrium point, see Figure 13.

6) Finally, we identified that the equilibrium point of our system obtains global asymptotic stability.

Acknowledgements

The authors are acknowledged to the anonymous reviewer for his constructive suggestions to improve the manuscript. The author M. Kamrujjaman research
was partially supported by TWAS, Grant: 2019_19-169 RG/MATHS/AS_I.

**Author’s Contributions**

This work was carried out in collaboration among all authors. All authors read and approved the final manuscript.

**Conflicts of Interest**

The authors declare no conflicts of interest regarding the publication of this paper.

**References**


DOI: 10.4236/jamp.2020.86073
Additional Mass: Orthotropic Membrane Material with Four Sides Fixed in Air Flow

Zhiying Tian¹, Weiju Song²*, Yafei Zhang², Huimin Yin², Xinxin Wang²

¹College of Civil Engineering, City College of Science and Technology, Chongqing University, Chongqing, China
²College of Civil Engineering, Hebei University of Engineering, Handan, China
Email: *nimrodsong@126.com

Abstract
When the membrane material in the air field vibrates, it will drive the movement of the surrounding air. The aerodynamic force generated by the moving air will act on the membrane material in turn, resulting in the change of dynamic characteristics such as membrane vibration frequency. In this paper, the additional air mass produced by membrane vibration in air is studied. Firstly, under the assumption that the incoming flow is uniform and incompressible ideal potential flow, the additional air mass acting on the surface is derived by using the thin airfoil theory and potential flow theory respectively. Then, according to the first law of thermodynamics and the principle of aeroelasticity, the analytical expression of the additional air mass is derived. Finally, through a specific example, the variation of the additional air mass with the membrane material parameters and pretension, as well as the influence of the aerodynamic force on the vibration frequency and amplitude of the membrane is obtained.

Keywords
Membrane Material, Additional Mass, Incompressible Fluid Motion, Wind Speed

1. Introduction
In the process of its vibration, the membrane material in the air fluid will cause the movement of air in a certain range around. Obviously, the acceleration or deceleration of the membrane material in the process of vibration will drive the acceleration or deceleration of the surrounding air. However, the accelerating or decelerating motion of the surrounding air will act on the vibrating membrane, so it will cause the change of its vibration frequency and other dynamic charac-
teristics. It is inevitable that the mass of the membrane should include the added value brought by the movement of the surrounding air in addition to its own properties. For most traditional structures, such as steel structure and concrete structure, the value of additional mass is very small compared with the structure itself, so it can be ignored in the analysis of structural response. But unfortunately, for the structure of membrane material, the additional mass is equal to or even more than the mass of membrane itself. Obviously, for the study of nonlinear vibration of membrane, the additional mass becomes a more important factor.

Miyake (1992) studied the added mass of the flexible plate in the stable fluid by analytical method. The results show that the added mass decreases with the increase of the number of vibration modes of the plate [1]; Li (2011), Wang (2011), Zhou (2014) studied the additional mass of films with different shapes under different air densities through theoretical calculation, numerical analysis and a series of tests, and gave the additional mass coefficient of about 0.65 [2] [3] [4]; Kim (2011), and the research of scholars such as Wu (2003) and Yang (2008, 2010) proposed a calculation method to simulate the wind structure interaction effect of tension structure. The added mass and aerodynamic damping are measured by experiments. Furthermore, the change of interaction parameters and its influence on free vibration characteristics are studied. And the mechanism of wind structure interaction is discussed. The results show that the ratio of added mass to self mass is about 0.2 to 1.2 [5] [6] [7]. Sun deduced the analytical expressions of additional mass and aerodynamic damping based on the principle of energy conservation, and analyzed the dynamic effects of open membrane structure and closed membrane structure respectively [8]. Li deduced the analytical expressions of additional mass, radiation damping and air bearing stiffness, and analyzed the influence parameters systematically [9].

In this paper, the additional mass produced by the vibration of membrane materials in the air is studied by analytical method with the consideration of the orthotropic characteristics of the membrane. Firstly, assuming that the incoming flow is a uniform incompressible ideal potential flow, the aerodynamic forces acting on the surface are derived by using the thin airfoil theory and the potential flow theory [10] [11]. Then, using the first law of thermodynamics and the principle of aeroelasticity, the analytical expression of additional mass is derived. Finally, the variation of the additional mass with the membrane material parameters and pretension is obtained by a specific example. The research results provide reasonable design reference for engineers, enrich the design theory of membrane structure, and further ensure the structural safety of tension membrane structure under wind load.

2. Derivation of Aerodynamic Force of Membrane Materials

The membrane material with four sides fixed in the air flow field is shown in Figure 1.
For the membrane material of air, due to the small membrane thickness, air flows from both sides of the membrane surface, which can be approximately determined by the thin wing theory. In the analysis, the vortex surface is used instead of the membrane surface, as shown in Figure 2. The vortex distribution per unit area of the membrane is defined as \( \gamma_c (x, y, t) \); the wake vortex force generated by the trailing edge of the membrane per unit time \((x > A) \) is defined as \( \gamma_w (x, y, t) \). Within the span of the membrane, the force generated by the surface vortex per unit length is represented by the vortex intensity (ring density), as shown in Figure 3.

According to thin wing theory [1],

\[
\gamma_c = \lim_{\Delta S \to 0} \frac{\int V dS}{\Delta S}
\]

(1)

where \( \Delta S \) is the width of the vortex surface enclosed by the circumference.

Suppose that \( p_1 \) is the indoor air pressure of the lower surface of the membrane, \( p_2 \) is the outdoor air pressure of the upper surface of the membrane, and that the incoming flow with the velocity \( V \) moves in the \( X \) direction without rotation. Then, according to the Bernoulli equation, the following equation can be obtained [12]:

\[
\rho_0 \left[ \frac{\partial \phi}{\partial t} + \frac{1}{2} (v_1^2 + v_y^2 + v_z^2) \right] + p_1 = \rho_0 \left[ \frac{\partial \phi}{\partial t} + \frac{1}{2} (v_2^2 + v_y^2 + v_z^2) \right] + p_2
\]

(2)
where, \( \phi \) is the velocity potential function of the upper and lower surfaces of the membrane, and \( v_{zi}, v_{ji}, v_{ai} \) are the velocity components of the upper and lower surfaces of the membrane respectively. Assuming that the incoming flow is along the \( Y \) direction of the structure, the velocity is \( V \), and the disturbance velocity in all directions is \( \bar{v}_{zi}, \bar{v}_{ji}, \bar{v}_{ai} \) when the flow field encounters obstacles [13], the velocity component in Equation (2) is

\[
v_{zi} = \bar{v}_{zi}, \quad v_{ji} = V + \bar{v}_{ji}, \quad v_{ai} = \bar{v}_{ai}
\]

Generally, \( \bar{v}_{zi}, \bar{v}_{ji}, \bar{v}_{ai} \ll V \), then, omitting high order small quantity term has general

\[
v_{zi}^2 + v_{ji}^2 + v_{ai}^2 = \bar{v}_{zi}^2 + (V + \bar{v}_{ji})^2 + \bar{v}_{ai}^2 \approx V^2 + 2V\bar{v}_{ji}
\]

(3)

Substituting Equation (3) into Equation (2) and simplifying it, then,

\[
p_1 - p_2 = \rho_b \left[ \frac{\partial \phi}{\partial t} - \frac{\partial \phi}{\partial t} \right] + V \left( \bar{v}_{zi} - \bar{v}_{ji} \right)
\]

(4)

Velocity and velocity potential can be approximately considered as functions of horizontal coordinates \( x, y \) and \( t \).

\[
\phi' = \int_0^x v_x \, dx + \int_0^y v_y \, dy \approx \int_0^x v_x \, dx
\]

(5)

Then,

\[
\frac{\partial \phi}{\partial t} - \frac{\partial \phi}{\partial t} = \int_0^y (v_{zi} - v_{ji}) \, dy
\]

(6)

Assuming that the vortex density on the surface element \( dx \, dy \) is \( \gamma_c(x, y, t) \), applying the thin airfoil theory, then:

\[
v_{zi} - v_{ji} = \gamma_c(x, y, t)
\]

(7)

Substituting Equations (6) and (7) into Equations (4), the aerodynamic force acting on the membrane unit can be obtained as follows:

\[
p = p_1 - p_2 = \rho_b \frac{\partial}{\partial t} \int_0^y \gamma_c(x, \eta, t) \, d\eta + \rho_b V \gamma_c
\]

(8)

The vortex lattice method is used to solve the expression of \( \gamma_c(x, y, t) \) in Equation (8). The projection area \( \{0 \leq x \leq a, 0 \leq y \leq b\} \) of the membrane on the \( xoy \) plane is divided into \( M \times N \) vortex grids. The dimensionless vortex strength \( \Gamma / aV \) is expressed by \( \gamma_1, \gamma_2, \gamma_3, \cdots, \gamma_{M \times N} \). The induced velocity \( v_{aij} \) at the control point \( i \) of the \( j \)th vortex grid is

\[
\frac{v_{aij}}{V} = C_{ij} \gamma_j
\]

(9)

where \( C_{ij} \) is the value of \( v_{aij}/V \) generated by \( \gamma_j \) at point \( i \), and the \( Z \)-induced velocity generated by all horseshoe vortices at point \( i \) is

\[
\frac{v_{ai}}{V} = \sum_{j=1}^{M \times N} C_{ij} \gamma_j
\]

(10)

The horseshoe vortex on the membrane surface is shown in Figure 4. The coordinates of corner points 1 and 2 are \( (x_{i1}, y_{i1}) \) and \( (x_{i2}, y_{i2}) \) respectively, and the coordinates of control point \( i \) are \( (x_i, y_i) \), then:
Figure 4. Horseshoe vortex over the membrane surface.

\[ C_{ij} = \left( \frac{v_{ij}}{V} \right)_{x_j = 1} = \frac{a}{4\pi} \left\{ 1 - \left( \frac{y_j - y_i}{y_j - y_i} \right) \left( x_j - x_i \right) - \left( y_j - y_i \right) \left( x_j - x_i \right) \right\} \]

\[ \times \left( \frac{y_{1} - y_i}{y_j - y_i} \right) \left( x_j - x_i \right) + \left( x_j - x_i \right) \left( y_j - y_i \right) \right\} \]

\[ - \left( y_{2j} - y_i \right) \left( y_j - y_i \right) + \left( x_j - x_i \right) \left( y_j - y_i \right) \right\} \]

\[ \frac{1}{\sqrt{x_j - x_i}^{2} + \sqrt{x_j - x_j}^{2}} \]

\[ + \frac{1.0}{x_j - x_i} \left[ \frac{y_j - y_i}{\sqrt{y_j - y_j}^{2} + \sqrt{x_j - x_i}^{2}} \right] \]

\[ - \frac{1.0}{x_j - x_i} \left[ \frac{y_j - y_i}{\sqrt{y_j - y_j}^{2} + \sqrt{x_j - x_j}^{2}} \right] \]

The expressions of coordinates of each point in the formula are as follows:

\[ x_i = x_{[i-1]N+k} = \frac{a}{M} \left( l - \frac{1}{2} \right) \quad (1 \leq l \leq M) \]

\[ y_i = y_{[i-1]N+k} = \frac{b}{N} \left( k - \frac{1}{4} \right) \quad (1 \leq k \leq N) \]

\[ x_j = x_{[(l-1)N+k]} = \frac{a}{M} \left( l - 1 \right) \quad (1 \leq l \leq M) \]

\[ y_j = y_{[(l-1)N+k]} = \frac{b}{N} \left( k - \frac{3}{4} \right) \quad (1 \leq k \leq N) \]

\[ x_{2j} = x_{[(l-1)N+k]} = \frac{a}{M} \left( l \right) \quad (1 \leq l \leq M) \]

\[ y_{2j} = y_{[(l-1)N+k]} = \frac{b}{N} \left( k - \frac{3}{4} \right) \quad (1 \leq k \leq N) \]

Applying boundary conditions (11) to the 4th control point, then

\[ \frac{v_{ij}}{V} = \sum_{j=1}^{M \times N} C_{ij} = \left[ \frac{\partial v}{\partial y} + \frac{1}{V} \frac{\partial v}{\partial t} \right] = \left[ \frac{\partial v}{\partial y} + \frac{\partial w}{\partial y} + \frac{1}{V} \frac{\partial w}{\partial t} \right] \]

The vibration displacement of the membrane is assumed to be:
where, \( T(t) \) is a function of time in the process of vibration, and \( W(x, y) \) is a function of mode shape.

\[
\sum_{j=1}^{MN} C_j \gamma_j = \frac{\partial W}{\partial y} T(t) + \frac{T'(t)}{V} W \quad (i = 1, 2, \ldots, M \times N)
\]

The value of \( \gamma_j \) can be obtained by combining the equations on \( M \times N \) control points.

Let the expression of \( \gamma_j \) be

\[
\gamma_j = a_j T(t) + a_{2j} \frac{T'(t)}{V}, \quad (j = 1, 2, \ldots, M \times N)
\]

Substituting Equation (15) into Equation (14), then

\[
\sum_{j=1}^{MN} C_j \left( a_j T(t) + a_{2j} \frac{T'(t)}{V} \right) = \frac{\partial W}{\partial y} T(t) + \frac{T'(t)}{V} W
\]

So,

\[
\sum_{j=1}^{MN} C_j a_j = \frac{\partial W}{\partial y}, \quad (i = 1, 2, \ldots, M \times N)

\sum_{j=1}^{MN} C_j a_{2j} = W, \quad (i = 1, 2, \ldots, M \times N)
\]

\( \gamma_j \) in Equation (14) is dimensionless vortex strength, then the expression of vortex strength \( \gamma_c \) is:

\[
\gamma_c = aV \gamma_j = aV \left( a_j T(t) + a_{2j} \frac{T'(t)}{V} \right), \quad (j = 1, 2, \ldots, M \times N)
\]

### 3. Analytical Derivation of Additional Mass

It is assumed that the air is an incompressible and in viscid ideal fluid, and the motion of the air fluid is caused by the membrane vibration [12]. Therefore, the kinetic energy increment of any part of the fluid is equal to the work done by the surface pressure \( P \) in \( dt \) time [8].

\[
dE_i = -\iint_S V \rho dS dt
\]

where, \( E_i \) is the kinetic energy, \( \rho \) is the aerodynamic pressure on the membrane surface, \( V \) is the normal velocity of the air particle, which is equal to the normal vibration velocity of the membrane, and the integral region is \( S \in \{0 \leq x \leq a, 0 \leq y \leq b\} \), then

\[
dE_i = -\iint_S \frac{\partial w(x, y, t)}{\partial t} \rho dxdy dt
\]

According to reference [11], the kinetic energy change rate of the additional air mass can be expressed as

\[
\frac{dE_i}{dt} = m \iint_S \frac{\partial^2 w(x, y, t)}{\partial t^2} \rho dS
\]
where \( m_a \) is the additional air mass of the membrane (kg/m²), substituting Equations (20) into (21), we can get:

\[
m_a = \frac{-\int_0^b \int_0^a p \left( \frac{\partial w(x,y,t)}{\partial t} \right) dx dy}{\int_0^b \int_0^a \left( \frac{\partial w(x,y,t)}{\partial t} \right) \cdot \left( \frac{\partial^2 w(x,y,t)}{\partial t^2} \right) dx dy}
\]  

(22)

4. Parameter Analysis and Discussion

Let the displacement function satisfy the boundary conditions as [14] [15] [16] [17] [18].

\[ w(x,y,t) = a_0 \sin \omega t \cdot \sin \frac{m\pi x}{a} \sin \frac{n\pi y}{b} \]  

(23)

By substituting the displacement function (23) into the aerodynamic expression (8) of the open membrane, the aerodynamic pressure acting on the membrane surface can be obtained as follows:

\[
p = \rho_0 \frac{\partial}{\partial t} \int_0^b \int_0^a V \left( a_{ij} a_0 \sin \omega t + a_{ij} \frac{\omega a_0 \cos \omega t}{V} \right) d\eta
\]

\[
+ \rho_0 V^2 a \left( a_{ij} a_0 \sin \omega t + a_{ij} \frac{\omega a_0 \cos \omega t}{V} \right)
\]

\[
= \rho_0 a V a_0 \cos \omega t \int_0^b \int_0^a d\eta - a a_0 \omega^2 \sin \omega t \int_0^b \int_0^a a_{ij} d\eta
\]

\[
+ \rho_0 V^2 a a_0 \sin \omega t + \rho_0 V a a_0 \omega \cos \omega t
\]

(24)

By substituting Equation (24) with Equation (22), the additional air mass expression of the membrane can be obtained

\[
m_a = \frac{\int_0^b \int_0^a \rho_0 a V a_0 \cos \omega t \left( a_{ij} a_0 \sin \omega t - a a_0 \omega^2 \sin \omega t \int_0^b \int_0^a a_{ij} d\eta + \rho_0 V a a_0 a_{ij} V^2 \sin \omega t + \rho_0 V a a_0 \omega \cos \omega t \right) W dx dy}{\int_0^b \int_0^a \int_0^b \int_0^a W \cdot a_{ij} \omega^2 \sin \omega t \cdot W dx dy}
\]

\[
= \frac{4 \rho_0}{b \omega^2} \beta_1 \cdot V^2 + \frac{4 \rho_0}{b \omega} \cos \omega t \cdot \left( \beta_2 + \beta_3 \right) \cdot V - \frac{4}{b} \beta_4
\]

(25)

where,

\[
\beta_1 = \int_0^b \int_0^a a_{ij} \sin \frac{m\pi x}{a} \sin \frac{n\pi y}{b} dx dy
\]

\[
\beta_2 = \int_0^b \int_0^a \frac{m\pi x}{a} \sin \frac{n\pi y}{b} dx dy
\]

\[
\beta_3 = \int_0^b \int_0^a \frac{m\pi x}{a} \sin \frac{n\pi y}{b} d\eta dx dy
\]

\[
\beta_4 = \int_0^b \int_0^a a_{ij} \sin \frac{m\pi x}{a} \sin \frac{n\pi y}{b} d\eta dx dy
\]

4.1. Static Air Field (V = 0)

When the wind speed \( V = 0 \), the variation relationship between the additional air mass and the vibration mode of the open membrane under different cross wind direction span ratio is shown in Figure 5.
It can be seen from Figure 5 that with the increase of mode number, the additional mass decreases gradually, and the influence of mode on the additional air mass decreases gradually. The reason is that the vibration amplitude of high-order mode of membrane is small, and the influence of aerodynamic force on membrane vibration is relatively weak. The motion of the first-order mode of membrane is volume expansion state, and the vibration amplitude is the largest. In addition, the additional mass increases with the increase of the span ratio of the membrane under the same mode.

4.2. Flowing Air Field (\(V \neq 0\))

Assuming that the wind speed is along the \(X\) direction, let \(\lambda = b/a\) is the span ratio of the transverse (\(Y\)) and longitudinal (\(X\)) wind directions; \(\gamma = N_{0x}/N_{0y}\) is the pretension ratio of the longitudinal (\(X\)) and transverse (\(Y\)) wind directions. Take the parameters of membrane as \(a = 20\) m, \(f = 1\) m, \(N_{0x} = 2\) kN/m, \(\gamma = 1\), \(\lambda = 1\).

The curve of the additional mass ratio with the wind speed under different modes is shown in Figure 6. With the increase of wind speed, the additional mass ratio decreases gradually, because in the process of membrane vibration caused by air, the excitation effect of wind increases, while the composition of membrane self-excited vibration decreases. Therefore, the aeroelastic coupling effect between membrane and air is reduced. Still, it can be seen that the first mode has the strongest additional aerodynamic force between the membrane and the surrounding air, which is much larger than the second and the third modes.

It can be seen from Figure 7 that with the increase of pretension, the additional air mass of the membrane decreases gradually, but the reduction amplitude is weaker than that of wind speed, because the increase of pretension makes the vibration stiffness of the membrane increase, so that the vibration response of the membrane becomes smaller under the same wind speed, thus reducing the interaction between the wind and the membrane. On the other hand, it also reflects that when the membrane stiffness is lower, the influence of the additional mass on the vibration process of the membrane cannot be ignored.

![Figure 5. Additional air mass for open membrane.](image-url)
Figure 6. Influence to additional mass of flat membrane with wind speed.

Figure 7. Influence to additional mass of flat membrane with the initial pretension.
5. Conclusions

According to the first law of thermodynamics and the principle of aeroelasticity, the additional mass of the membrane is solved analytically in this paper, and the parameters are analyzed. It provides a theoretical basis for further study of the aerodynamic damping of membranes and the influence of additional aerodynamic force on the vibration characteristics of membranes. The main conclusions are as follows:

1) The additional mass of the first mode of membrane is the largest, while that of the other three modes is relatively small.

2) With the increase of mode number, the additional mass decreases gradually.

3) In the same mode, the additional mass increases with the increase of the membrane span ratio.

4) The additional mass of the membrane decreases with the increase of pretension, and the decrease amplitude of the high-order mode is smaller than that of the low-order mode.

Acknowledgements

The work is supported by the National Natural Science Foundation of China (Grant No. 51608060) and the Innovation fund of Hebei University of Engineering (Grant No. SJ010002159).

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References


**Nomenclatures**

- \( a \): length of the membrane
- \( b \): width of the membrane
- \( E_k \): the kinetic energy
- \( M_s \): areal density of membrane
- \( M_a \): the additional air mass of the membrane
- \( N_{0x} \): the initial stress of membrane in \( X \) direction
- \( N_{0y} \): the initial stress of membrane in \( Y \) direction
- \( P \): aerodynamic pressure on the membrane surface
- \( P_1 \): the indoor air pressure of the lower surface of the membrane
- \( P_2 \): the outdoor air pressure of the upper surface of the membrane
- \( T(t) \): a function of time in the process of vibration
- \( W(x, y) \): a function of mode shape
- \( V \): wind velocity
- \( v_{xi}, v_{yi}, v_{zi} \): the velocity components of the upper and lower surfaces
- \( \phi \): the velocity potential function of the upper and lower surface
- \( \gamma = N_{0x}/N_{0y} \): the pretension ratio of the longitudinal (\( X \)) and transverse (\( Y \)) wind directions
- \( \gamma_v(x, y, t) \): the vortex density on the surface element
- \( \gamma_w(x, y, t) \): the wake vortex force
- \( \lambda = b/a \): the span ratio of the transverse (\( Y \)) and longitudinal (\( X \)) wind directions;
- \( \rho \): air density.
Some Chaotic Properties of a Kind of Coupled Map Lattices

Xiaofang Yang1, Tianxiu Lu1,2*, Guo Liu1

1Department of Mathematics and Statistics, Sichuan University of Science and Engineering, Zigong, China
2Artificial Intelligence Key Laboratory of Sichuan Province, Zigong, China

Email: xf_suse@163.com, *lubeltx@163.com, lg0506@outlook.com

Abstract
This paper is concerned with some chaotic properties of a kind of coupled map lattices, which is proposed by Kaneko. First, this research discussed the sensitivity, infinite sensitivity, transitivity, accessibility, densely Li-Yorke sensitivity and exact of coupled map lattices. Then, some sufficient conditions under which \( \Delta^\infty_\varepsilon, d, F \) is Kato chaotic, positive entropy chaotic and Ruelle-Takens chaos are obtained.

Keywords
Coupled Map Lattices, Sensitivity, Chaos, Accessibility

1. Introduction
In 1983, Kaneko [1] proposed coupled map lattices (Short for CMLs). Then, in biophysics, materials, chaos, image processing, CMLs are intensively discussed (Refer to literature [2]-[8] and others). In 2005, the literature [9] showed that CMLs have some topology and ergodic properties. In 2010, Juan Lu [10] presented a definition of distributional chaos on a sequence (DCS) for CML systems and stated two different sufficient conditions for having DCS. In 2010, Juan Luis [11] proved that this CML system has positive topological entropy for zero coupling constant. In 2016, Risong Li [12] [13] had obtained some relevant conclusions for the zero coupling constant and proved that the system has three kinds of chaos. In this paper, the following CML from [14] is considered.

\[
x_{m+1,n} = (1 - \varepsilon) f(x_{m,n}) + \varepsilon f(x_{m,n-1})
\]

where \( x_{m,n} \in I \), \( m \in \mathbb{N}_0 = \{0,1,2,\ldots\} \), \( n \in \mathbb{Z} = \{\ldots,-1,0,1,\ldots\} \), \( I \) is a non-degenerate compact interval, \( f \) is a map on \( I \), and \( \varepsilon \in [0,1] \) is a constant.
For $t \in \mathbb{Z}$, let $\mathbb{N}_0 = \{t, t+1, \cdots \}$ and $\Omega = \{(0, n) : n \in \mathbb{Z}\} = \{(0, -1), (0, 0), (0, 1), \cdots \}$. For any sequence $\phi = \{\phi_{0,n}\}_{n=0}^\infty$ on $\Omega$, by induction, one can obtain a double-indexed sequence $x = \{x_{m,n} : m = 0, 1, 2, \cdots ; n = \cdots, -1, 0, 1, \cdots \}$, which is said to be a solution of the above system (1) with initial condition $\phi$.

Let $I$ be a subset of real number set, write

$$I_\infty^x = \{(a_n)_{n=0}^\infty = (\cdots, a_{-1}, a_0, a_1, \cdots) : a_n \in I, n \in \mathbb{Z}\}$$

and

$$\Delta_\infty^x = \{(\cdots, a_{-1}, a_0, a_1, \cdots) : a_i = a_j \in I, i, j \in \mathbb{Z}\}$$

which is called the diagonal set of $I_\infty^x$.

For arbitrary, two sequences $x_1 = \{x_{1,n}\}_{n=0}^\infty, x_2 = \{x_{2,n}\}_{n=0}^\infty \in I_\infty^x$, it is easy to prove that

$$d(x_1, x_2) = \sup\{|x_{1,n} - x_{2,n}| : n = \cdots, -1, 0, 1, \cdots\} \quad (2)$$

is a metric on $I_\infty^x$.

Let $f : I \mapsto I$ be a continuous map and $x = \{x_{m,n} : m \in \mathbb{N}_0, n \in \mathbb{Z}\}$ be a solution of the above system (1) with initial condition $\phi = \{\phi_{0,n}\}_{n=0}^\infty \in I_\infty^x$.

Let

$$x_m = \{x_{m,n}\}_{n=-\infty}^\infty = (\cdots, x_{m,-1}, x_{m,0}, x_{m,1}, \cdots), \forall m \in \mathbb{N}_0,$$

and

$$x_{m+1} = \{x_{m+1,n}\}_{n=-\infty}^\infty = (\cdots, x_{m+1,-1}, x_{m+1,0}, x_{m+1,1}, \cdots) = F(x_m), \forall m \in \mathbb{N}_0,$$

where

$$x_0 = \phi = \{x_{0,n}\}_{n=-\infty}^\infty \quad \text{and} \quad x_{m+1} = (1-\varepsilon) f(x_m) + \varepsilon f(x_{m,n-1}), \forall m \in \mathbb{N}_0, n \in \mathbb{Z}.$$

Then, one can see that the above system (1) is equivalent to the following system

$$x_{m+1} = F(x_m), x_m \in I_\infty^x, m = 0, 1, 2, \cdots \quad (3)$$

For the above system (3), the map $F$ is said to be induced by the system (1). Obviously, a double-indexed sequence $\{x_{m,n} : m \in \mathbb{N}_0, n \in \mathbb{Z}\}$ is a solution of the above system (1) if and only if the sequence $\{x_m = \{x_{m,n}\}_{n=-\infty}^\infty : m \in \mathbb{N}_0\}_{m=0}^\infty$ is a solution of the above system (3).

Next section, the definitions of sensitive, infinite sensitive, transitive, accessibility, densely Li-Yorke sensitive and exact will be reviewed. And then, in section 3, it is proved that the system $\left(\Delta_\infty^x, d, \{F\}_{\mathbb{N}_0}^\infty\right)$ satisfies three definitions of chaos (Kato chaotic, positive entropy chaotic and Ruelle-Takens chaos) under the conditions that $f$ is chaos in these sense.

2. Preliminaries

After T. Y. Li and J. A. Yorke [15] first put forward the mathematical definition of “chaos”, many other definitions of chaos appeared later. For example, seni-
tivity, infinite sensitivity, transitivity, accessibility, densely Li-Yorke sensitivity, Kato chaotic, positive entropy chaotic, Ruelle-Takens chaos, and so on.

**Definition 1.** Let \((X, \rho)\) be a metric space and \(f : X \mapsto X\) be a continuous function. \(f\) is said to be

1) transitive if for any nonempty open subsets \(U_1, U_2 \subset X\), \(f^n(U_1) \cap U_2 \neq \emptyset\) for some integer \(n \in \mathbb{N}\) (see [16]).

2) sensitive if there exist \(\eta > 0\) such that for any \(x \in X\) and \(\varepsilon > 0\), there exists \(y \in B(x, \varepsilon)\) and \(n \in \mathbb{N}\) such that \(\rho(f^n(x), f^n(y)) \geq \eta\) (see [17]).

3) infinitely sensitive if there exist \(\eta > 0\) such that for any \(x \in X\) and \(\varepsilon > 0\), there exists \(y \in B(x, \varepsilon)\) and \(n \in \mathbb{N}\) such that \(\limsup_{n \to \infty} \rho(f^n(x), f^n(y)) \geq \eta\) (see [17]).

4) accessible if for any \(\varepsilon > 0\) and any two nonempty open subsets \(U_1, U_2 \subset X\), there are two points \(x \in U_1\) and \(y \in U_2\) such that \(\rho(f^n(x), f^n(y)) < \varepsilon\) for some integer \(n > 0\) (see [16]).

5) exact if for any open subset \(U \subset X\), there is \(m \in \mathbb{N}\) such that \(f^m(U) = X\) (see [18]).

**Remark 1.** [19] There is another equivalent definition of transitivity: \(f : X \mapsto X\) is said to be transitivity, if there is an \(x_0 \in X\) such that \(\text{Orb}_f(x_0) = X\). Where, \(\text{Orb}_f(x_0) = \{f^n(x_0) | n = 0, 1, 2, \cdots\}\) is called the orbit of the point \(x_0\).

**Definition 2.** 1) A dynamic system \((X, f)\) (or the map \(f : X \mapsto X\)) is Li-Yorke sensitive, if for any \(x \in X\) has \(x \in Q_\delta(f)\) for some \(\delta > 0\).

2) A dynamic system \((X, f)\) (or the map \(f : X \mapsto X\)) is densely Li-Yorke sensitive if \(Q_\delta(f)\) is dense in \(X\) for some \(\delta > 0\). Among them, \(Q_\delta(f) = \{x \in X : \forall \varepsilon > 0, \exists y \in B(x, \varepsilon)\text{ such that } (x, y) \in LY_\delta(f, \delta)\}\)

\[LY_\delta(f, \delta) = \{(x, y) \in X \times X : \limsup_{n \to \infty} \rho(f^n(x), f^n(y)) > \delta\}\]

and \(\liminf_{n \to \infty} \rho(f^n(x), f^n(y)) = 0\).

**Definition 3.** 1) A dynamic system \((X, f)\) (or the map \(f : X \mapsto X\)) is Kato chaotic if it is sensitive and accessible (see [20]).

2) A dynamic system \((X, f)\) (or the map \(f : X \mapsto X\)) is chaotic in the sense of Ruelle and Takens (short for R-T chaotic) if it is transitive and sensitive (see [21]).

**Proposition 1.** A dynamic system \((X, f)\) (or the map \(f : X \mapsto X\)) is Li-Yorke sensitive if and only if \(P_\delta(f) = X\) for some \(\delta > 0\). Among them, \(P_\delta(f) = \{x \in X : \forall \varepsilon > 0, \exists y \in B(x, \varepsilon), \exists n \in \mathbb{N}\text{ such that } \rho(f^n(x), f^n(y)) > \delta\}\)

**Proposition 2.** [17] A dynamical system \((X, f)\) is infinitely sensitive if and only if it is sensitive.

**Proposition 3.** [22] A dynamical system \((X, f)\) is dense Li-Yorke sensitivity, then it is Topological mixing (or its topological entropy is positive).
3. Main Results

In this section, let $X = I$. The metric $\rho$ in $I$ is defined by
$\rho(a,b) = \left| a - b \right| (\forall a,b \in I)$. The metric $d$ in $I^\infty$ is defined by (2).

**Theorem 1.** If $f$ is transitive, then the system $\left( \Delta^\infty_x, d, F^\infty \right)$ is transitive.

**Proof.** Since $f$ is transitive, then there exist $a \in I$ satisfying $\text{Orb}_f(a) = I$. Then for any $b \in I$ and any $\varepsilon > 0$, $B(b,\varepsilon) \cap \text{Orb}_f(a) \neq \emptyset$. That is, there exists a $k_0 > 0$ such that $\rho\left( f^{k_0}(a), b \right) = \left| f^{k_0}(a) - b \right| < \varepsilon$. Take $x_0 = (\cdots, a, a, a, \cdots) \in \Delta^\infty_x$. It is easy to see, for any $k \in \mathbb{N}$,
$F^k(x_0) = \left\{ f^k(a) \right\}^\infty_{n=-\infty}$. Then,
$\text{Orb}_f(x_0) = \left\{ f^k(a) \right\}^\infty_{n=-k}$. For any $y = (\cdots, b, b, b, \cdots) \in \Delta^\infty_x$ and above $k_0 > 0$,
$d\left( F^{k_0}(x_0), y \right) = \sup\left\{ \left| f^{k_0}(a) - b \right| : n \in \mathbb{N} \right\} = \left| f^{k_0}(a) - b \right| < \varepsilon.
So, $B(y,\varepsilon) \cap \text{Orb}_f(x_0) \neq \emptyset$.

Thus, the system $\left( \Delta^\infty_x, d, F^\infty \right)$ is transitive.

**Theorem 2.** If $f$ is sensitive, then the system $\left( \Delta^\infty_x, d, F^\infty \right)$ is sensitive.

**Proof.** Take $\Delta^\infty_x = \{\cdots, x_1, x_0, x_0, \cdots : x_n = a \in I, n \in \mathbb{Z} \} \subset I^\infty_x$,
$\forall x = (\cdots, a, a, a, \cdots) \neq y = (\cdots, b, b, b, \cdots) \in \Delta^\infty_x$. It is easy to know that, for any $k \in \mathbb{N}$,
$F^k(x) = \left\{ f^k(a) \right\}^\infty_{n=-\infty}, F^k(y) = \left\{ f^k(b) \right\}^\infty_{n=-\infty}.
So, for $\forall k \in \mathbb{N}$,
$d\left( F^k(x), F^k(y) \right) = d\left( \left\{ f^k(a) \right\}^\infty_{n=-\infty}, \left\{ f^k(b) \right\}^\infty_{n=-\infty} \right) = \sup\left\{ \left| f^k(a) - f^k(b) \right|, k = \cdots, 1, 0, 1, \cdots \right\} = \left| f^k(a) - f^k(b) \right|.$

Since $f$ is sensitive, so there exists a $\varepsilon_0 > 0$ such that for any $p \in I$ and any $\delta > 0$, there exists a $q_{p,\delta} \in B(p,\delta)$ and $n_{p,\delta} \in \mathbb{N}$ such that $\left| f^{n_{p,\delta}}(p), f^{n_{p,\delta}}(q_{p,\delta}) \right| > \varepsilon_0$. So for any fixed $x = (\cdots, p, p, p, \cdots) \in \Delta^\infty_x$ and any $\delta > 0$, taking $y = (\cdots, q_{p,\delta}, q_{p,\delta}, q_{p,\delta}, \cdots) \in \Delta^\infty_x$, one has that,
$\left| f^{n_{p,\delta}}(x), f^{n_{p,\delta}}(y) \right| = \left| f^{n_{p,\delta}}(p), f^{n_{p,\delta}}(q_{p,\delta}) \right| > \varepsilon_0, \left| p - q_{p,\delta} \right| < \delta,$
that is $y \in B(x,\delta)$. And because
$d\left( F^{n_{p,\delta}}(x), F^{n_{p,\delta}}(y) \right) = \left| f^{n_{p,\delta}}(p) - f^{n_{p,\delta}}(q_{p,\delta}) \right| > \varepsilon_0,$
so $F^\infty_{\Delta^\infty_x}$ is sensitive.

**Corollary 1.** If $f$ is chaotic in the sense of Ruelle and Takens, then the system $\left( \Delta^\infty_x, d, F^\infty \right)$ is chaotic in the sense of Ruelle and Takens.

**Proof.** According to Theorem 1, Theorem 2 and the definition of R-T chaos,
the conclusion is obvious.

According to Proposition 2 and Theorem 2, the following Corollary is hold.

**Corollary 2.** If \( f \) is infinitely sensitive, then the system \( \left( \Delta^\infty_\alpha, d, F|_{\Delta^\infty_\alpha} \right) \) is sensitive.

In fact, there is a stronger conclusion.

**Theorem 3.** If \( f \) is infinitely sensitive, then the system \( \left( \Delta^\infty_\alpha, d, F|_{\Delta^\infty_\alpha} \right) \) is infinitely sensitive.

**Proof.** Since \( f \) is infinitely sensitive, then there exists a \( \delta > 0 \) such that for any \( a \in I \) and any \( \epsilon \), there exists \( b_{a, \epsilon} \in B(a, \epsilon) \) and \( n_{a, \epsilon} \in \mathbb{N} \) such that
\[
\limsup_{n \to \infty} \rho\left(f^{n_{a, \epsilon}}\left(a\right), f^{n_{a, \epsilon}}\left(b_{a, \epsilon}\right)\right) \geq \delta.
\]
So for any fixed \( x = (\cdots, a, a, a, \cdots) \in \Delta^\infty_\alpha \), and any \( \epsilon > 0 \), taking \( x = (\cdots, b_{a, \epsilon}, b_{a, \epsilon}, b_{a, \epsilon}, \cdots) \in \Delta^\infty_\alpha \), one has that
\[
d\left(x, y\right) = \sup\{a - b_{a, \epsilon} \mid d - b_{a, \epsilon} \mid \} = |a - b_{a, \epsilon}| \leq \epsilon,
\]
that is \( y \in B\left(x, \epsilon\right) \). And because
\[
\limsup_{n_{a, \epsilon} \to \infty} d\left(F^{n_{a, \epsilon}}\left(x\right), F^{n_{a, \epsilon}}\left(y\right)\right) = \limsup_{n_{a, \epsilon} \to \infty} \rho\left(f^{n_{a, \epsilon}}\left(a\right), f^{n_{a, \epsilon}}\left(b_{a, \epsilon}\right)\right) \geq \delta.
\]
So \( F|_{\Delta^\infty_\alpha} \) is infinitely sensitive.

**Theorem 4.** If \( f \) is accessible, then the system \( \left( \Delta^\infty_\alpha, d, F|_{\Delta^\infty_\alpha} \right) \) is accessible.

**Proof.** For any open subset
\[
\left( \Delta^\infty_1\right) = \left\{x_1, x_2, \ldots, x_{1, \infty}, x_{\infty} \mid a \in U_1 \cap I, n \in \mathbb{Z} \right\} \subset I^\infty_1
\]
and
\[
\left( \Delta^\infty_2\right) = \left\{y_1, y_2, \ldots, y_{1, \infty}, y_{\infty} \mid b \in U_2 \cap I, n \in \mathbb{Z} \right\} \subset I^\infty_2,
\]
since \( f \) is accessible, then, for the above \( U_1, U_2 \subset I \), there exist \( a \in U_1, b \in U_2 \) such that
\[
\rho\left(f^k\left(a\right), f^k\left(b\right)\right) = |f^k\left(a\right) - f^k\left(b\right)| < \epsilon
\]
for some \( k > 0 \). Take
\[
x = (\cdots, a, a, a, \cdots) \in \left( \Delta^\infty_1\right), y = (\cdots, b, b, b, \cdots) \in \left( \Delta^\infty_2\right),
\]
then
\[
d\left(F^k\left(x\right), F^k\left(y\right)\right) = |f^k\left(a\right) - f^k\left(b\right)| < \epsilon.
\]
So, the system \( \left( \Delta^\infty_\alpha, d, F|_{\Delta^\infty_\alpha} \right) \) is accessible.

**Corollary 3.** If \( f \) is Kato chaotic, then the system \( \left( \Delta^\infty_\alpha, d, F|_{\Delta^\infty_\alpha} \right) \) is Kato chaotic.

**Proof.** According to Theorem 2 and Theorem 4, the conclusion is obvious.

**Theorem 5.** If \( f \) is exact, then the system \( \left( \Delta^\infty_\alpha, d, F|_{\Delta^\infty_\alpha} \right) \) is exact.

**Proof.** Since \( f \) is exact, for any open subset \( D \subset I \), there exist \( m \in \mathbb{N} \) such that
\[
f^m\left(D\right) = I.
\]
That is, for any \( a \in D \), there exists an \( m > 0 \) such that
\[
B\left(f^m\left(a\right), \epsilon\right) \cap I \neq \emptyset
\]
for any \( \epsilon > 0 \). So there is a \( b \in X \) such that
\[
\rho\left(f^m\left(a\right), b\right) = |f^m\left(a\right) - b| < \epsilon.
\]
Take \((\Lambda^\ast)^\circ\) is arbitrary open subset of \(\Delta^\circ\), and \(x_0 = (\cdots, a, a, a, \cdots) \in (\Lambda^\ast)^\circ\).

Clearly, for any \(k \in \mathbb{N}\), \(F^k(x_0) = \{f^k(a)\}_{n=-\infty}^\infty\). For any \(y_0 = (\cdots, b, b, b, \cdots) \in \Delta^\circ\), \(d(F^n(x_0), y_0) = |f^n(a) - b| < \varepsilon\). That is to say, there exist an \(m \in \mathbb{N}\), \(F^n(\Lambda^\ast)^\circ) = \Delta^\circ\). So, the system \((\Lambda^\circ, d, F|_{\Lambda^\circ})\) is exact.

In [23] we had proved that, \(f\) is Li-Yorke sensitive implies that the system \((\Lambda^\circ, d, F|_{\Lambda^\circ})\) is Li-Yorke sensitive. Inspired by this, the following conclusion can be drawing.

**Theorem 6.** If \(f\) is densely Li-Yorke sensitive, then the system \((\Lambda^\circ, d, F|_{\Lambda^\circ})\) is densely Li-Yorke sensitive.

**Proof.** Since \(f\) is densely Li-Yorke sensitive, then for any \(a \in Q_\varepsilon(f)\) and any \(\varepsilon > 0\). Then there exists a \(b \in B(a, \varepsilon)\) such that \((a, b) \in LY_{\varepsilon}(f, \delta)\). Take \(x^* = \{x_n = a\}_{n=-\infty}^\infty\), \(y^* = \{y_n = b\}_{n=-\infty}^\infty\). One has that

\[
\limsup_{n \to \infty} d(F^n(x^*), F^n(y^*)) = \limsup_{n \to \infty} d(f^n(a), f^n(b)) > \delta
\]

and

\[
\liminf_{n \to \infty} d(F^n(x^*), F^n(y^*)) = \liminf_{n \to \infty} d(f^n(a), f^n(b)) = 0
\]

Thus there is an \(x^* \in Q_\delta(F)\).

Any fixed \(x \in \Delta^\circ\), write \(x = (\cdots, x_{m-1}, x_m, x_{m+1}, \cdots)\), where \(x_{m,p} = x_{m,p+1}, p \in \mathbb{Z}\). Because \(f : I \mapsto I\) is densely Li-Yorke sensitive, then for any \(\varepsilon > 0\) and the above \(x_{m,0}\), \(B(x_{m,0}, \varepsilon) \cap Q_\delta(f) \neq \emptyset\). Take \(a \in B(x_{m,0}, \varepsilon) \cap Q_\delta(f)\), then

\[d(x, x^*) = \sup \{|x_{m,p} - a|\} = |x_{m,p} - a| < \varepsilon.\]

So \(x^* \in B(x, \varepsilon)\). This suggests that \(Q_\delta(F) = \Delta^\circ\).

So, the system \((\Lambda^\circ, d, F|_{\Lambda^\circ})\) is densely Li-Yorke sensitive.

According to Proposition 3 and Theorem 6 the following is right.

**Corollary 4.** If \(f\) is dense Li-Yorke sensitivity, then the system \((\Lambda^\circ, d, F|_{\Lambda^\circ})\) it is Topological mixing (or its topological entropy is positive).

### 4. Conclusion

Inspired by the literature [23], this paper further studies the chaoticity of coupled map lattices. Some sufficient conditions of sensitivity, accessibility and transitivity are obtained. However, the study of coupled map lattices is still a hot topic. Based on the conclusions of this paper and others, one can consider some questions, such as the form of CMLs, the measurement of CMLs, and discuss the chaos of CMLs in other systems, which are worthy of studying.

**Conflicts of Interest**

The authors declare no conflicts of interest regarding the publication of this paper.

DOI: 10.4236/jamp.2020.86075
References


Criterion for the Emergence of Meta-Stable States in Traffic Systems

Liuhua Zhu1,2

1School of Physics and Telecommunication Engineering, Yulin Normal University, Yulin, China
2Guangxi Colleges and Universities Key Laboratory of Complex System Optimization and Big Data Processing, Yulin, China

Email: zhuliuhua421@sina.com

Abstract

The measurements on actual traffic have revealed the existence of meta-stable states with high flow. Such nonlinear phenomena have not been observed in the classic Nagel-Schreckenberg traffic flow model. Here we just add a constraint to the classic model by introducing a velocity-dependent randomization. Two typical randomization strategies are adopted in this paper. It is shown that the Matthew effect is a necessary condition to induce traffic meta-stable states, thus shedding a light on the prerequisites for the emergence of hysteresis loop in the fundamental diagrams.

Keywords

Traffic Flow, Cellular Automaton, Matthew Effect, Hysteresis Loop

1. Introduction

In the past decades, a lot of attention has been devoted to the study of traffic flow. Since the seminal work of Nagel and Schreckenberg in the early 1990s [1], a number of cellular automata models describing traffic flow have been proposed in order to consider the real traffic scenes such as road blocks, intersections, adverse weather conditions, and so on. These cellular automata models can be used in real-time simulation very effectively, and they successfully replicate many nonlinear phenomena which are consistent with the actual traffic.

In recent years, the cellular automata models have been extended to investigate the meta-stable states in traffic systems [2]-[8]. The meta-stable states are usually represented as a hysteresis loop in the fundamental diagram. The latter is a consequence of phase separation within a certain density range. The slow-to-start rule can reduce the outflow of the traffic congestion area and keep
the inflow unchanged, so the congestion area gradually expands, resulting in phase separation [9] [10]. The slow-to-start rule was once considered as a necessary condition to induce traffic meta-stable states.

The study of meta-stable states is of great practical significance. On the smooth road, using cruise constant velocity can keep the vehicles running at a constant velocity and reduce fuel consumption. Based on this fact, D. Chowdhury et al. proposed a cruise-control limit model, which successfully reproduced the meta-stable states in traffic system [11]. This meta-stable state strategy has been applied to Lincoln and Holland tunnels in New York City to reduce frequent traffic congestions. Both the slow-to-start and the cruise-control limit can keep the vehicles in a meta-stable state with high flow, rather than transition to a congestion state, which effectively relieves the traffic pressure.

2. Model

For the sake of completeness, let us briefly recall the evolution rules of the classic Nagel-Schreckenberg model. This set of rules describes the principles that the vehicles must follow when driving on a one-dimensional ring road. The road is divided into a series of cells. Each cell is either empty or occupied by just one vehicle with a discrete velocity \( v_i(t) \in [0, v_{\text{max}}] \). Here \( v_{\text{max}} \) is the velocity limit of the vehicles. The density \( \rho \) of the road is defined as the ratio of the number \( N \) of vehicles to the length \( L \) of the road, i.e., \( \rho = N/L \).

The configurations of the vehicles are updated in parallel according to the following four rules. R1: Acceleration, \( v_i(t+1/3) = \min\left(v_i(t) + 1, v_{\text{max}}\right) \); R2: Braking, \( v_i(t+2/3) = \min\left(v_i(t+1/3), d_i(t)\right) \); R3: Randomization (with probability \( p \)), \( v_i(t+1) = \max\left(v_i(t+2/3) - 1, 0\right) \); R4: Location updating, \( x_i(t+1) = x_i(t) + v_i(t+1) \). Here \( x_i(t) \) and \( v_i(t) \) are the position and the velocity of the \( i \)th vehicle at time \( t \). The parameter \( d_i(t) \) is the empty cells between vehicle \( i \) and the nearest neighbor vehicle \( i+1 \) in front of it.

Although these rules seem simple, they can simulate some complex traffic phenomena such as the free flow and ghostly congestion. Rule 1 characterizes a driver’s trait to drive as fast as possible without exceeding the maximum velocity limit. Rule 2 is designed to avoid collisions between vehicles. Rule 3 requires drivers to slow down randomly and change their visual angle, so as to effectively alleviate visual fatigue. The randomization is crucial for the spontaneous emergence of traffic jams.

In the classic Nagel-Schreckenberg traffic flow model, the meta-stable states will not occur, because the indiscriminate randomization makes the homogeneous structure of the system difficult to maintain. In this paper, we do not modify the evolution rules of Nagel-Schreckenberg model, but add a specific function to control the random slowing probability of each vehicle. The determination of random slowing probability is placed before the acceleration step. R0: \( p_i(t) = p_o g\left(v_i(t)\right) \). Here \( g(x) \) is a bounded sine or cosine function.
3. Main Results

For simplicity, only one type of vehicles is considered in this paper and therefore the same maximum velocity $v_{\text{max}} = 5$ is applied to all vehicles. For a realistic description of highway traffic, the length of a cell is set to 7.5 m, which is interpreted as the length of one vehicle plus the average gap between two adjacent vehicles in a jam.

First, we show the differences of fundamental diagrams for three different control strategies, as shown in Figure 1(a). For the sine or cosine control strategy, its peak value $p_0$ is set to 0.35. As a contrast, the middle curve in Figure 1(a) is controlled by a constant control strategy, i.e., $p = 0.25$, because the effective value of the sine or cosine wave is 0.707 times of its peak value.

The cosine control strategy mainly limits the low-velocity vehicles, but has no limit to the vehicles with maximum velocity. At the low density, only vehicles with the cosine control strategy can drive at maximum velocity, as shown in Figure 1(b).

Under the control of the sinusoidal law, the vehicles traveling at maximum velocity are most restricted, while the stationary vehicle can start at any time as long as there is free space in front of it. Therefore, in the middle and high density areas, the traffic flow under the control of sine function is significantly higher than the other two strategies, as shown in Figure 1(a).

On one hand, the cosine control strategy requires that the vehicles running at the maximum velocity do not slow down randomly, which corresponds to the phenomenon of “rich get richer” in life. On the other hand, it demands the stationary vehicles to slow start, which is equivalent to the phenomenon of “poor get poorer” in life. In this sense, the cosine control strategy is equivalent to Matthew effects, which is conducive to the separation of phases and the emergence of meta-stable states.

In Figure 2, $p_0$ is set to 0.3 because the maximum flows obtained from the two different control strategies are 1872 vehicles per hour and 2160 vehicles per hour respectively, which is consistent with the current single lane capacity of expressways. In Figure 2(a), we obtain two different fundamental diagrams starting with two different initial configurations, labeled as homogeneous and random initializations, respectively. The upper branch corresponds to the calculation starting with a homogeneous initialization, while the lower branch is obtained starting from a random configuration. In certain regions, the traffic flow is no longer a single value function of density, which is known as meta-stable regions. It is obvious that Matthew effects induce meta-stable states in traffic systems. However, for the sine control strategy the two curves in Figure 2(b) show a typical second-order phase transition with a perfect match between homogeneous and random initializations.

4. Minor Perturbation

In this section, we show the effect of a minor perturbation on the traffic system.
Figure 1. Flow-density and average velocity-density diagrams under three different control strategies (a) The fundamental diagrams for different randomization functions. As a contrast, the middle curve adopts a constant control strategy. (b) Under the same conditions, the evolution of average velocity with density.

Figure 2. The fundamental diagrams obtained from two different control strategies. (a) the cosine control strategy; (b) the sine control strategy. The hollow dots are obtained from a homogeneous initial state, while the solid dots from a random initial configuration. The parameter $p_0$ is set to 0.3.

Figure 3(a) shows that when there is no perturbation the queue of vehicles is orderly and runs at the maximum velocity. A minor perturbation leads to the spontaneous emergence and cascading effects of traffic congestions. In this uniform and dense region, the gap between adjacent vehicles just meets the demand of maximum velocity driving. In the region all following vehicles have to slow down to avoid rear end collision once a vehicle in front is disturbed. After being disturbed, the orderly pace of the vehicles is broken. This kind of damage is irreversible, and cannot be repaired by itself with the passage of time, resulting in a sharp decline in traffic flow, as shown in Figure 3(b). According to the control function of randomization, the slower the velocity is, the greater the random slowing probability is. The chain reaction of following vehicles eventually leads to the expansion of traffic congestion.

For the random initial configuration, at the same traffic density there are already traffic congestions in the lane. After being disturbed, the chain reaction of subsequent vehicles leads to new traffic congestions. The new traffic congestions here relieve the traffic pressure elsewhere, so the width of the early traffic congestions in the system becomes narrow as shown in Figure 4(a). This mechanism of “as one falls, another rises” leads to the fluctuation of traffic flow, as shown in Figure 4(b).

For the sine control strategy, at the same traffic density some small traffic
congestions are randomly scattered on the lane, as shown in Figure 5(a). According to the control function of randomization, the slower the velocity is, the smaller the random slowing probability is. Static vehicles start fast, high-velocity vehicles slow down fast, so traffic congestions happen and disappear from time to time. Figure 5(a) is also disturbed after 100 evolution time steps. However, we can not even find the disturbed position accurately. This kind of spatial-temporal distribution, which has little difference before and after the perturbation, will not cause the great change of traffic flow naturally, as shown in Figure 5(b). Compared with Figure 4(b), the traffic flow here is more stable and the anti-interference ability is stronger.

5. Conclusion

In this paper, the main factors and limiting conditions of meta-stable and hysteretic phenomena are explored. Although the Nagel-Schreckenberg model itself cannot reflect the meta-stable and hysteretic phenomena found in real traffic, it can capture more complex traffic phenomena with a little modification. For any traffic flow model, it is a challenge to describe the possibility of hysteresis loop. Our study generalizes some of the previous results and extends the possibility of

Figure 3. Spatiotemporal evolutions under a minor disturbance and corresponding change of flow. (a) The spatial-temporal diagrams at $\rho = 0.15$ for the upper branch of the cosine control strategy. The vehicles are moving from left to right. A vehicle is plotted by a black dot. The time axis is vertical down. After the system reaches a steady state, a randomly selected vehicle is forbidden to move within 5 time steps. (b) Under the same conditions, the evolution of traffic flow with time.

Figure 4. Spatiotemporal evolutions under a minor disturbance and corresponding change of flow. (a) The spatial-temporal diagrams at $\rho = 0.15$ for the lower branch of the cosine control strategy. The vehicles are moving from left to right. A vehicle is plotted by a black dot. The time axis is vertical down. After the system reaches a steady state, a randomly selected vehicle is forbidden to move within 5 time steps. (b) Under the same conditions, the evolution of traffic flow with time.
Figure 5. Spatiotemporal evolutions under a minor disturbance and corresponding change of flow. (a) The spatial-temporal diagrams at $\rho = 0.15$ for the sine control strategy. The vehicles are moving from left to right. A vehicle is plotted by a black dot. The time axis is vertical down. After the system reaches a steady state, a randomly selected vehicle is forbidden to move within 5 time steps. (b) Under the same conditions, the evolution of traffic flow with time.

meta-stable states in traffic systems to a general criterion. Only when the Matthew effect is embedded in the evolution rules, the meta-stable state will appear as scheduled. Our results will pave the way for the research with the same dynamic background in the real traffic systems.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grant Nos. 61563054 and 10562020, the Natural Science Foundation of Guangxi under Grant No. 2019GXNSFAA245023, the open fund of Guangxi Colleges and Universities Key Laboratory of Complex System Optimization and Big Data Processing under Grant No. 2015CSOBDP0101, Initiation Fund of Doctoral Research of Yulin Normal University under Grant No. G20150003.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References


An Efficient Projected Gradient Method for Convex Constrained Monotone Equations with Applications in Compressive Sensing

Yaping Hu, Yujie Wang

College of Science, Tianjin University of Science and Technology, Tianjin, China
Email: huyaping@tust.edu.cn

Abstract
In this paper, a modified Polak-Ribière-Polyak conjugate gradient projection method is proposed for solving large scale nonlinear convex constrained monotone equations based on the projection method of Solodov and Svaiter. The obtained method has low-complexity property and converges globally. Furthermore, this method has also been extended to solve the sparse signal reconstruction in compressive sensing. Numerical experiments illustrate the efficiency of the given method and show that such non-monotone method is suitable for some large scale problems.

Keywords
Projection Method, Monotone Equations, Conjugate Gradient Method, Compressive Sensing

1. Introduction
This paper is dedicated to solving the following nonlinear convex constrained monotone equations:

\[ F(x) = 0, \quad x \in \Omega, \] (1)

where \( F : \mathbb{R}^n \rightarrow \mathbb{R}^n \) is a continuous nonlinear mapping and the feasible region \( \Omega \subset \mathbb{R}^n \) is a nonempty closed convex set, namely, \( \Omega = \{ x \in \mathbb{R}^n : l \leq x \leq u \} \). Monotone means that

\[ \langle F(x) - F(y), x - y \rangle \geq 0, \quad \forall x, y \in \mathbb{R}^n, \] (2)

where the \( \langle \cdot , \cdot \rangle \) denotes the inner product of vectors. The problems (1) emerges in many fields such as economic equilibrium problems [1], chemical equilibrium...
systems [2] and the power flow equations [3]. Based on the work of Solodov and Svaiter [4], Wang et al. [5] proposed a projection type method to solve Equation (1). The obtained method in [5] possesses global convergence property without any regularity assumptions. Nevertheless the method needs to solve a linear equation at each iteration. To avoid solving the linear equation and improving the effectiveness, some projected conjugate gradient methods [6] [7] [8] [9] are studied based on the projection technique of Solodov and Svaiter [4]. The numerical results gained in [6] [7] [8] [9] indicate that the projected conjugate gradient type methods for solving problem (1) are indeed efficient and promising. In this paper, by combining the well-known Polak-Ribière-Polyak [10] [11] method with the projection technique of Solodov and Svaiter [4], a conjugate gradient projected method with fast convergent property is proposed for the nonlinear monotone equations with convex constraints. Under some mild conditions, the global convergence results are established for the given method. The obtained method possesses the following three beneficial properties: 1) The search direction satisfies the sufficient descent condition, 2) The global convergence is independent of any merit function, and 3) It is derivative-free method and is effective for large scale nonlinear convex constrained monotone equations (with a maximum dimension of 100,000). Furthermore, the obtained method is extended to solve the \( l_1 \)-norm problem by reformulating it as non-smooth monotone equations.

In Section 2, the modified PRP-type conjugate gradient projected method is proposed, and some preliminary properties are studied. The global convergence results are established in Section 3. The numerical experiments, and the applications of the obtained method for \( l_1 \)-norm regularized compressive sensing problems are discussed in Section 4. Finally, we have a conclusion section.

2. The Proposed Method and Corresponding Algorithm

We firstly introduce the definition of the projection operator \( P_\Omega [\cdot] \) which is defined as the mapping from \( R^n \) to \( \Omega \),

\[
P_\Omega [x] = \operatorname{arg\ min} \{\|y - x\| \mid y \in \Omega\}, \quad \forall x \in R^n,
\]

where \( \|\cdot\| \) denotes the Euclidean norm of vectors, \( \Omega \) is a nonempty closed convex subset of \( R^n \).

The projection operator is non-expansive, namely, for any \( x, y \in R^n \), the following condition holds

\[
\|P_\Omega [y] - P_\Omega [x]\| \leq \|x - y\|.
\]  (3)

Let’s review the Polak-Ribière-Polyak [10] [11] conjugate gradient method briefly. The PRP method is firstly designed for solving the unconstrained optimization problem:

\[
\min \left\{ f(x) \mid x \in R^n \right\},
\]  (4)
where \( f : \mathbb{R}^n \to \mathbb{R} \) is continuously differentiable. It generates the iteration sequence \( \{x_k\} \) in the form
\[
x_{k+1} = x_k + \alpha_k d_k,
\]
where \( x_k \) is the current iteration point, \( \alpha_k > 0 \) is a step-length, and \( d_k \) is the search direction given by
\[
d_k = \begin{cases} -g_k + \beta_{k-1}^{PRP} d_{k-1}, & \text{if } k > 0, \\ -g_k, & \text{if } k = 0, \end{cases}
\]
where \( \beta_{k-1}^{PRP} = \frac{g_k^T y_{k-1}}{\|g_k\|^2}, \ y_{k-1} = g_k - g_{k-1}. \)

Combining the projected technique of Solodov and Svaiter \[4\] with the PRP method formed by Equation (5) and Equation (6), the following modified PRP formula is defined given in this paper
\[
d_k = \begin{cases} -g_k + \frac{g_k^T y_{k-1} d_{k-1} - d_{k-1}^T g_k y_{k-1}}{\max \{2 \gamma \|d_{k-1}\|y_{k-1}^T d_{k-1} - g_k^T y_{k-1}\} \|g_k\|^2}, & \text{if } k > 0, \\ -g_k, & \text{if } k = 0, \end{cases}
\]
where \( y_{k-1} = g_k - g_{k-1} \) and \( \gamma > 0 \) is a constant.

It is show be noted that the proposed direction formula Equation (7) reduces to PRP formula if the exact line search is used. Furthermore, the sufficient descent condition automatically holds for all \( k \), since \( d_k^T g(x_k) = -\|g(x_k)\|^2 \). There are some conjugate gradient methods with similar idea concerning Equation (7) have been studied in the papers \[12\]-[19].

The corresponding modified PRP conjugate gradient projection algorithm for solving problem (1) starts as follows.

**Algorithm 1:**

**Step 0** Choose any initial point \( x_0 \in \Omega \), and select constants \( \rho \in (0,1), \gamma > 0, \sigma > 0, \zeta > 0, \epsilon \in (0,1) \) and \( d_0 = -F(x_0). \) Let \( k := 0 \).

**Step 1** If \( \|F(x_k)\| \leq \epsilon \), stop. Otherwise compute search direction \( d_k \) by Equation (7) with \( g_k \) and \( g_{k-1} \) replaced by \( F_k \) and \( F_{k-1} \), respectively.

**Step 2** Let \( z_k = x_k + \alpha_k d_k \), where \( \alpha_k = \max \{\gamma \delta^i \mid i = 0, 1, \ldots\} \) such that
\[
-\langle F(x_k + \alpha_k d_k), d_k \rangle \geq \sigma \alpha_k \|d_k\|^2.
\]

**Step 3** If \( \|F(z_k)\| \leq \epsilon \), stop and let \( x_{k+1} = z_k \). Otherwise compute the next iteration by
\[
x_{k+1} = P_\Omega \left[ x_k - \beta_k F(z_k) \right],
\]
where
\[
\beta_k = \frac{\langle F(z_k), x_k - z_k \rangle}{\|F(z_k)\|^2}
\]

**Step 4** Let \( k := k + 1 \), and go to Step 1.

**Remark 1:** In the algorithm 1, the step size \( \alpha_k \) given by Equation (8) satisfies
\[ \langle F(z_k), x_k - z_k \rangle > 0, \]
where \( z_k = x_k + \alpha_k d_k \), \( d_k \) is the search direction. Moreover, for any \( x^* \) such that \( F(x^*) = 0 \),
\[ \langle F(z_k), x^* - z_k \rangle \leq 0. \]
comes from the monotonicity property of \( F(x) \). This means that the hyperplane
\[ H_k = \{ x \in \mathbb{R}^n | \langle F(z_k), x - z_k \rangle = 0 \} \]
strictly separates the current point \( x_k \) from the solution set of the problem. The above facts and Step 3 indicate that the next iteration \( x_{k+1} \) is computed by projecting \( x_k \) onto the intersection of the feasible set \( \Omega \) with the halfspace \( H_k \).

3. Convergence Analysis

In this section, we are going to discuss the convergence property of the given method. Before that, there are some basic assumptions on problem (1) needs to be given.

**Assumption 1:** The mapping \( F \) is Lipschitz continuous with constant \( L > 0 \) in a set \( \Omega \), written \( F \in \text{Lip}(\Omega) \), for every \( x, y \in \Omega \),
\[ \| F(x) - F(y) \| \leq L \| x - y \|. \]  (11)

**Assumption 2:** The solution set of the problem (1), denoted by \( S \), is nonempty convex.

For conjugate gradient method, the sufficient descent property is essential in the convergence analysis, the following lemma shows that the search direction \( \{d_k\} \) generated by Algorithm 1 satisfies the sufficient descent condition independent of line search.

**Lemma 1:** Let the sequence \( \{x_k\} \) and \( \{d_k\} \) be generated by Algorithm 1. Then, for all \( k \geq 0 \),
\[ F(x_k)^T d_k = -\| F(x_k) \|^2; \]  (12)
and
\[ \|d_k\| \leq \left(1 + \frac{1}{\gamma}\right) \| F(x_k) \|. \]  (13)

**Proof:** For \( k = 0 \), Equation (12) and Equation (13) follows from the direct application of \( d_0 = -g(x_0) \). For \( k \geq 1 \), using Equation (7), the definition of the search direction \( d_{k+1} \), it follows that
\[ d_{k+1}^T F_{k+1} = -\| F_{k+1} \|^2 + \left[ \frac{F_{k+1}^T y_k d_k - d_{k+1}^T F_{k+1} y_k}{\max \left\{2\gamma \|d_k\|, \|d_k^T y_k\|, \|F_{k+1}\|^2\right\}} \right]^T \]
\[ F_{k+1} = -\| F_{k+1} \|^2, \]
similarly,
\[
\|d_{k+1}\| = \|F_{k+1} + \frac{F_{k+1}^T y_k}{\max \left\{ 2\gamma \|d_k\| y_k, \|d_k\| F_{k+1} \|y_k\| \right\}} \| \\
\leq \|F_{k+1}\| + \frac{\|F_{k+1}\| \|y_k\| \|d_k\| + \|d_k\| \|F_{k+1}\| \|y_k\|}{\max \left\{ 2\gamma \|d_k\| \|y_k\|, \|d_k\| \|F_{k+1}\| \|y_k\| \right\}} \\
\leq \left( 1 + \frac{1}{\gamma} \right) \|F_{k+1}\|,
\]

where the last inequality follows from the fact
\[
\max \left\{ 2\gamma \|d_k\| \|y_k\|, \|d_k\| \|F_{k+1}\| \|y_k\| \right\} \geq 2\gamma \|d_k\| \|y_k\|.
\]

In the remaining part of this paper, we assume that \( F_k \neq 0 \) for all \( \forall k \geq 0 \), otherwise, the solution of the problem (1) has been found.

**Lemma 2:** Let the sequence \( \{x_k\} \) and \( \{z_k\} \) be generated by Algorithm 1. Suppose that the Assumption 1 holds. Then there exists a positive number \( \alpha_k \) satisfying Equation (8) for all \( k \geq 0 \).

**Proof:** The line search ensure that if \( \alpha_k \xi \neq 0 \), then \( \alpha_k = \rho^{-1} \alpha_k \) does not satisfy Equation (8), namely,
\[
-F(z'_k), d_k < \sigma \alpha_k \|d_k\|^2,
\]
where \( z'_k = x_k + \alpha_k d_k \). From Equation (12) and Assumption 1 we have
\[
\|F_i\|^2 = -\|F_i, d_k\| = \|F(z'_k) - F(x_k), d_k\| - \|F(z'_k), d_k\| \\
\leq \lambda \alpha_k \|d_k\|^2 + \sigma \alpha_k \|d_k\|^2 \leq \rho^{-1} \alpha_k (L + \sigma) \|d_k\|^2
\]
which means that
\[
\alpha_k \geq \min \left\{ \xi, \frac{\rho}{\lambda + \sigma} \|F_i\|^2 \right\} \quad \text{(14)}
\]

The above result Equation (14) shows that the line search procedure Equation (8) always terminates in a finite number of steps.

**Lemma 3:** Let sequences \( \{x_k\} \) and \( \{z_k\} \) be generated by Algorithm 1. Suppose that Assumptions 1 and 2 hold. Then both \( \{x_k\} \) and \( \{z_k\} \) are bounded. Moreover, we have
\[
\lim_{k \to \infty} \|x_k - z_k\| = 0, \quad \text{(15)}
\]
and
\[
\lim_{k \to \infty} \|x_{k+1} - x_k\| = 0. \quad \text{(16)}
\]

Particularly, Equation (15) implies that
\[
\lim_{k \to \infty} \alpha_k \|d_k\| = 0. \quad \text{(17)}
\]

**Proof:** \( x^* \in S \) denotes any arbitrary solution of the problem (1). The monotonicity of \( F \) and the line search Equation (8) deduce
\[ \langle F(z_k), x_k - x^* \rangle \geq \langle F(z_k), x_k - z_k \rangle \geq \sigma \alpha_k^2 \| d_k \|^2 \geq 0. \]  

Equation (3), Equation (9) and Equation (18) imply
\[
\| x_{k+1} - x^* \|^2 = \| x_k - \beta_k F(z_k) - x^* \|^2 - 2 \beta_k \langle F(z_k), x_k - x^* \rangle + \beta_k^2 \| F(z_k) \|^2 \\
\leq \| x_k - x^* \|^2 - 2 \beta_k \langle F(z_k), x_k - z_k \rangle + \beta_k^2 \| F(z_k) \|^2 \\
\leq \| x_k - x^* \|^2 - \frac{\langle F(z_k), x_k - z_k \rangle^2}{\| F(z_k) \|^2} \\
\leq \| x_k - x^* \|^2 - \sigma^2 \| x_k - z_k \|^2 \| F(z_k) \|^2.
\] 

Since the sequence \( \{ \| x_k - x^* \| \} \) is decreasing and convergent, the sequence \( \{ x_k \} \) is bounded. Equation (19) shows that \( \| x_k - x^* \| \leq \| x_0 - x^* \| \) for all \( k \). Then, by Assumption 1, we have
\[
\| F(x_k) \| = \| F(x_k) - F(x^*) \| \leq L \| x_k - x^* \| \leq L \| x_0 - x^* \|. 
\] 

Let \( M_1 = L \| x_0 - x^* \|. \)
\[
\| F(x_k) \| \leq M_1, \forall k \geq 0. 
\] 

From the Cauchy-Schwarz inequality, the line search Equation (8), the monotonicity of \( F \) and Equation (18), it follows that
\[
0 < \sigma \| x_k - z_k \|^2 \leq \langle F(z_k), x_k - z_k \rangle \leq \langle F(x_k), x_k - z_k \rangle \leq \| F(x_k) \| \| x_k - z_k \| \\
\sigma \| x_k - z_k \|^2 \leq \| F(x_k) \| \leq M_1, 
\] 

which shows that the sequence \( \{ z_k \} \) is bounded. Furthermore, the sequence \( \{ \| F_k - x^* \| \} \) is also bounded, there exists \( M_2 > 0, \ k_0 \geq 0 \), such that
\[
\| z_k - x^* \| \leq M_2, \forall k \geq k_0. 
\] 

Based on Equation (23) and Assumption 1 it follows
\[
\| F(z_k) \| = \| F(z_k) - F(x^*) \| \leq L \| z_k - x^* \| \leq LM_2.
\] 

Substituting the above relationship into Equation (19), it deduces
\[
\frac{\sigma^2}{(LM_2)^2} \sum_{k=0}^{\infty} \| x_k - z_k \|^2 \leq \sum_{k=0}^{\infty} \left( \| x_k - x^* \|^2 - \| x_{k+1} - x^* \|^2 \right) < \infty,
\] 

which implies
\[
\lim_{k \to \infty} \| x_k - z_k \| = 0.
\]

From the definition of \( z_k \) and Equation (15), it holds that
\[
\lim_{k \to \infty} \| d_k \| = 0.
\]

Combining the definition of \( \beta_k \), Equation (3), and the Cauchy-Schwarz in-
equality, we have
\[ \|x_{k+1} - x_k\| = \|P_{T_k} [x_k - \beta_k F(z_k)] - x_k\| \]
\[ \leq \|x_k - \beta_k F(z_k) - x_k\| \]
\[ = \frac{\langle F(z_k), x_k - z_k \rangle}{\|F(z_k)\|} \]
\[ \leq \|x_k - z_k\| \]
which together with Equation (15), proves Equation (16).

**Theorem 1:** Let sequences \( \{x_k\} \) and \( \{z_k\} \) be generated by Algorithm 1. Suppose that Assumptions 1 and 2 hold. Then
\[ \liminf_{k \to \infty} \|F_k\| = 0. \]  
(26)

**Proof:** We prove this Theorem by contradiction. Assume that Equation (26) does not hold, namely, there exists \( \varepsilon > 0 \) such that
\[ \|F_k\| \leq \varepsilon, \quad \forall k \geq 0. \]  
(27)

From Equation (12) and Equation (27),
\[ \|d_k\| \leq \|d_k + F_k\| \]
\[ = \|d_k + F_k\| - 2\langle d_k + F_k, F_k \rangle + \|F_k\|^2 \]
\[ \geq -2\langle d_k, F_k \rangle - \|F_k\|^2 \]
\[ = \|F_k\|^2, \]
which implies
\[ \|d_k\| \geq \varepsilon, \quad \forall k \geq 0. \]  
(28)

On the other hand, Equation (13), Equation (21) and the definition of \( d_k \) deduce
\[ \|d_k\| \leq \left(1 + \frac{1}{\gamma}\right)\|F_k\| \leq \left(1 + \frac{1}{\gamma}\right)M_1, \quad \forall k \geq 0. \]

Finally, from Equation (14), Equation (27) and Equation (28),
\[ \alpha_k \|d_k\| \geq \min \left\{ \xi, \frac{\rho \|F_k\|^2}{L + \sigma \|d_k\|^2} \right\} \|d_k\| \]
\[ \geq \min \left\{ \xi \varepsilon, \frac{\rho \varepsilon^2}{(L + \sigma)(1 + \gamma^{-1})M_1} \right\} \]
which contradicts with Equation (17). Thus, Equation (26) holds.

### 4. Numerical Experiments

The numerical performances of the proposed Algorithm 1 for large scale non-linear convex constrained monotone equations with various dimensions and different initial points are studied in this section. Furthermore, the given Algorithm 1 is extended to solve the \( l_1 \)-norm regularized problems which de-
code a sparse signal in compressive sensing. The algorithm is coded in MATLAB R2015a and run on a PC with Core i5 CPU and 4 GB memory.

4.1. Experiments on Nonlinear Convex Constrained Monotone Equations

The testing problems are listed as follows.

**Problem 1.** (Wang et al. [5]) The elements of \( F(x) \) are given by
\[
F_i(x) = e^x - 1, \quad i = 1, 2, 3, \ldots, n.
\]
and \( \Omega = R^n \).

**Problem 2.** The example is taken from [7]. The elements of \( F(x) \) are given by
\[
F_i(x) = 2x_i - \sin(x_i), \quad i = 1, 2, 3, \ldots, n.
\]
and \( \Omega = R^n \).

**Problem 3.** The example is taken from [9].
\[
g_i(x) = x_i - e^{-\cos\left(\sum_{k=1}^{i} x_k / n + 1\right)},
g_i(x) = x_i - e^{-\cos\left(\sum_{k=1}^{i} x_k / n + 1\right)}, \quad i = 2, 3, \ldots, n-1,
g_n(x) = x_n - e^{-\cos\left(\sum_{k=1}^{n} x_k / n + 1\right)}.
\]
and \( \Omega = R^n \).

**Problem 4.** The example is taken from [20].
\[
F_i(x) = x_i - \sin(|x_i - 1|), \quad i = 1, 2, 3, \ldots, n.
\]
and \( \Omega = \{x \in R^n | \sum_{i=1}^{n} x_i \leq n, x_i \geq -1, i = 1, 2, \ldots, n\} \).

For convenience, MPRP denotes the proposed Algorithm 1. We compare the MPRP method with CGD method [8] on problems 1-4. For both methods, set \( \xi = 1, \quad \rho = 0.4, \quad \sigma = 10^{-4} \). In order to evaluate the efficiency and the robustness of both methods, we test the Problems 1-4 with various dimensions
\[
n = 10000, 50000, 100000 \quad \text{and different initial points: } x_i = \left[0.5, \ldots, \frac{1}{n}\right],
x_1 = \frac{1}{n} \text{ones}(n,1), \quad x_3 = \text{ones}(n,1), \quad x_4 = 2 \text{ones}(n,1), \quad x_5 = \text{rand}(n,1),\]
where \( \text{ones}(n,1) \) returns a \( n \)-by-1 array of ones and \( \text{rand}(n,1) \) returns a \( n \)-by-1 array of random values in MATLAB.

Numerical results are shown in Tables 1-4, in which Init (Dim), NI and NF denote initial points (dimension), the number of iterations and the number of function evaluations respectively. \( \|F(x)\| \) is the final Euclidean norm of the function values, and CPU-time in seconds.

Tables 1-4 indicate that the dimension of the problem has little effect on the number of iterations of the algorithm. However, the computing time is relatively large in high dimension cases. Moreover, we can see from the results of Tables...
that Algorithm 1 is more competitive than CGD algorithm as Algorithm 1 can get the solution of all the test data at a smaller number of iterations and smaller CPU time. So the results of Tables 1-4 show that our method is very efficient.

The numerical performances of the both methods are also evaluated by using the performance profile tool of tool of Dolan and Moré [21]. Figure 1 shows the performance of two methods, it is obviously that the proposed MPRP method is more efficient and robust than CGD method.

**Table 1.** Numerical results for MPRP/CGD on problem 1.

<table>
<thead>
<tr>
<th></th>
<th>MPRP</th>
<th></th>
<th>CGD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Init (Dim)</td>
<td>NI/ NF/</td>
<td></td>
<td>f(x)</td>
</tr>
<tr>
<td>x1(10000)</td>
<td>11/25/3.84830e-006</td>
<td>0.13</td>
<td>16/185/5.84301e-006</td>
</tr>
<tr>
<td>x1(10000)</td>
<td>5/11/9.76376e-006</td>
<td>0.08</td>
<td>5/11/9.59824e-006</td>
</tr>
<tr>
<td>x1(10000)</td>
<td>11/24/2.98167e-006</td>
<td>0.15</td>
<td>10/21/4.63815e-006</td>
</tr>
<tr>
<td>x1(10000)</td>
<td>11/27/6.30895e-006</td>
<td>0.14</td>
<td>12/28/5.36760e-006</td>
</tr>
<tr>
<td>x1(10000)</td>
<td>14/31/8.16510e-006</td>
<td>0.15</td>
<td>30/61/9.55151e-006</td>
</tr>
<tr>
<td>x1(50000)</td>
<td>11/25/3.84897e-006</td>
<td>0.42</td>
<td>19/261/5.42831e-006</td>
</tr>
<tr>
<td>x1(50000)</td>
<td>5/11/4.36715e-006</td>
<td>0.21</td>
<td>5/11/4.29309e-006</td>
</tr>
<tr>
<td>x1(50000)</td>
<td>11/24/6.66722e-006</td>
<td>0.45</td>
<td>11/23/2.07424e-006</td>
</tr>
<tr>
<td>x1(50000)</td>
<td>12/29/3.52681e-006</td>
<td>0.56</td>
<td>13/30/2.40047e-006</td>
</tr>
<tr>
<td>x1(50000)</td>
<td>15/33/5.67816e-006</td>
<td>0.63</td>
<td>32/65/7.69276e-006</td>
</tr>
<tr>
<td>x1(100000)</td>
<td>11/25/3.84906e-006</td>
<td>0.78</td>
<td>16/192/7.19276e-006</td>
</tr>
<tr>
<td>x1(100000)</td>
<td>5/11/3.08810e-006</td>
<td>0.39</td>
<td>5/11/3.03573e-006</td>
</tr>
<tr>
<td>x1(100000)</td>
<td>11/24/9.42888e-006</td>
<td>0.86</td>
<td>11/23/2.93342e-006</td>
</tr>
<tr>
<td>x1(100000)</td>
<td>12/29/4.98767e-006</td>
<td>1.03</td>
<td>13/30/3.39477e-006</td>
</tr>
<tr>
<td>x1(100000)</td>
<td>15/33/8.06362e-006</td>
<td>1.22</td>
<td>33/67/6.52318e-006</td>
</tr>
</tbody>
</table>

**Table 2.** Numerical results for MPRP/CGD on problem 2.

<table>
<thead>
<tr>
<th></th>
<th>MPRP</th>
<th></th>
<th>CGD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Init (Dim)</td>
<td>NI/ NF/</td>
<td></td>
<td>f(x)</td>
</tr>
<tr>
<td>x1(10000)</td>
<td>10/21/5.34065e-006</td>
<td>0.09</td>
<td>17/141/7.30891e-006</td>
</tr>
<tr>
<td>x1(10000)</td>
<td>5/11/3.20000e-006</td>
<td>0.07</td>
<td>5/11/9.60000e-006</td>
</tr>
<tr>
<td>x1(10000)</td>
<td>10/21/3.73273e-006</td>
<td>0.09</td>
<td>11/23/4.43164e-006</td>
</tr>
<tr>
<td>x1(10000)</td>
<td>9/20/5.61741e-006</td>
<td>0.11</td>
<td>11/23/8.36494e-006</td>
</tr>
<tr>
<td>x1(10000)</td>
<td>13/27/3.34488e-006</td>
<td>0.11</td>
<td>31/63/9.40955e-006</td>
</tr>
<tr>
<td>x1(50000)</td>
<td>10/21/5.34094e-006</td>
<td>0.21</td>
<td>16/125/9.89342e-006</td>
</tr>
<tr>
<td>x1(50000)</td>
<td>4/9/7.15542e-006</td>
<td>0.12</td>
<td>5/11/4.29325e-006</td>
</tr>
<tr>
<td>x1(50000)</td>
<td>10/21/8.34663e-006</td>
<td>0.21</td>
<td>11/23/9.90944e-006</td>
</tr>
<tr>
<td>x1(50000)</td>
<td>10/22/2.51218e-006</td>
<td>0.22</td>
<td>12/25/2.16294e-006</td>
</tr>
<tr>
<td>x1(50000)</td>
<td>13/27/7.50144e-006</td>
<td>0.27</td>
<td>19/67/3.81389e-006</td>
</tr>
<tr>
<td>x1(100000)</td>
<td>10/21/5.34097e-006</td>
<td>0.12</td>
<td>16/125/8.77333e-006</td>
</tr>
<tr>
<td>x1(100000)</td>
<td>4/9/5.05964e-006</td>
<td>0.06</td>
<td>5/11/3.03579e-006</td>
</tr>
<tr>
<td>x1(100000)</td>
<td>11/23/2.36078e-006</td>
<td>0.14</td>
<td>12/25/2.80281e-006</td>
</tr>
<tr>
<td>x1(100000)</td>
<td>10/22/3.55276e-006</td>
<td>0.13</td>
<td>12/25/3.05886e-006</td>
</tr>
<tr>
<td>x1(100000)</td>
<td>14/29/2.96451e-006</td>
<td>0.18</td>
<td>19/95/7.73838e-006</td>
</tr>
</tbody>
</table>
Table 3. Numerical results for MPRP/CGD on problem 3.

<table>
<thead>
<tr>
<th>MPRP</th>
<th>CGD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Init (Dim)</td>
<td>NI/ NF/</td>
</tr>
<tr>
<td>x₁(10000)</td>
<td>13/27/4.52180e-006</td>
</tr>
<tr>
<td>x₁(10000)</td>
<td>13/27/4.52370e-006</td>
</tr>
<tr>
<td>x₁(10000)</td>
<td>13/27/2.86185e-006</td>
</tr>
<tr>
<td>x₁(10000)</td>
<td>12/25/4.29188e-006</td>
</tr>
<tr>
<td>x₁(10000)</td>
<td>13/27/4.47547e-006</td>
</tr>
<tr>
<td>x₁(50000)</td>
<td>13/27/7.25565e-006</td>
</tr>
<tr>
<td>x₁(50000)</td>
<td>13/27/7.25860e-006</td>
</tr>
<tr>
<td>x₁(50000)</td>
<td>13/27/4.57547e-006</td>
</tr>
<tr>
<td>x₁(50000)</td>
<td>12/25/6.94196e-006</td>
</tr>
<tr>
<td>x₁(50000)</td>
<td>13/27/8.32203e-006</td>
</tr>
<tr>
<td>x₁(10000)</td>
<td>13/27/9.92424e-006</td>
</tr>
<tr>
<td>x₁(10000)</td>
<td>14/29/2.85975e-006</td>
</tr>
<tr>
<td>x₁(10000)</td>
<td>13/27/4.59774e-006</td>
</tr>
<tr>
<td>x₁(10000)</td>
<td>12/25/6.91496e-006</td>
</tr>
<tr>
<td>x₁(10000)</td>
<td>14/29/2.82988e-006</td>
</tr>
</tbody>
</table>

Table 4. Numerical results for MPRP/CGD on problem 4.

<table>
<thead>
<tr>
<th>MPRP</th>
<th>CGD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Init (Dim)</td>
<td>NI/ NF/</td>
</tr>
<tr>
<td>x₁(10000)</td>
<td>16/49/3.81881e-006</td>
</tr>
<tr>
<td>x₁(10000)</td>
<td>11/34/4.30699e-006</td>
</tr>
<tr>
<td>x₁(10000)</td>
<td>11/34/2.49346e-006</td>
</tr>
<tr>
<td>x₁(10000)</td>
<td>11/32/8.70634e-006</td>
</tr>
<tr>
<td>x₁(10000)</td>
<td>13/40/3.26840e-006</td>
</tr>
<tr>
<td>x₁(50000)</td>
<td>16/49/9.28932e-006</td>
</tr>
<tr>
<td>x₁(50000)</td>
<td>11/34/9.68593e-006</td>
</tr>
<tr>
<td>x₁(50000)</td>
<td>11/34/5.57554e-006</td>
</tr>
<tr>
<td>x₁(50000)</td>
<td>12/35/4.06815e-006</td>
</tr>
<tr>
<td>x₁(50000)</td>
<td>13/40/7.33466e-006</td>
</tr>
<tr>
<td>x₁(10000)</td>
<td>17/52/4.28438e-006</td>
</tr>
<tr>
<td>x₁(10000)</td>
<td>12/37/2.86250e-006</td>
</tr>
<tr>
<td>x₁(10000)</td>
<td>11/34/7.88501e-006</td>
</tr>
<tr>
<td>x₁(10000)</td>
<td>12/35/5.75324e-006</td>
</tr>
<tr>
<td>x₁(10000)</td>
<td>14/43/2.89653e-006</td>
</tr>
</tbody>
</table>

Figure 1. Performance profiles for two methods MPRP and CGD, where the left and the right figures are represented as the number of function evaluations and the CPU time, respectively.
4.2. Experiments on the $l_1$-Norm Regularization Problem

The problem of the combination of $l_2$ and $l_1$ norms in the cost function often emerges for the signal reconstruction, i.e.:

$$
\min \frac{1}{2} \| y - Ax \|_2^2 + \lambda \| x \|_1,
$$

(28)

where $\| \cdot \|_2$ is the Euclidean norm, and$
\| x \|_1 = \sum_{j=1}^{m} |x_j|$

is the $l_1$ norm, $A$ is a system matrix, $y \in \mathbb{R}^m$ is the observed data, $x \in \mathbb{R}^n$ is the signal to be reconstructed, and $\lambda$ is a positive regularization parameter.

The optimization problems of the form Equation (28) appear in several signal reconstruction problems, such as sparse signal de-blurring [22], medical image reconstructions [23], compressed sensing [24], and super-resolution [25]. Iterative line search method or fixed point iteration schemes are commonly used to solve problem (28). By using the technique proposed by Figueiredo et al. [26], we can reformulate problem (28) as a convex quadratic program problem. Let $x = u - v$, $u \geq 0$, $v \geq 0$, where $u, v \in \mathbb{R}^n$, $u_i = \max(0, x_i)$ for all $i = 1, \ldots, n$ and $v_i = -\min(0, x_i)$ for all $i = 1, \ldots, n$. The $l_1$ norm can be formulated as $\| x \|_1 = e_n^T u + e_n^T v$, where $e_n = (1,1,\ldots,n)^T$. The problem (28) is expressed as the bound-constrained quadratic program:

$$
\min_{u,v} \frac{1}{2} \| y - A(u-v) \|_2^2 + \lambda e_n^T u + \lambda e_n^T v, \text{ s.t. } u \geq 0, v \geq 0.
$$

(29)

Furthermore, the problem (29) can be rewritten as a standard convex quadratic program problem:

$$
\min_{z} \frac{1}{2} z^T B z + c^T z, \text{ s.t. } z \geq 0,
$$

(30)

where

$$
z = \begin{pmatrix} u \\ v \end{pmatrix}, \quad c = \lambda e_n + \begin{pmatrix} u \\ v \end{pmatrix}, \quad b = A^T y, \quad B = \begin{pmatrix} A^T A & -A^T A \\ -A^T A & A^T A \end{pmatrix},
$$

$B$ is a semi-definite positive matrix. Recently, the problem (30) was reformulated as a linear variable inequality (LVI) problem by Xiao et al. [8] [27]. They pointed out that this LVI problem is equivalent to a linear complementary problem, and $z$ is a solution of the linear complementary problem if and only if it is a solution of the following nonlinear monotone equations:

$$
F(z) = \min \{ z, Bz + c \} = 0,
$$

(31)

where $F(z)$ is Lipschitz continuous. This result indicates that problem (28) can be solved by MPRP projection method.

In this part of numerical experiments, a compressive sensing scenario is considered, which aims to reconstruct a length-$n$ sparse signal from significantly fewer $m$ observations, where $m \ll n$. The quality of restoration is measured by the mean of squared error (MSE) to the original signal $\bar{x}$, that is...
\[ \text{MSE} = \frac{1}{n} \left\| \mathbf{r} - \mathbf{x}^* \right\|, \]

where \( \mathbf{x}^* \) is the restored signal. In practice, \( n = 2^{12} \) and \( m = 2^{10} \), and the original contains \( 2^n \) randomly non-zero elements. \( \mathbf{A} \) is the Gaussian matrix generated by Matlab’s code \( \text{rand}(m,n) \), the measurement \( \mathbf{y} \) contains noise, 
\[ \mathbf{y} = \mathbf{A} \mathbf{x} + \mathbf{\omega}, \]

where \( \mathbf{\omega} \) is the Gaussian noise distributed as \( N(0,10^{-4}) \). The merit function is
\[ f(x) = \frac{1}{2} \left\| \mathbf{y} - \mathbf{A} \mathbf{x} \right\|^2 + \tau \left\| \mathbf{x} \right\|, \]

where \( \tau \) is forced to decrease as the measure in. The experiment starts at the measurement image, i.e. \( \mathbf{x}_0 = \mathbf{A}^\dagger \mathbf{y} \), and terminates when the relative change of the iteration satisfies:
\[ \text{Tol} = \left\| f_k - f_{k-1} \right\| < 10^{-5}, \]

where \( f_k \) is the function value at \( \mathbf{x}_k \).

We compare the proposed MPRP method with CGD method for this problem. In both methods, the parameters are taken as \( \xi = 10 \), \( \sigma = 10^{-4} \) and \( \rho = 0.5 \). The same initial point and continuation technique on parameter \( \tau \) are used in both methods.

**Figure 2** shows simulation results of MPRP and CGD for a signal sparse reconstruction. As we can see in **Figure 2**, the original sparse signal is restored highly exactly both by MPRP and CGD. **Figure 3** provides a series of comparisons among the objective function values and relative error as the iteration numbers and computing time increase. As we can see in **Figure 3**, the descent rates of MSE and objective function values of MPRP method are faster. The experiments are repeated for 15 random different noise samples in **Table 5**. We report the

![Figure 2](image)

**Figure 2.** From top to bottom: the original signal, the measurement, and the recovery signals by two methods MPRP and CGD, respectively.
Figure 3. Comparison results of MPRP and CGD methods. From left to right: the changed trends of MSE and the changed trends of the objective function values goes along with the number of iterations and CPU time in seconds, respectively.

Table 5. The experiment results for MPRP/CGD on $l_1$-norm regularization problem.

<table>
<thead>
<tr>
<th>MSE</th>
<th>Niter</th>
<th>CPU(s)</th>
<th>MSE</th>
<th>Niter</th>
<th>CPU(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.152e-006</td>
<td>119</td>
<td>2.69</td>
<td>2.278e-005</td>
<td>227</td>
<td>6.73</td>
</tr>
<tr>
<td>1.562e-005</td>
<td>120</td>
<td>3.23</td>
<td>6.210e-005</td>
<td>172</td>
<td>4.72</td>
</tr>
<tr>
<td>6.780e-006</td>
<td>127</td>
<td>3.47</td>
<td>2.520e-005</td>
<td>209</td>
<td>5.83</td>
</tr>
<tr>
<td>8.236e-006</td>
<td>124</td>
<td>3.05</td>
<td>3.367e-005</td>
<td>236</td>
<td>7.48</td>
</tr>
<tr>
<td>1.446e-005</td>
<td>120</td>
<td>3.09</td>
<td>8.207e-005</td>
<td>167</td>
<td>4.64</td>
</tr>
<tr>
<td>9.091e-006</td>
<td>110</td>
<td>2.25</td>
<td>4.870e-005</td>
<td>221</td>
<td>6.09</td>
</tr>
<tr>
<td>8.346e-006</td>
<td>122</td>
<td>3.31</td>
<td>5.382e-005</td>
<td>174</td>
<td>5.47</td>
</tr>
<tr>
<td>8.669e-006</td>
<td>117</td>
<td>3.23</td>
<td>4.233e-005</td>
<td>216</td>
<td>5.91</td>
</tr>
<tr>
<td>6.977e-006</td>
<td>123</td>
<td>3.33</td>
<td>3.839e-005</td>
<td>210</td>
<td>5.78</td>
</tr>
<tr>
<td>8.973e-006</td>
<td>122</td>
<td>3.70</td>
<td>3.789e-005</td>
<td>225</td>
<td>6.88</td>
</tr>
<tr>
<td>1.050e-005</td>
<td>119</td>
<td>3.30</td>
<td>5.531e-005</td>
<td>208</td>
<td>5.86</td>
</tr>
<tr>
<td>1.204e-005</td>
<td>128</td>
<td>2.63</td>
<td>5.370e-005</td>
<td>204</td>
<td>5.27</td>
</tr>
<tr>
<td>6.265e-006</td>
<td>111</td>
<td>3.52</td>
<td>1.873e-005</td>
<td>202</td>
<td>6.66</td>
</tr>
<tr>
<td>8.977e-006</td>
<td>129</td>
<td>3.70</td>
<td>3.035e-005</td>
<td>222</td>
<td>6.28</td>
</tr>
<tr>
<td>7.975e-006</td>
<td>126</td>
<td>3.47</td>
<td>6.946e-005</td>
<td>172</td>
<td>4.78</td>
</tr>
</tbody>
</table>
number of iterations (Niter) and the CPU time (in second) required for the whole testing process. From Table 5, we can see that MPRP method is better than CGD method. For example, the new method’s iteration number and CPU time are much less than those of the CGD method. To summarize, these experiment results show that the proposed algorithm MPRP can work well in an efficient manner.

5. Conclusion

In this paper, we proposed a conjugate gradient projection algorithm for solving large-scale nonlinear convex constrained monotone equations based on the well-known Polak-Ribière-Polyak conjugate gradient method which is one of the most effective conjugate gradient methods to solve the unconstrained optimization problems. The algorithm combines CG technique with projection scheme and is a derivative-free method, so it can be applied to solve large-scale non-smooth equations for its low storage requirement. Under some technical conditions, we have established the global convergence. Another contribution of this paper is to use the given method to solve the $l_1$-norm regularized problems in compressive sensing.

Acknowledgements

This work was supported by the Scientific Research Project of Tianjin Education Commission (No. 2019KJ232).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References


DOI: 10.4236/jamp.2020.86077


The Confirmation of Hypothesis of the Absolute Reference System

Konstantinos Patrinos

National Technical University of Athens, Athens, Greece
Email: kpatr@central.ntua.gr

Abstract

The purpose of the research in this article is the examination of the agreement of the hypothesis of the absolute reference system with the results of experiments that have been implemented in the past in order to confirm the special theory of relativity. To achieve this goal, we have chosen for discussing a theoretical topic of electromagnetism, that of electromagnetic mass calculation, and some experiments, some of which concern the transverse Doppler effect in a rotated system, two experiments that concern the kinetic energy measurement of accelerated electrons, one of which is the well-known Bertozzi experiment, one experiment that concerns the propagation of Coulomb fields and one more experiment that concerns the effect of annihilation. The basic principles of the hypothesis of the absolute reference system, and the electromagnetic theory derived from these principles, are used to explain the experimental results. In these examples, the hypothesis of the absolute reference system is confirmed, since the experimental results agree with the predictions of this hypothesis. Also, in the discussion of calculation of electromagnetic mass is addressed the difficulty of solving this problem, when someone tries to solve this according to the energy-mass relation of the theory of relativity.

Keywords

Absolute Reference System, Electromagnetic Mass, Transverse Doppler Effect, Kinetic Energy of Accelerated Electrons

1. Introduction

In all the experiments carried out in order to examine the agreement of the special theory of relativity with the corresponding experimental results, the theoretical predictions that are commented on in relation to this agreement are those of the theory of relativity and the Newtonian physics. This means that the theories
examined in all these experiments are the special theory of relativity and Newtonian physics. Bertozzi’s experiment is one such example. What is being proven in this experiment is that the speed of the accelerating electrons can never reach the speed of light in vacuum. However, as can be seen from a closer look at this article, the expected relativistic kinetic energy values resulting from the measured velocities of the electrons deviate greatly from the measured kinetic energies, which were measured by a thermodynamic method. But because this deviation for the corresponding Newtonian energies was much greater, this was enough to convince the scientific community of the correctness of the theory of special relativity.

Of course, Newtonian physics deviates significantly from experimental reality as particle speed and energy increase. This article examines a new theory in terms of its agreement with these experimental results, and this theory is expressed in the newly introduced hypothesis of the absolute reference system. The previous comparison, therefore, concerns in this case the predictions of the special theory of relativity in comparison with the predictions of the hypothesis of the absolute reference system, regarding these experimental results.

It was deemed appropriate in this article to first consider a theoretical issue, that of the problem of incompatibility of the electromagnetic mass with the relation $E = mc^2$ of the special theory of relativity, because this incompatibility is eliminated by applying the principles of hypothesis of the absolute reference system for the electromagnetic interactions. This is a much-discussed problem, but we will refer specifically to the relevant analysis of this issue by Feynman, as it is examined in [1], and in this analysis also are presented the efforts to solve this problem. Despite these efforts, however, the relativistic problem remains, as one can see by reading the current literature.

In the study of the experiment of positron-electron annihilation process, it seems that the theoretical results of the hypothesis of the absolute reference system are the same as those of the special relativity. In all other experiments studied in this article, the agreement of the newly introduced hypothesis with the experimental results is examined, because there is no such agreement of the experimental results and the corresponding relativistic theoretical predictions (for example, the experiment for Measuring Propagation Speed of Coulomb Fields, in [2]).

The attempt to interpret natural phenomena on the basis of the hypothesis of an absolute reference system does not based on the attempt to interpret the phenomena using the relativistic conception, which was formulated in Galilean relativity and subsequently in Einstein’s theory of relativity. The relativistic notion is based on the evaluation of physical quantities by the observer of the natural phenomenon at the observation point. On the basis of the absolute reference system, the space-time is Newtonian and the physical quantities such as momentum, force and kinetic energy are not determined on the basis of the relative velocity of the observed object by the observer’s reference system, but on the basis of the object’s reference system and the inertial reference system of the
source of the field that affects the state of the object.

Electromagnetic interactions based on this hypothesis are carried out by force carriers, which are real photons with some peculiar behavior, unknown to this day, although this theoretical result is derived from classical electromagnetic theory. Also as explained in [3] the origin of creation of mass is electromagnetic and the same is true for the origin of the quantum nature of matter particles, as described in detail in [4]. Using this hypothesis, we will then provide answers to unresolved, by the modern physics theory, issues.

2. Electromagnetic Mass

The very old problem of electromagnetic mass in relation to the equation of mass and energy in the special theory of relativity is the first issue that we will consider in this paper. We will begin first with the study of this subject according to classical electromagnetic theory.

We consider a particle of charge \( e \) moving at velocity \( v \) with respect to the inertial laboratory reference system. We will calculate the energy of the electromagnetic field in space between two concentric spheres of radii \( a \) and \( b \). We consider that the center of the spheres is the center of mass of the charged particle. We will then calculate the momentum and mass of the electromagnetic field in the space between the two spheres. Assuming \( ab \) the electromagnetic energy is:

\[
U_{\text{elec}} = \frac{1}{2} \frac{e^2}{4 \pi \epsilon_0} \left( \frac{1}{a} - \frac{1}{b} \right)
\]

when the speed \( v \) is much less than the speed of light in the vacuum, the momentum of that field is:

\[
p = \frac{2}{3} \frac{e^2}{4 \pi \epsilon_0 c^2} v \left( \frac{1}{a} - \frac{1}{b} \right)
\]

The corresponding electromagnetic mass is:

\[
m_{\text{elec}} = \frac{2}{3} \frac{e^2}{4 \pi \epsilon_0 c^2} \left( \frac{1}{a} - \frac{1}{b} \right) = \frac{4 U_{\text{elec}}}{3 c^2}
\]

According to the absolute reference system hypothesis the previous equation is fully justified. It turns out that, on the basis of this hypothesis, the force carriers of Coulomb electric field are real photons. It also turns out that part of the energy \( h\omega \) of these photons is the energy of an elementary mass. The amount of energy of this elemental mass, as shown by the relevant calculations, is equal to \( \frac{2}{3} h\omega \). If we denote by \( m_{\text{ph}} \) the mass of this photon, the energy equation of this photonic mass is 

\[
m_{\text{ph}} c^2 = \frac{2}{3} h\omega.
\]

This result has been obtained by making use of Maxwell’s equations and the

---

\[1\] The solution to this problem is analogous to that in [1], section 28.

\[2\] [3], subsection 2.4 (equations 2.44).
Coulomb field spherical symmetry. Also, according to the same hypothesis, the total kinetic energy of all the photons of the field in the space between the two spheres, for \( N \) photons, is 
\[ U_{\text{dec}} = \frac{1}{2} \hbar \sum_{i=1}^{N} \omega_i. \]
Therefore, in accordance with all the foregoing, which derive on the basis of the absolute reference system hypothesis, the electromagnetic mass is:
\[ m_{\text{elec}} = \sum_{i=1}^{N} m_{\text{pho}} = \frac{2}{3} \frac{\hbar}{c^2} \sum_{i=1}^{N} \omega_i = \frac{4}{3} \frac{U_{\text{dec}}}{c^2} \]
(4)

3. Transverse Doppler Effect in a Rotated System

The origin of this problem lies in Walter Kundig’s 1963 experiment entitled “Measurement of the Transverse Doppler Effect in an Accelerated System”, in [5]. A re-analysis of the experimental data of this scientific work, by Alexander L Kholmetskii, Tolga Yarman, Oleg V Missevitch and Boris I Rogozev, in 2008, showed that they conclude that “instead of the relative energy shift \( \Delta E/E = -\left(1.0065 \pm 0.011\right)v^2/2c^2 \) reported by Kundig (\( v \) being the linear velocity of absorber and \( c \) being the light velocity in vacuum), we derive from his results \( \Delta E/E = -\left(1.192 \pm 0.011\right)v^2/2c^2 \)” and the details of this data analysis are in [6].

The experiment was repeated in 2009 by the same scientific team, and the result of the experiment, exposed to [7], showed that the results of the previous analysis are confirmed by the new experimental results. In particular, as a result, for the equation \( \Delta E/E = -kv^2/c^2 \), they obtained the overall estimation \( k = 0.68 \pm 0.03 \).

One attempt to explain this experimental result on the basis of theory of relativity, in 2015, in [8] was unsuccessful, as detailed in [9], in the response given by the aforementioned authors.

The experiment has been repeated several times. In 2011 by A. L. Kholmetskii, T. Yarman and O. V. Missevitch. (Int. J. Phys. Sci. 6, 84 (2011)), the experimental result yields the coefficient \( k = 0.66 \pm 0.03 \) within the expression frame for the relative energy shift between emission and absorption lines \( \Delta E/E = -kv^2/c^2 \).

We will then use the theory of the absolute reference system to address this problem. According to this hypothesis in the Coulomb field of an atomic nucleus of the absorber material, the energy \( h\omega \) of the absorbed photons is divided into two components of energy, due to the two components of their helical motion within this field, as explained in [3]. This is due, as we have already mentioned, to the spherical symmetry of the electric field of the nucleus and to the wave equation of the electric field, which results from Maxwell’s equations. One of the two components of the motion is radial and if the corresponding angular frequency is denoted by \( \omega_r \), the corresponding energy is equal to \( h\omega_r = \frac{1}{3}h\omega \).

The other component is circular and the corresponding energy, characterized as the energy of mass of the photon, is equal to \( h\omega_\theta = \frac{2}{3}h\omega \).
In the case of the experiment in [5] the absorber rotates at a speed \( u \). The rate of energy absorption of the radial component of the motion of the photons is determined by the Doppler effect. So the corresponding absorbed energy is estimated from the relation 
\[
h \omega' = \gamma h \omega, \quad \gamma = \left(1 - \frac{u^2}{c^2}\right)^{1/2}
\]
and the angular frequency measured by the clock of the momentarily inertial reference system of absorber is denoted by \( \omega' \). Also we denote by \( \omega_r \) the angular frequency measured by the clock of the laboratory reference system. The energy supplied by the source, due to the time contraction, will be \( \gamma \) times less than that measured by the clock of the absorber, and are given by the relation:
\[
E_{ar} = \frac{1}{\gamma} h \omega' = h \omega
\]  
(5)

The amount of energy \( h \omega_r \), characterized as energy of mass, obeys the kinetic energy conservation relation of an elementary photon mass in the source-absorber system:
\[
\frac{1}{2} h \omega = \frac{1}{2} h \omega_v + E_{phkin}
\]  
(6)

where \( E_{phkin} \) is the kinetic energy of the photonic mass due to the rotary motion of the absorber and the angular frequency of the photonic mass, measured by the clock of the instantaneous inertial system of the absorber, is denoted by \( \omega_v \). If the absorbed photon mass is denoted by \( m_{va} \), the previous relation becomes:
\[
\frac{1}{2} h \omega = \frac{1}{2} m_{va} c^2 + \frac{1}{2} m_{va} \gamma^2 u^2
\]  
\[
= \frac{1}{2} m_{va} c^2 \left(1 + \gamma^2 \beta^2\right)
\]  
\[
= \frac{1}{2} \gamma^2 h \omega_v
\]  
(7)

where \( m_{va} = h \omega_v c^2 \) and \( \beta = u/c \). The angular frequency of the absorbed mass according to the last relation is \( \omega_v = \omega_r / \gamma^2 \). The change in absorption rate due to the Doppler effect is canceled by the time contraction as in the case of calculating the radial motion energy \( h \omega_r \), previously mentioned. The energy supplied by the source will be given by the relation:
\[
E_{av} = h \omega_v = \frac{1}{\gamma^2} h \omega_r
\]  
(8)

Therefore, the total absorbed photon energy is \( E_a = E_{ar} + E_{av} \), while the total energy of a source photon is \( E_s = h \omega \) and only an amount of kinetic energy, equal to \( E_{phkin} = \frac{1}{2} m_{va} \gamma^2 u^2 \), is offered by the rotated absorber, whereby, the equation of the relative energy shift between emission and absorption lines, is:

---

3The rate of absorption of the photons results from the relation (4.31) of section 4.2, Doppler Effect, in [3], in the inertial reference system of the laboratory instead of the absolute reference system, where we set \( u_e = 0 \), \( \phi = 0 \), \( c_e = c \), \( u_e = u \), \( \gamma_e = \gamma = (1 - u'/c')^{1/2} \) and the frequencies \( \nu_e \) and \( \nu_l \) are replaced by the angular frequencies \( \omega' \) and \( \omega_r \) according to clock of the absorber and the laboratory clock respectively.
This result is in excellent agreement with the above experimental results.


In order to experimentally test the correctness of the absolute reference system hypothesis, we will study two experiments measuring the kinetic energy of electrons, which simultaneously measure the beam electron velocity in a linear accelerator. The peculiarity of these experiments is that the velocity is measured by measuring the flight time of the electrons at a given travel length. One of these two experiments is the historical experiment of Bertozzi, described in [10]. The other experiment is that of M. Lund and U. I. Uggerhøj and the relevant details and experimental results are reported in [11].

According to the absolute reference system hypothesis, kinetic energy which is transferred to the detection device, which is the target of accelerated electrons, when the kinetic energy $E_a$ of the electrons is high enough so that the scattering of the force carriers by atoms of target to be no longer elastic, equals $E_a/2$. The remaining half of the kinetic energy of the electrons is converted into mass absorbed by the atoms of the target. This absorbed mass is derived from the photon mass of the electromagnetic field force carriers of the accelerated electrons. As we have mentioned, this electromagnetic mass of the force carriers are real photons that exhibit such behavior under the given conditions in the Coulomb field.

A first estimate of this phenomenon in the Bertozzi experiment is made by studying the experimental results presented in [4]. There are two very useful measurements of the detectable kinetic energy of the electrons in this experiment, derived from the corresponding measurements of the target heat gain by simultaneously measuring the charge, that is, the number of beam electrons that hit the target. These measurements are shown in Table 1.

Table 1. Experimental results of W. Bertozzi’s measurements, as set out in his work entitled “The Ultimate Speed”, in 1964. The measured increase of the target heat per electron is denoted by $E_a$. 

<table>
<thead>
<tr>
<th>Measured energy $E_a$, MeV</th>
<th>Electron velocity $u$, x10^6 m/sec</th>
<th>Relativistic kin. energy $E_r$, MeV</th>
<th>Detectable kin. energy $E_d$, MeV</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.6</td>
<td>2.88</td>
<td>1.3</td>
<td>1.5</td>
</tr>
<tr>
<td>4.8</td>
<td>2.96</td>
<td>2.7</td>
<td>4.8</td>
</tr>
</tbody>
</table>
The transmitted kinetic energy of the beam is equal to the increase in target heat. This energy per electron, measured by the thermal arrangement of the Bertozzi experiment, is denoted by \( E_m \) in Table 1. The relativistic kinetic energy calculated from the measured velocity \( u \) is denoted by \( E_k \) and is equal to \( m(\gamma - 1)c^2 \), where \( \gamma = \left(1 - u^2/c^2\right)^{-1/2} \) and \( m \) is the mass of the electron. The corresponding kinetic energy transferred according to the absolute reference system hypothesis is \( E_{det} = \frac{1}{4}m\gamma^2u^2 \), while the kinetic energy of the beam electrons is \( E_a = \frac{1}{2}m\gamma^2u^2 \). The values of these kinetic energies are in Table 1, where it appears that the energy values \( E_{det} \), predicted by the absolute reference system hypothesis, are in excellent agreement with the measured values \( E_m \).

The transfer of kinetic energy in the form of heat to the target is accomplished by the transfer of the kinetic energy of the electromagnetic mass of the force carriers of the electron beam electromagnetic field of the linear accelerator. The electromagnetic mass \( m_{el} \), passing through the active scattering cross section of the force carriers with the bound electron of an atom in the target material, transfers all the kinetic energy of the beam electron and is equal to the mass of this electron from which it is derived, that is, \( m = m_{el} \). In order for the scattering to be completely inelastic, it must be possible to incorporate half the kinetic energy of the beam’s electron, that is, energy equal to \( E_a/2 \), as photonic mass in the atoms of the target material. There is a limitation to be implemented this completely inelastic interaction. Assuming that \( N \) force carriers transfer the kinetic energy of a beam electron, the corresponding electromagnetic mass, according to the relation (4), is \( m_{el} = \frac{2}{3}\frac{\hbar}{c^2} \sum_{i=1}^{N} \omega_i \). Therefore for fully inelastic scattering the energy \( E_a/2 \) should be greater than \( \hbar \sum_{i=1}^{N} \omega_i \), which is the total energy of the Coulomb field in the area of the specific electromagnetic mass \( m_{el} \). Therefore the threshold for a fully inelastic scattering is determined by the relation:

\[
\frac{E_a}{2} = \frac{1}{4}(\gamma^2 - 1)mc^2 = \hbar \sum_{i=1}^{N} \omega_i = \frac{3}{2}m_{el}c^2
\]

Since \( m = m_{el} \), the last relation gives \((1/4)(\gamma^2 - 1) = 3/2\) and therefore this relation is satisfied for \( \gamma \) values greater than a critical value of \( \gamma \) which is \( \gamma_c = \sqrt{\frac{3}{2}} \) and a corresponding critical speed \( u_c \) given by the relation \( u_c = \frac{\sqrt{6}}{\sqrt{\gamma_c}} c \). The corresponding kinetic energy is \( E_c = \frac{1}{4}(\gamma_c^2 - 1)mc^2 = \frac{3}{2}mc^2 \). Therefore the criterion for completely inelastic scattering is \( \gamma \geq \sqrt{\frac{3}{2}} \) or equivalently \( \frac{1}{2}E_a \geq \frac{3}{2}mc^2 \).

For \( \gamma < \sqrt{\frac{3}{2}} \) an amount of energy equal to \( \frac{1}{4}(\gamma^2 - 1)mc^2 \) is not sufficient to fully incorporate all the electromagnetic mass \( m_{el} \), but only one part of it, which
is equal to $am_d$, where the coefficient $a$ is equal to the ratio of the energy values $(E_d/2)/E$, that is, $a = (\gamma^2 - 1)/(\gamma^2 - 1)$. In this case the amount of kinetic energy incorporated as a mass is equal to $\frac{1}{4}(\gamma^2 - 1)am_d c^2$. Therefore the detectable kinetic energy transferred to the target will be:

$$E_{det} = \frac{1}{2}(\gamma^2 - 1)mc^2 - \frac{1}{4}(\gamma^2 - 1)\frac{\gamma^2 - 1}{6}mc^2$$

$$= \left(\frac{\gamma^2 - 1}{2} - \frac{(\gamma^2 - 1)^2}{24}\right)mc^2$$

(11)

According to the latter relation, at low speeds, that is, $u \ll c$ and $\gamma \to 1$, the detectable value of energy tends to the Newtonian value of kinetic energy, that is, $E_{det} \to \frac{1}{2}mu^2$.

The electron beam measured energy values of the experiment by M. Lund and U. I. Uggerhøj are in excellent agreement with the values of the energy curve $E_{det}$ in the two aforementioned speed ranges, that is, for $\gamma \geq \sqrt{7}$, where $E_{det} = E_d/2$, and for $\gamma \leq \sqrt{7}$, where the relation (11) applies, that is, the experimental data in [10] and [11] (see Figure 1) are in excellent agreement with the detectable energy curves $E_{det} = \frac{1}{4}(\gamma^2 - 1)mc^2$ for $u/c \geq \frac{6}{\sqrt{7}}$ and $E_{det} = \left((\gamma^2 - 1)/2 - (\gamma^2 - 1)^2/24\right)mc^2$ for $u/c \leq \frac{6}{\sqrt{7}}$.

**Figure 1.** The kinetic energy of the electron obtained with 0.42 MeV electrons from a linear accelerator as a function of the ratio of the electron velocity and the speed of light $c$. The continuous line is a curve based on the expression of detectable energy in region $u/c \geq \frac{6}{\sqrt{7}}$ where the equation of detectable energy is $E_{det} = \frac{1}{4}(\gamma^2 - 1)mc^2$ and $u/c \leq \frac{6}{\sqrt{7}}$ where the relation $E_{det} = \left((\gamma^2 - 1)/2 - (\gamma^2 - 1)^2/24\right)mc^2$ holds. The dashed line is a curve based on the relativistic expression, $E_d = m(\gamma - 1)c^2$. 
In the case of very high energies of electrons, for example energies of the order of several tens of TeV, due to the very high value of the coefficient $\gamma$, the actual physical contraction of the particle contributes to the decomposition of the particle, and, so, the particle no longer behaves as a solid body. This has the effect of hitting the target by a photon beam instead of the particle beam and these photons of the total mass of the particles have frequencies denoted by $\nu_\parallel$, measured in the laboratory, which are $\gamma$ times smaller than the corresponding measured frequencies in the particle reference system.

It should be noted that according to the hypothesis of the absolute reference system, the Lorentz contraction is a real physical contraction and not a geometric contraction, in Minkowski’s space-time, which concerns the observer’s inertial system, as inferred from the special theory of relativity.

In this case the kinetic energy is $E_a = \frac{1}{2} \gamma h \sum \nu_\parallel = \frac{1}{2} \left( \gamma^2 - 1 \right) mc^2$ and the total energy transferred to the target is equal to $\hbar \sum \nu_\parallel \approx \gamma mc^2$.

5. Propagation of Coulomb Fields

The theoretical calculations in this section concern the measurement of the potential difference that results from the effect of the maximum transverse electric field on the detector of the experiment in [2]. The detection device comprises a Data Acquisition System described as “Data Acquisition System by means of fast, terminated coax cables. To record the sensors waveforms we used a Switched Capacitor Array (SCA) circuit (CAEN mod V1472) able to sample the input signal at 5 GHz. In addition to the sensors output, the SCA stored also the LINAC-RF trigger and the toroid pulse …”

According to the following equation ([12], page 664, relation (14.14)):

$$[E]_{ret} = \frac{e}{4\pi\epsilon_0} \left[ \frac{n - \beta}{\gamma^2 \left( 1 - \beta \cdot n \right)^2} \right]_{ret}$$  \hspace{1cm} (12)

the transverse component of the electric field, in the unit system SI, at a constant height $y = b$, with the beam extending along the $Z$-direction, considering that the point $M$ is the origin of the axes ([12], figure 14.2, page 664), is given by the equation:

$$E_z = \frac{e}{4\pi\epsilon_0} \frac{b}{\gamma^2 \left( \sqrt{b^2 + z^2} + \beta z \right)^3}$$  \hspace{1cm} (12)

and the maximum transverse component of the electric field is:

$$E_{z,\text{max}} = \frac{e}{4\pi\epsilon_0} \frac{\gamma}{b^2}$$  \hspace{1cm} (13)

We first assume that a free electron beam propagates in a linear range of about 1000 meters, with the sensor in the middle of this distance. In Figure 2 is shown the curve $E_z/E_{z,\text{max}}$, in such a lab, according to the previous analysis in [12].

Concerning the electron beam of the accelerator it is mentioned “… the
electron beam produced at the DAΦNE Beam Test Facility (BTF) [11], a beam line built and operated at the Frascati National Laboratory to produce a well-defined number of electrons (or positrons) with energies between 50 MeV and 800 MeV;” and its description is given in the phrase “At maximum intensity the facility yields, at a 50 Hz repetition rate, 10 ns long beams with a total charge up to several hundreds pCoulomb.”.

Therefore the length of a portion of the beam coming from a burst is \( \ell = 10 \text{ ns} \times 3 \text{ m} \). We assume that the sensor is above the middle of this segment and that \( \phi \) is the angle formed by \( R'(t') \) defining the position of the electron giving the signal, and by the linear segment \( OM \) ([12], figure 14.2, page 664). The point \( O \) is assumed at the bottom of the sensor.

According to the relation \( R'(t') - (u/c)R(t') \cdot u = R(t)\sqrt{1 - \left(\frac{u^2}{c^2}\right)\sin^2(\theta(t))} \) ([13], §63) and the relation \( R'(t') - (u/c)R(t') = R(t) \) the electric field as a function of observation time \( t \) is:

\[
E(t) = \frac{e}{4\pi\varepsilon_0} \frac{R(t)}{R(t)^2} \left( \frac{1 - \frac{u^2}{c^2}}{1 - \frac{u^2}{c^2}\sin^2(\theta(t))} \right)^\frac{3}{2}
\]

The differential component derived from an elementary charge \( dq = \lambda dz \), where \( \lambda \) is the linear charge density of the segment of the beam in time \( t \), taking into account that \( y = b = R(t)\cos(\phi(t)) \), is:

\[
dE_y(t) = \frac{\lambda}{4\pi\varepsilon_0\gamma^2} \frac{\cos(\phi(t))dz}{R(t)^2\left(1 - \frac{u^2}{c^2}\sin^2(\theta(t))\right)^\frac{3}{2}}
\]

Since \( R(t)\cos(\phi(t)) = y \) and \( d\tan(\phi(t)) = dz(t)/y = d\phi(t)/\cos^2(\phi(t)) \), the following relation applies:

\[
\frac{dz(t)}{R(t)^2} = \frac{d\phi(t)}{y}
\]
The angles \( \phi \) and \( \theta \) are complementary, so:

\[
dE_j(t) = \frac{\lambda}{4\pi\varepsilon_0 y^2} \frac{d\sin \phi}{\left(\frac{1}{\gamma^2} + \frac{u^2}{c^2}\sin^2(\phi)\right)^{\frac{3}{2}}}
\]

(18)

For convenience, we set \( \beta \sin \phi = x \), where \( \beta = \frac{u}{c} \), and so:

\[
dE_j(t) = \frac{\lambda}{4\pi\varepsilon_0 y^2} \frac{1}{\beta} \frac{dx}{\left(\frac{1}{\gamma^2} + x^2\right)^{\frac{3}{2}}}
\]

(19)

In a linear infinite distribution, because of the symmetry, the integration gives a transverse component of the electric field, at a height \( y \), at time \( t \), equal to:

\[
E_j(t) = \frac{2\lambda}{4\pi\varepsilon_0 y^2} \frac{1}{\beta} \int_0^\beta \frac{dx}{\left(\frac{1}{\gamma^2} + x^2\right)^{\frac{3}{2}}} = \frac{\lambda}{2\pi\varepsilon_0 y} \left[ \frac{1}{\beta} \left(\frac{x}{\gamma^2\sqrt{x^2 + \frac{1}{\gamma^2}}}\right) \right]_0^{\beta} = \frac{\lambda}{2\pi\varepsilon_0 y}
\]

(20)

Since the length of this beam segment is finite (\( \ell \simeq 3 \text{ m} \)), the maximum angle \( \phi_{\text{max}} \) is 79° and the corresponding sine, at a height of \( y = 30 \text{ cm} \), is \( \sin \phi_{\text{max}} = 0.98 \). Therefore, the term in square brackets, of the previous relation, becomes:

\[
\left[ \frac{1}{\beta} \left(\frac{x}{\gamma^2\sqrt{x^2 + \frac{1}{\gamma^2}}}\right) \right]_0^{0.98\beta} \simeq 0.99999998y^2 \approx y^2
\]

(21)

Therefore, this finite length portion of the beam gives a transverse electric field component equal to that which is given by an infinite length beam.

The maximum potential difference at the ends of the sensor \( y_1 = y, y_2 = y + 14 \text{ cm} \) is:

\[
V_{\text{max}} = \frac{\lambda}{2\pi\varepsilon_0} \ln \left(\frac{y + 14 \text{ cm}}{y}\right)
\]

(22)

If we multiply the second member of the previous relationship by a calibration coefficient \( \eta \), we get the relation (8) of the bibliographic reference [2].

The transverse component of the electric field derived from a single electron, based on the Equation (15), and the relation \( R(t) = y/\sin \theta \), is given by the relation:

\[
E_z(t) = \frac{e}{4\pi\varepsilon_0} \frac{\sin^2 \theta}{y^2} \frac{1 - u^2/c^2}{\left(1 - u^2/c^2\sin^2(\theta(t))\right)^{\frac{3}{2}}}
\]

(23)

The ratio \( E_z(t)/E_{z_{\text{max}}} \) is therefore given by the relation:
\[
\frac{E_z(t)}{E_{2\text{max}}} = \frac{\sin^3(\theta(t))}{\gamma^3(1 - \frac{u^2}{c^2}\sin^2(\theta(t)))^{\frac{3}{2}}}
\] (24)

In Figure 3 is shown the graph of the function \( E_z(t)/E_{2\text{max}} \) of the previous relation. This function looks like a delta function.

As can be seen from Figure 3 only the electrons beneath the sensor, at a very short distance from the point \( M \), contribute in the transverse component of the electric field. This means that the corresponding signals depart from a position \( Z \) near the position \( z = -300 \) m in the negative semi axis.

In order to better understand this, we will calculate the relation between \( z(t') \) and \( z_i = z(t) \). From the known relation \( z(t) = z(t') + \beta R(t') = z(t') + \beta\sqrt{b^2 + z(t')^2} \), by setting \( z = z(t') \), \( z_i = z(t) \), we obtain the following equation:

\[
z^2 - 2\gamma^2 z_i z + \gamma^2 z_i^2 - \beta^2\gamma^2 b^2 = 0
\] (25)

The resulting values of \( z \) are \( z = \gamma^2 z_i \pm \beta\gamma\sqrt{\gamma^2 z_i^2 + b^2} \). The value with (+) is rejected, since for \( z_i = 0 \) we have the relation \( z = -\beta\gamma b \). This solution shows that for very short distances from \( M \) we have large absolute values of \( z \) and that is why we take the curve of Figure 2.

However, since in the laboratory of this experiment the electronic beam extends at distances of less than 10 m, the sensor should receive signals departing from this region, and so the transverse component of the electric field should be much smaller relative to the maximum we calculated in the previous relation.

In Figure 4 is shown the graph of the ratio \( E_z/E_{2\text{max}} \) in the range \(-10 < z < 0\). It is obvious that the ratio \( E_z/E_{2\text{max}} \) in the range \(-10 < z < 0\) is of the order of \( 10^{-4} \), and such a signal is not detectable by the experimental device. Therefore, the measured signal is many orders of magnitude larger than the predicted by relativistic theory.

From the Equation (25), if we consider \( b = y \), we obtain an acceptable solution \( z_i = z + \beta\sqrt{z^2 + y^2} \). Assuming that the left end of the portion of the electron beam is in the initial position where the charges generating the field are not be shielded by conductors and it is possible to emit an electromagnetic signal, then the signal will arrive at the sensor at time \( t \) and this end of the beam will be at the corresponding position \( z_{11}(t) \). The right end of the beam at the same time \( t \) will be at the position \( z_{12} = z_{11} + 3 \) m. The limits of integration of the relation (20) are determined by the relation \( x_i = \beta\sin(\arctan(z_{1i}/y)) \).

The calculation of the ratio of the relativistic maximum electrical potential differences to measured maximum electrical potential at three different positions of the sensor in the direction of \( Z \), at 1720 mm, 3295 mm, and 5525 mm, but at a constant height from the lower end of sensor end at 30 cm, gives us:

\[
\frac{V_{1720}}{V_{\text{max}}} = 4.6914\times10^{-5}
\]

\[
\frac{V_{3295}}{V_{\text{max}}} = 1.6953\times10^{-4}
\]
Figure 3. The curve of the function $E_x/E_{x_{\text{max}}}$ in time $t$ looks like a delta function.

Figure 4. The ratio $E_x/E_{x_{\text{max}}}$ in the range $-10 \, \text{m} < z < 0$ is of the order of $10^{-4}$.

$$V_{5525}/V_{\text{max}} = 4.7477\times 10^{-4}$$

where $V_{1720}$, $V_{1295}$, $V_{5525}$, the calculated electrical potential differences according to the special theory of relativity, that is, at the height $y = 30 \, \text{cm}$ at which we consider that is the location of the sensor, the measured $V_{\text{max}}$ is many orders of magnitude higher than the one we would expect from the calculations of relativistic electric potential differences.

In Figure 5 is shown the graph of the ratio $V_{\text{rel}}(t)/V_{\text{max}}$ versus time, within time interval $-10 \, \text{ns} < t < 20 \, \text{ns}$, at $z_{45,46} = 1.72 \, \text{m}$, transverse distance 30 cm.

We consider that $t = 0$ is the time the initial signal is received at the sensor.

However, the measurements of this experiment show that the maximum electric potential differences are in agreement with the relation (22). Therefore, as shown in Figure 5, at $z_{45,46} = 1.72 \, \text{m}$, transverse distance 30 cm, given that $V_{1720}/V_{\text{max}} = 4.6914\times 10^{-5}$ according to previous calculations, the highest value of the relativistic potential difference is about 20,000 times smaller than the one measured in the experiment.
Theoretical Study According to the Hypothesis of the Absolute Reference System

In the abstract of my published work entitled “The Physics of an Absolute Reference System” ([3] [4]), the first of the basic principles of the absolute reference system hypothesis is described as “… The first of these principles is that the electromagnetic field quantitative estimates are made in the inertial reference system of the source of the electromagnetic field …” Also, according to the hypothesis of the absolute reference system, there are no transformations of physical magnitudes, because the Lorentz contraction is physical, not geometric, that is, it is not a space-time view of the observer, as in the special theory of relativity. 

In the present experiment of the reference [2] the inertial reference systems are two, namely the first one is the reference system of the laboratory and the other is that of the electronic beam. Therefore, because the source of the electric field is the beam electrons, the electric field may be calculated in the inertial reference system of the electron beam and this calculated value is equal to the calculated value of the electric field in the reference system of laboratory. This principle contradicts the basic principles of special theory of relativity, since it does not take into account the relative velocity between the beam electrons and the sensor, but only the speed of the beam inertial system relative to the inertial reference system of laboratory (because the inertial system of the beam comes from the action of electromagnetic fields whose sources are in the laboratory). This means that according to the hypothesis of the absolute reference system there is no the relativistic concept of the evaluation of the electric field at the retarded time giving the corresponding electrical and magnetic potential at the observation point. Also in the beam reference system, the sensor “scans” the electric field region, while in the laboratory reference system there is no the limitation of the maximum velocity $c$ at the speed of the interaction photons. Therefore under this condition, an observer of the laboratory will see an electric field rigidly carried by the beam itself.
We will now calculate the electric field derived from a portion of the electron beam with a linear density $\lambda$ in the beam reference system, according to the hypothesis of the absolute reference system. The maximum value of the calculated electric field at a height of $y$, at the observation point, that is, at the sensor, will be when the sensor is above the mean of the linear charge distribution. Since the sensor is stationary in the laboratory, the calculation of electric field is derived from Maxwell’s equations. So, it is calculated according to the relation:

$$E_M = \frac{\lambda}{2\pi\epsilon_o y} \hat{j}$$

(26)

where the index $M$ means that the electric field is calculated according to Maxwell’s equations and $\hat{j}$ is a unit vector in the $Y$-direction.

The electric potential difference between the $y_1$ and $y_2$ sensor edges is calculated as follows:

$$V_{\text{max}} = \frac{\lambda}{2\pi\epsilon_o} \left[ \ln\left(\frac{y_{\text{max}}}{y}\right) \right]$$

(27)

so for $y_1 = y$ and $y_2 = y + 14$ cm this relation becomes:

$$V_{\text{max}} = \frac{\lambda}{2\pi\epsilon_o} \ln\left(\frac{y + 14\text{ cm}}{y}\right)$$

(28)

6. Positron-Electron Annihilation Process

An experiment implemented in order to experimentally investigate relativistic kinematics in an undergraduate student laboratory is the article in [14], and it is showed that it is possible to use the flight positron annihilation for the experimentally study of the special relativistic kinematic relation between momentum and energy.

In this article we give the theoretical background for this process using the kinematics of the hypothesis of the absolute reference system. We consider a positron with kinetic energy $E$ and momentum $p$ that annihilates with an electron at rest with the emission of two photons as shown in Figure 6. The energy and the momentum of the emitted photons are $E_i, p_i$ and $E_2, p_2$. The energy and momentum relations of the emitted photons are $E_i = p_i c = \nu_i$ and $E_2 = p_2 c = \nu_2$, where $\nu_i, \nu_2$ are the frequencies of them and the corresponding kinetic energies are $\frac{1}{2}h\nu_1 = \frac{1}{2}E_i, \frac{1}{2}h\nu_2 = \frac{1}{2}E_2$.

The total kinetic energy of the positron is $E = (1/2)mc^2$ and because the frequencies of the bound photons measured with the clock of the inertial reference system of the laboratory are $\gamma$ times smaller than the corresponding ones measured in the reference system of the positron, the contribution of positron to energy of the emitted photons is equal to $E_i = E/\gamma = (1/2)mc^2$, while the total kinetic energy of the electron that is at rest in the laboratory is equal to $(1/2)mc^2$. Conservation of momentum and energy give
Figure 6. Energy and momentum of the positron-electron annihilation in flight.

\[ p = p_1 + p_2 \]  \hspace{1cm} (29)

\[ E_1 + \frac{1}{2} mc^2 = \frac{1}{2} E_1 + \frac{1}{2} E_2 \]  \hspace{1cm} (30)

Squaring both sides of the Equation (29) we get

\[ p^2 = p_1^2 + p_2^2 + 2 \rho_1 \rho_2 \cos \theta \]  \hspace{1cm} (31)

Using the previous relations and the relation \( \left( \frac{\gamma m}{c^2} \right)^2 = p^2 c^2 + m^2 c^4 \), which is valid in relativistic kinematics and in kinematics of hypothesis of absolute reference system, and after some simple algebraic calculations, we get the following relation,

\[ \frac{1}{E_1} + \frac{1}{E_2} = \frac{1 - \cos \theta}{mc^2} \]  \hspace{1cm} (32)

which is the same as that derived from relativistic kinematics. Also, the last equation is the main result that was experimentally verified in [14].

7. Conclusion

In all experiments studied in this article on the basis of the hypothesis of an absolute reference system, it appeared unequivocally that the experimental results are in excellent agreement with the theoretically predicted results of this hypothesis. There are further similar issues to be discussed, but the most of them have already been discussed in the initial publications of the hypothesis of the absolute reference system. In the present study some issues discussed in order to test the validity of this hypothesis. However, as much as possible experimental research is suggested, mainly taking as many as possible experimental data on the speed and energy values of the Bertozzi’s experiment, as the results of this experiment have a one-sided interpretation in the scientific world, in order for us to reach scientific conclusions, fully accepted by the scientific community, regarding to the agreement of the hypothesis of absolute reference system with this experiment.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References


A Complex Algorithm for Solving a Kind of Stochastic Programming

Yunpeng Luo*, Xinshun Ma

Department of Mathematics and Physics, North China Electric Power University, Baoding, China
Email: *15176368820@163.com, maxs@sina.com

Abstract
Considering that the probability distribution of random variables in stochastic programming usually has incomplete information due to a perfect sample data in many real applications, this paper discusses a class of two-stage stochastic programming problems modeling with maximum minimum expectation compensation criterion (MaxEMin) under the probability distribution having linear partial information (LPI). In view of the nondifferentiability of this kind of stochastic programming modeling, an improved complex algorithm is designed and analyzed. This algorithm can effectively solve the nondifferentiable stochastic programming problem under LPI through the variable polyhedron iteration. The calculation and discussion of numerical examples show the effectiveness of the proposed algorithm.

Keywords
Stochastic Programming with Recourse, Probability Distribution with Linear Partial Information, Maximized Minimum Expectation, Complex Algorithm

1. Introduction
The stochastic programming with recourse, as an important method for solving optimization problems with uncertain parameters, was first proposed by G. Dantzig, the founder of linear programming. In the design of the optimal number of airline flights, he first considered a two-stage stochastic programming problem with recourse [1], considering the randomness of the passenger flow. With a lot of research by scholars such as G. Tinter and D. Walkup [2] [3], the theory and application of stochastic programming have been systematically developed, gradually showing its advantages in practical applications. Stochastic programming has the characteristics of many variables, many constraints, and large-scale problems. Therefore, solving stochastic programming problems has always been...
a difficult point, and it is also one of the problems that many scholars at home and abroad are keen to study. The dual decomposition L-algorithm given in [4] is considered to be one of the classical algorithms for solving stochastic programming problems. This method is a cutting plane method which belongs to the external linear approximation. By using the constraints of feasibility cutting and optimal cutting, it gradually reduces the feasibility area, and finally makes the algorithm converge to obtain the optimal solution. With the birth of the L-algorithm, a large number of researches have focused on the improvement of the L-algorithm, including the improvement of the simplicity multiplier, the improvement of the optimal cutting scheme and so on. The research of L-algorithm is becoming more and more mature [5] [6].

Generally, in the study of the stochastic programming with recourse, the second stage function with recourse is determined by the expectation criterion on the premise that the probability distribution of the random variable has complete information, so that the stochastic programming problem is equivalent to a definite mathematical programming problem. However, in practical problems, due to the lack of historical data and the limitation of statistical methods, the probability distribution information of random variables is not easily obtained, and only partial information may be obtained. Thus, the classic stochastic programming algorithms are no longer applicable. In order to solve this problem, the linear partial information theory (LPI) was proposed in reference [7], which can determine the fuzzy variables through the form of linear constraints. This method provides ideas for dealing with the problem of incomplete random variable information in stochastic programming. In reference [8], the paper introduced in detail the use of $\alpha$-cut technology to transform the probability constraints with fuzzy relations in stochastic programming into LPI form, and obtained the stochastic programming model under the linear partial information probability distribution.

Considering that the stochastic programming problem is transformed into the corresponding equivalent problem, the problem can be regarded as a kind of nonlinear programming problem, which can be solved by nonlinear programming method. With the continuous development of nonlinear programming methods, a large number of nonlinear programming methods have been applied to stochastic programming problems. In reference [9], the paper studied the quadratic stochastic programming model with recourse, which was transformed into quadratic programming model by B-regular [10] and semi smooth concept [11] [12]. Finally, quasi Newton method was introduced to solve the stochastic programming problem. The complex method [13] [14] is a kind of variable polyhedron algorithm that can determine the search direction only by comparing the value of the objective function to solve the constrained nonlinear programming problem. This method is simple and easy to implement for the requirements of the function. It can effectively solve the problem that the stochastic programming model with probability constraints can’t be derived. Therefore, the
complex method becomes a solution to the two-stage stochastic programming problem with uncertain probability distribution.

Aiming at the problem of stochastic programming under the uncertain probability distribution, this paper discusses a kind of stochastic two-stage programming model based on the maximal minimum expectation criterion under LPI based on the literature [8] [13] [14], which is a robust decision-making model under the linear partial information probability distribution and has high practical value. Considering the discreteness of the random variables and the uncertainty of the probability distribution in the model, this paper introduces the complex method and designs a stochastic programming algorithm based on the complex method. The validity of the algorithm is verified by solving the examples.

2. The Stochastic Programming Model with Recourse under LPI

Let $\left( \Omega, 2^\Omega, P \right)$ be a probability space, where $\Omega = \{\omega_1, \omega_2, \cdots, \omega_l\}$ is a finite sample space, $2^\Omega$ is the power set of sample space $\Omega$, and $P = (p_1, p_2, \cdots, p_l)^T$ is the probability distribution corresponding to sample set $\Omega = \{\omega_1, \omega_2, \cdots, \omega_l\}$. That is, $P_i = \Pr(\{w = x_i\}), i = 1, \cdots, l$, where $\Pr(\theta)$ is the probability function of event $\theta$, $P(\Omega) = 1$. In reference [9], the following stochastic programming problems are considered:

\[
\begin{align*}
\min_{x \in \mathbb{R}^n} & \quad f(x) + g(y) \\
\text{s.t.} & \quad Cx \leq b,
\end{align*}
\]

(1)

where,

\[
g(y) = \mathbb{E}_f(\phi(x, \omega))
\]

\[
\phi(x, \omega) = \max_{y \in \mathbb{R}^m} -\frac{1}{2} y^T H y + (\sigma(\omega) - x)^T y
\]

(2)

Here, $x \in \mathbb{R}^n, y \in \mathbb{R}^m$ are the decision variables in the first and second stages, $H \in \mathbb{R}^{m \times m}$ is a symmetric positive definite matrix, $\sigma(\omega) \in \mathbb{R}^m$ is a random variable in space $\Omega$, $C \in \mathbb{R}^{k \times n}, b \in \mathbb{R}^k, W \in \mathbb{R}^{r \times m}, q \in \mathbb{R}^{r \times l}$ are all known coefficient matrices, $f(x)$ is a convex function with $x$ as the decision variable, and $g(y)$ is the second stage function with recourse.

Assuming that the random variables in the model are finitely discrete, the second stage compensation function $g(y)$ can be expressed as

\[
g(y) = \sum_{i=1}^l p_i\phi(x, \omega_i)
\]

(3)

The establishment of the above model is based on the assumption that the probability distribution information of the random variables in the model is complete, that is, $P = (p_1, p_2, \cdots, p_l)^T$ is completely determined. But due to the limitations of historical data, such complete probability distribution information
is not easy to obtain. Based on the literature [8], the paper considers comprehensively the structure of the stochastic programming model and its application scenarios in practical problems, and makes the following assumptions about the probability distribution of random variables in stochastic programming:

Suppose that the probability distribution information of random variables aren’t known completely, but have linear partial information, that is, the following constraint condition is satisfied:

$$
\varphi = \left\{ P = (p_1, \cdots, p_l)^T \in R^l \mid BP \leq d, \sum_{i=1}^{l} p_i = 1; p_i \geq 0, i = 1, \cdots, l \right\}
$$

(4)

In the formula, $B \in R_{m \times l}$ and $d \in R^m$ are both known matrices.

From the above assumption, it can be concluded that the solution space $\varphi$ composed of LPI (P) of probability distribution $P$ of random variables is a bounded convex polyhedron. The value on this convex polyhedron $\varphi$ is the probability distribution of random variables in the model.

Since the probability distribution of random variables has linear partial information, simply using the expectation criterion to determine the second stage function with recourse will no longer be applicable. The paper expands the second stage function with recourse of the model, and combines the maximal minimum expectation criterion in the expectation model to give the two-stage stochastic programming model with recourse under the LPI discussed in the paper:

$$
\begin{align*}
\min_{x \in R^n} & \quad f(x) + \max_{P \in \varphi} \sum_{i=1}^{l} p_i \varphi(x, \omega_i) \\
\text{s.t.} & \quad Cx \leq b,
\end{align*}
$$

(5)

where,

$$
\varphi(x, \omega) = \max_{y \in R^m} -\frac{1}{2} y^T Hy + (\sigma(\omega) - x)^T y \\
\text{s.t.} & \quad Wy \leq q
$$

(6)

$$
\xi = \left\{ P = (p_1, \cdots, p_l)^T \in R^l \mid BP \leq d, \sum_{i=1}^{l} p_i = 1; p_i \geq 0, i = 1, \cdots, l \right\}
$$

(7)

Models (5)-(7) are the stochastic programming models with linear partial information probability distributions given in the paper. It can be seen that this model is a generalized form of the stochastic programming model in reference [9]. When $P$ takes a certain value in $\varphi$, that is, the probability distribution information of the random variable is complete, the model is consistent with the model in [9]. The biggest difference between the two models is that the probability distribution information assumed in the paper is incomplete, so when determining the function with recourse in the second stage, the paper uses the method of maximizing the expectation of the compensation value. This strategy is a robust choice, which can ensure that the optimization goal of the final decision-making scheme is not worse than the optimization goal of any possible situation. This is a conservative and robust decision model.
Because the second stage compensation function \( \max_{i \in C} \sum_{p \in \mathcal{P}} p_i \phi(x, w) \) is not differentiable, the gradient information of the model does not exist, and the previous gradient-based method will not be applicable. In order to solve the two-stage stochastic programming problem with linear partial information probability distribution given in this paper, the complex optimization algorithm based on direct optimization method is introduced. By improving the complex method, it is adapted to the solution process of the model, and then a stochastic programming algorithm based on the improved complex method under the uncertain probability distribution is given. Then, several examples are used to verify the effectiveness of the designed model and the algorithm.

3. Complex Method

As a direct optimization algorithm, the complex method is simple and easy to implement, so it is widely used in engineering optimization problems [15] [16]. This method can be regarded as the variable polyhedron method derived from simplex method. The biggest difference between the method and simplex method is that this method can be directly used to solve the optimization problem with constraints, and does not limit the number of vertices of complex shape, so it is more widely used than simple method. Assuming that the variables of the optimization problem are in the \( n \)-dimensional space, the complex shape with iteration in the complex method is a polyhedron composed of more than \( n + 1 \) vertices, which is formed by the combination of multiple simplexes. Considering the nondifferentiability of the stochastic programming model established in this paper, the complex method is introduced into the solution of stochastic programming under LPI, and the optimal value of the problem is obtained by using the variable polyhedron iterative process of the complex method in the optimization.

The complex method is an optimization method that only needs to compare the objective value of the optimization function to determine the optimization direction. Its basic idea is that we should first construct an initial complex shape in the feasible region. Then by comparing the objective function values of each vertex, we can find a new point in the feasible region where the objective function values are improved, and use it to replace the vertices with poor objective function values to form a new complex shape. By repeating the above process, the complex shape is continuously deformed, transferred and shrunk, gradually approaching the best. When the objective function value of each vertex in the complex shape is not much different or the distance between each vertex is very close, the vertex with the lowest objective function value can be regarded as the best [17] [18]. The following is a detailed description of the iterative process of the complex method.

In \( n \)-dimensional space, a polyhedron composed of \( k \geq n + 1 \) points is called a complex shape. Referring to the previous literature, there are two main methods to generate initial complex shape: manual definition of initial complex shape and random generation of initial complex shape. Considering the complexity of the stochastic programming model, the paper uses the second method.
The following is the specific operation of randomly generating the initial complex shape:

1) Suppose that the vertices of the complex shape are \( n \)-dimensional, the number of vertices of the initial complex shape is determined to be \( k \), and an initial vertex is selected manually in a given feasible region;

2) Suppose that the upper and lower bounds of the vertices of the complex shape are \( \upb \in \mathbb{R}^n, \lob \in \mathbb{R}^n \) respectively, where \( \upb \) are the upper bounds of the vertices and \( \lob \) are the lower bounds of the vertices. Then the remaining \( k-1 \) vertices are generated by using the random number in \([0,1]\). The build rule is \( x_i = \lob + r_i (\upb - \lob) \), where \( r_i \) is the random number in interval \([0,1]\), \( i = 2, \ldots, k \);

3) Check whether the generated \( k \) vertices are in the feasible region: assuming that \( w \) vertices are in the feasible region and the remaining \( k-w \) vertices are not in the feasible region, the \( k-w \) vertices that are not in the feasible region can be translated into the feasible region by the following methods:

a) The geometric centers of \( w \) vertices in the feasible region are calculated and recorded as \( x_{gc} = \frac{1}{w} \sum_{i=1}^{w} x_i \);

b) If \( k-w \) vertices that are not in the feasible region are recorded as \( x_{out,j} \), \( j = 1, \ldots, k-w \), then a vertex \( x'_{out,j} \) in the feasible region can be found on the line between \( x_{gc} \) and \( x_{out,j} \). The specific searching method is as follows:

\[
x'_{out,j} = x_{gc} + \rho (x_{out,j} - x_{gc}), \quad \rho \in (0,1), \quad j = 1, \ldots, k-w
\]  

If the result \( x'_{out,j} \) is not in the feasible region, the formula \( \rho = 0.5\rho \) can be used to continuously reduce \( \rho \) until the vertex is translated into the feasible region. Through the above steps, we can get the initial complex shape that meets the conditions.

In the generated complex shape, let the worst point be recorded as \( x^b \), the secondary bad point as \( x' \), and the best point as \( x' \). The centroid of other vertices with the worst points removed in the complex shape is calculated by formula \( x' = \frac{1}{n} \sum_{i \neq h} x_i \), which is recorded as \( x' \). In the process of the complex shape optimization, several methods of vertex transformation for polyhedron in the iterative process are as follows:

1) Mapping method:

Transformation thought: We expect to find a better value in the opposite of the worst point \( x^b \), to replace \( x^b \).

Search direction: It searches along the direction from the worst point \( x^b \) to the centroid \( x' \), i.e. along the direction of \( x^b \rightarrow x' \).

Step factor: Mapping factor \( \alpha : \alpha > 1 \), representing the step size of the mapping.

Mapping iteration formula: \( x' = x' + \alpha (x' - x^b) \), where \( x' \) is called mapping point.

Rule of judgement: If \( x' \) is in the feasible region and \( f(x') < f(x^b) \), \( x' \)
will be used instead of $x^h$ to form a new complex shape and carry out the next iteration.

2) Expansion method:

Transformation thought: According to the advantages and disadvantages of mapping point $x'$ obtained by mapping method, we expect to get better transformation vertices. If the function value of the mapping point is less than the function value of the best $x'$, i.e. $f(x') < f(x')$, then the direction from $x'$ to $x'$ is the current optimal direction and it can be expanded in this direction.

Expansion iteration formula: $x' = x' + \beta \left(x' - x^h\right)$.

Expansion coefficient $\beta: \beta \geq 1$.

Rule of judgement:

a) if $f(x') < f(x')$, the expansion is successful, and $x'$ replaces $x^h$ to form a new complex shape.

b) If $f(x') > f(x')$, expansion fails, and $x'$ replaces $x^h$ to form a new complex shape.

3) Shrinkage method:

Transformation thought: If $f(x^h) < f(x')$ in the mapping method, it indicates that the step size of the mapping method is too large, let $\alpha = 0.5\alpha$, and we repeat the mapping method. If it still fails until $\alpha < 10^{-5}$, it indicates that the current optimization direction is not right. In this case, shrinkage method is considered to find the search direction in the complex shape.

Shrinkage direction: through the failure of the mapping method, it shows that the optimization direction $x^h \rightarrow x'$ of the mapping method is not correct, so the complex shape is shrunk along the direction from the center of the centroid $x'$ to the worst point $x^h$, i.e. along the direction of $x' \rightarrow x^h$.

Shrinkage coefficient: $\gamma: 0 < \gamma < 1$.

Shrinkage formula: $x^h = x' - \gamma \left(x' - x^h\right)$.

Rule of judgement: If $f(x^h) < f(x')$, we use shrinkage point $x^h$ to replace the worst point $x^h$ to form a new polyhedron; If the shrinkage fails, we carry out the compression step.

4) Compression method:

Transformation thought: shrinkage failure means that the effect of iteration points in the search direction composed of the most nearly $x^h$ and the center of mass $x'$ is not good. In this case, we generally compress the compound shape to the best point $\mu'$, so as to find the compound shape with good performance.

Compression formula: $x' = x' + \delta \left(x' - x^h\right)$, $i = 1, \ldots, k$, use this formula to replace all points except the best point in the current composite shape.

Compression factor: $\delta: 0 < \delta < 1$.

The basic thought of the complex method is to change the complex shape step by step through continuous iteration, so that the final approximation of complex shape can be compressed to the optimal solution, and the iteration can be completed [19]. Therefore, the termination condition of the complex method is giv-
en here, that is
\[
\sqrt{\frac{1}{k} \left[ \sum_{i=1}^{k} \left[ f(x^i) - f(x') \right]^2 \right]} \leq \varepsilon
\]
(9)

where \( x' = \frac{1}{k} \sum_{i=1}^{k} x_i, i = 1, \ldots, k \).

The following is the specific steps of the complex method: Set the parameter \( \alpha, \beta, \gamma, \delta \), and the convergence parameter \( \varepsilon > 0 \). The number of vertices of the complex shape is determined. If the decision variable is \( n \)-dimensional, the number of vertices of the complex shape should be between \( n + 1 \) and \( 2n \).

1) Generate the initial complex shape. The steps of generating the initial complex shape by using the random method given in this paper are used to get the initial complex shape satisfying the requirements;

2) Calculate the function value of each vertex in the current complex shape, and sort out the worst point \( x^w \), the secondary bad point \( x^s \), the best point \( x^l \), and calculate the centroid \( x^c \) of the current complex shape;

3) According to the mapping coefficient \( \alpha \) and the mapping formula, the mapping point \( x' \) is calculated:
   a) If the mapping point \( x' \) is within the feasible region, step 4) is carried out;
   b) If the mapping point \( x' \) is not in the feasible region, we reduce the mapping coefficient \( \alpha \), that is \( \alpha = 0.5 \alpha \), and then repeat step 3);

4) Calculate the function value of the mapping point \( x' \), and compare the function value of \( x' \) with the vertex of the current complex shape:
   a) If \( f(x') < f(x^w) \), the expansion step is carried out. Using the expansion formula, the expansion point \( x' \) can be got. If \( f(x') < f(x') \), then we replace \( x^w \) with \( x' \) to get a new polyhedron, and carry out step 6); otherwise, we replace \( x^w \) with \( x' \) to get a new polyhedron, and carry out step 6);
   b) If \( f(x') < f(x') < f(x^w) \), \( x' \) is used instead of \( x^w \) to get a new polyhedron, and step 6) is carried out;
   c) If \( f(x') \geq f(x^w) \), compare the value of mapping coefficient \( \alpha \) : if \( \alpha > 10^{-5} \), we reduce \( \alpha \), and set \( \alpha = 0.5 \alpha \). Then step 3) is carried out; otherwise, we carried out the contraction step of the complex method, and use the contraction formula \( x^c = x' - \gamma(x' - x^w) \) to calculate the contraction point \( x^c \).

Then step 5) is carried out;

5) Compare the function values of the contraction point and the worst point \( x^w \) : if \( f(x^c) < f(x^w) \), we replace \( x^w \) with \( x^c \) to get a new polyhedron, and carry out step 6); otherwise, the compression step of the complex method is carried out to get a new complex shape. Then step 2) is carried out;

6) Judge whether the current complex shape meets the termination condition
\[
\sqrt{\frac{1}{k} \left[ \sum_{i=1}^{k} \left[ f(x^i) - f(x') \right]^2 \right]} \leq \varepsilon .
\]
If it does, we stop the iteration. At this time, the best solution is the best solution and the best function value is the best value. Otherwise, step 2) is carried out.

Through the concrete steps of the complex method, the nonlinear program-
ming problem can be solved. The stochastic programming problem under LPI proposed in this paper can also be regarded as a nondifferentiable nonlinear programming problem. Therefore, the paper innovatively introduces the complex method into the solution of the model, which provides a feasible way for the stochastic programming algorithm under the uncertain probability distribution.

4. Example Analysis

Combined with the compensation two-stage stochastic programming model (5) - (6) given above, this paper presents a complex method of decision variable $x \in R^6$ to solve the stochastic programming model. At the same time, in view of the different probability distribution information of random variables, the paper discusses the examples according to different probability distribution information, so as to compare and analyze the two-stage stochastic programming model under different probability distribution information.

In the model (5) - (6), $f(x)$ in the first stage is a general convex function form. Here, it is set as a quadratic function form in the calculation example, in which the decision variable is $x \in R^6$. As for the random variable in the compensation function of the second stage, the capacity of the calculation example is set to 7, i.e. $l = 7$, so $P = (p_1, \cdots, p_7)^T \in R^7$. Therefore the paper considers the following stochastic programming problems:

$$\begin{align*}
\min_{x \in R^6} & \quad \frac{1}{2} x^T Ax + D^T x + \max_{p \in \mathcal{P}} \sum_{i=1}^{7} p_i \phi(x, \omega_i) \\
\text{s.t.} & \quad Cx \leq b,
\end{align*}$$

(10)

$$\begin{align*}
\phi(x, \omega_i) = & \max_{y \in R^6} \quad -\frac{1}{2} y^T Hy + (\sigma(\omega_i) - x)^T y \\
\text{s.t.} & \quad Wy \leq q
\end{align*}$$

(11)

The parameters of correlation matrix and variables used in the model are: $A = \text{diag}(2, 2, 3, 1, 2, 1)$, a diagonal matrix; $H \in R^{3 \times 3}$, a unit matrix; other parameters of correlation matrix are as follows:

$$D = \begin{pmatrix} 2 \\ 3 \\ 1 \\ 4 \\ 2 \\ 1 \end{pmatrix}; \quad C = \begin{pmatrix} 3 & 1 & 0 & 2 & 1 & 3 \\ 1 & 1 & 2 & 0 & 1 & 2 \\ 2 & 3 & 1 & 4 & 0 & 3 \end{pmatrix}; \quad b = \begin{pmatrix} 12 \\ 5 \\ 20 \end{pmatrix};$$

$$W = \begin{pmatrix} 1 & 0 & 2 & 1 & 1 & 3 \\ 2 & -1 & 0 & 3 & 1 & 2 \\ 3 & 2 & 1 & 0 & 1 & 1 \end{pmatrix}; \quad q = \begin{pmatrix} 7 \\ 7 \\ 7 \end{pmatrix}.$$  

In the paper, the corresponding value of random variable is fixed, and the probability of occurrence of random variable is uncertain information, that is, the probability of occurrence of random variable is variable. In order to make
the example more universal, the value of the random variable 
\[ \sigma(w_i) = (\sigma^1(w_i), \sigma^2(w_i), \sigma^3(w_i), \sigma^4(w_i), \sigma^5(w_i), \sigma^6(w_i))^T, i = 1, \cdots, 7 \] 
is generated by a random number with lower bound \( (1, 2, 3, 4, 5, 6)^T \) and upper bound \( (6, 7, 8, 9, 10, 11)^T \), and determined. The value of \( \sigma(\omega) \) is

\[
\begin{pmatrix}
3.0851 & 5.6016 & 3.0006 & 5.7338 & 6.4617 \\
5.8351 & 4.7362 & 7.8634 & 7.5741 & 8.4886 & 7.0804 \\
5.4643 & 3.6599 & 7.1061 & 4.2085 & 5.5383 & 8.9753 \\
1.3815 & 5.8996 & 5.1920 & 7.6173 & 9.8899 & 8.6925 \\
\end{pmatrix}
\]

Combined with the matrix parameters of the model given above, according to the completeness of the probability distribution information of the designed random variables, the paper analyzes and discusses the examples in three cases.

**Case (1):**

It is assumed that the probability distribution of random variables involved in the model does not have too much effective information, and only has the following linear partial information constraints:

\[
\xi = \left\{ P = (p_1, \cdots, p_7)^T \in \mathbb{R}^7 \mid \sum_{i=1}^{7} p_i = 1, p_i \geq 0, i = 1, \cdots, 7 \right\}
\]

This means that the occurrence of random variables in the case is accidental, and we cannot know the exact value of random variables in the case. For such a specific problem, we use the robust decision-making scheme designed in this paper to find the optimal decision-making result under the condition of maximizing the compensation function, so as to ensure that the actual result will not be worse than the expected decision-making result.

In this paper, the first initial point of the initial complex shape is taken as \( x_0 = (0, 0, 0, 0, 0)^T \), the number of vertices of the complex shape is set as 12. As the paper introduces in the vertex transformation method of complex shape, the mapping coefficient \( \alpha > 1 \), expansion coefficient \( \beta \geq 1 \), contraction coefficient \( 0 < \gamma < 1 \), the compression coefficient \( 0 < \delta < 1 \) and the smaller the convergence parameter \( \varepsilon \), the higher the accuracy of the algorithm. Therefore these parameters adopted in the complex method are respectively taken as \( \alpha = 1.3 \), \( \beta = 1 \), \( \gamma = 0.7 \), \( \delta = 0.5 \), \( \varepsilon = 10^{-6} \). Through the operation of the program, the iterative process is shown in Table 1.

As shown in the above table, the algorithm stops iteration after the 448th time, and the optimal solution \( x = (-2.1646, 0.7194, -0.3065, -0.4003, 1.3779, -0.7288)^T \) is obtained. At the same time, the optimal value \( W^* = 62.2188 \) of stochastic programming is obtained. In order to more intuitively explain the iterative process of the complex method in solving stochastic programming, the paper presents the iterative graph of the optimal value changing with the number of iterations \( w \), as shown in Figure 1.
Figure 1. Iterative figure of optimal value.

Table 1. Iterative process of optimal solution.

<table>
<thead>
<tr>
<th>Iteration times $w$</th>
<th>optimal solution $x$</th>
<th>optimal value $W$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$(-1.7197, 1.5025, 0.4358, 1.4467, 2.7328, -2.2133)$</td>
<td>71.8494</td>
</tr>
<tr>
<td>2</td>
<td>$(-1.7197, 1.5025, 0.4358, 1.4467, 2.7328, -2.2133)$</td>
<td>71.8494</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>19</td>
<td>$(-1.1580, 2.5627, -0.4518, -0.0830, 1.2015, -2.5338)$</td>
<td>69.2828</td>
</tr>
<tr>
<td>20</td>
<td>$(-0.0872, 1.3478, -0.0326, 0.7095, 1.5657, -0.2709)$</td>
<td>67.9572</td>
</tr>
<tr>
<td>21</td>
<td>$(-0.0872, 1.3478, -0.0326, 0.7095, 1.5657, -0.2709)$</td>
<td>67.9572</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>447</td>
<td>$(-2.1646, 0.7194, -0.3065, -0.4003, 1.3779, -0.7288)$</td>
<td>62.2188</td>
</tr>
<tr>
<td>448</td>
<td>$(-2.1646, 0.7194, -0.3065, -0.4003, 1.3779, -0.7288)$</td>
<td>62.2188</td>
</tr>
</tbody>
</table>

It can be seen that the optimal value of the model gradually decreases with the increase of the number of iterations, and keeps approaching to the optimal solution. The final optimal value converges to 62.2188, which shows that for the solution of stochastic programming, the complex method has good convergence and the designed algorithm is effective.

Case (2):

Compared with case (1), we set the probability distribution information of random variables in case (2) more complete, and its probability distribution has some linear constraint information. Under the constraint of case (1), case (2) supposes that the probability distribution of random variables has the following linear constraints:

$$
\begin{align*}
    p_1 + p_2 + p_3 &\leq \frac{1}{2} \\
    p_4 + p_5 &\leq \frac{1}{3} \\
    p_6 + p_7 &\leq \frac{1}{3} \\
    \frac{1}{9} &\leq p_8 \leq \frac{1}{5}
\end{align*}
$$
Let \( B = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \), \( d = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{3} \\ \frac{1}{5} \\ \frac{1}{9} \end{bmatrix} \), then the probability distribution of random variables in case (2) has the following linear partial information:

\[
\xi = \left\{ P = (p_1, \cdots, p_7)^T \in \mathbb{R}^7 \mid BP \leq d, \sum_{i=1}^{7} p_i = 1, p_i \geq 0, i = 1, \cdots, 7 \right\}
\]

In this case, the other relevant parameters set in case (1) are kept unchanged.

The complex method is used to solve the stochastic programming problem in case (2), and the robust decision scheme and result in case (2) are given. The results of the iterative process are shown in Table 2.

The program is terminated after 431 iterations, and the optimal solution \( x = (-2.0086, 0.6482, -0.4208, -0.7191, 0.9701, -0.2265) \). At this time, the optimal value of stochastic programming problem is obtained, that is \( W^2 = 56.1144 \). It can be seen that the optimal value of case (2) is better than that of case (1), which shows that when the probability distribution information of random variables is more complete, the decision result is better. The optimal value of the model changes with the number of iterations \( w \), as shown in Figure 2.

**Case (3):**

In order to compare the influence of the completeness of the probability distribution information of the random variables on the decision result, the probability of the random variables in case (3) is set as a fixed value. Next, the other parameters of the stochastic programming model are consistent with the situations (1) and (2), and the probability distribution of the random variables is set as \( P = \begin{bmatrix} \frac{3}{25} & \frac{3}{25} & \frac{1}{25} & \frac{3}{25} & \frac{1}{25} & \frac{3}{25} & \frac{1}{25} \end{bmatrix} \), that is, the example in this paper is strengthened to the classical stochastic programming model. In this paper, the result of case (3) obtained by the complex method under the condition that the probability information of the random variables is complete is shown in Table 3.

The experimental result shows that the program ends after 385 iterations. The optimal solution and the optimal value of the example are: \( x = (-1.6394, 0.1992, -0.1810, -1.0080, 0.5954, -0.6059) \), \( W^3 = 45.1761 \), respectively. At this time, the optimal value of case (3) is far less than that of case (1) and case (2), which also shows that when the probability distribution information of random variables in the stochastic programming problem is complete, the better decision result can be obtained. The trend chart of the optimal value iteration in case (3) is shown in Figure 3.

In order to illustrate the significance of stochastic programming model under uncertain probability distribution in reality, the paper brings the optimal solution of case (3) into the objective function of case (1), and the optimal value difference value between them is \( 64.3512 - 62.2188 \), that is 2.1324; similarly, the
Table 2. Iterative process of optimal solution.

<table>
<thead>
<tr>
<th>Iteration times $w$</th>
<th>optimal solution $x$</th>
<th>optimal value $W'$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$(0, 0, 0, 0, 0, 0)$</td>
<td>63.7740</td>
</tr>
<tr>
<td>2</td>
<td>$(0, 0, 0, 0, 0, 0)$</td>
<td>63.7740</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>29</td>
<td>$(0.1711, 0.3808, -0.1774, 0.7904, 0.3968, 0.2465)$</td>
<td>63.5492</td>
</tr>
<tr>
<td>30</td>
<td>$(0.1711, 0.3808, -0.1774, 0.7904, 0.3968, 0.2465)$</td>
<td>63.5492</td>
</tr>
<tr>
<td>31</td>
<td>$(-1.3699, 1.2283, -0.2804, 0.6253, -0.0063, -1.1811)$</td>
<td>60.6809</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>431</td>
<td>$(-2.0086, 0.6482, -0.4208, -0.7191, 0.9701, -0.2265)$</td>
<td>56.1144</td>
</tr>
<tr>
<td>431</td>
<td>$(-2.0086, 0.6482, -0.4208, -0.7191, 0.9701, -0.2265)$</td>
<td>56.1144</td>
</tr>
</tbody>
</table>
optimal solution of case (3) is brought into case (2), and the optimal value difference value between them is 57.1422 - 56.1144, that is 1.0278. It can be seen that the difference values 2.1324 and 1.0278 are the loss value caused by the inaccuracy of the probability distribution information of the random variables when using the classical stochastic programming model. This also fully shows the significance of the stochastic programming model based on the maximum minimum expectation criterion under the uncertainty probability distribution in the actual problem. The model can effectively reduce the loss caused by decision-making in the face of the stochastic programming problem with incomplete information of the probability distribution of random variables.

5. Conclusion

Under the guidance of linear partial information theory, the stochastic programming model with uncertain probability distribution is established based on the maximum minimum expectation criterion. According to the nondifferentiability of the model, the paper designs a solution method based on the complex method. Finally, the solution algorithm is used to solve several specific examples, which show the value of the model in practical problems and the effectiveness of the designed solution algorithm.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References


Comparison of Ising Model and Potts Model on Non-Local Directed Small-World Networks

M. A. Sumour1*, M. Kh. Srour1, S. M. Baraka2, M. A. Radwan1, R. J. Khozondar1, M. M. Shabat3

1Physics Department, Faculty of Science, Al-Aqsa University, Gaza, Gaza Strip, Palestine
2National Institute of Aerospace, Hampton, VA, USA
3Department of Physics, Faculty of Science, Islamic University of Gaza, Gaza Strip, Palestine, Palestine

Email: *ma.sumoor@alaqsa.edu.ps

Abstract

Further to the investigation of the critical properties of the Potts model with $q = 3$ and 8 states in one dimension (1D) on directed small-world networks reported by Aquino and Lima, which presents, in fact, a second-order phase transition with a new set of critical exponents, in addition to what was reported in Sumour and Lima in studying Ising model on non-local directed small-world for several values of probability $0 < P < 1$. In this paper the behavior of two models discussed previously, will be re-examined to study differences between their behavior on directed small-world networks for networks of different values of probability $P = 0.1, 0.2, 0.3, 0.4$ and $0.5$ with different lattice sizes $L = 10, 20, 30, 40,$ and $50$ to compare between the important physical variables between Ising and Potts models on the directed small-world networks. We found in our paper that is a phase transitions in both Ising and Potts models depending essentially on the probability $P$.

Keywords

Ising Model, Potts Model, Directed Small World, Probability, Magnetization, Susceptibility

1. Introduction

Networks of coupled dynamics systems have been used to numerically model many self-organizing systems, such as biological oscillators, neural networks, spatial games and genetic control networks [1] and the references therein. With resourceful computing fascicles using modeling techniques such as Monte Carlo codes [2], the simulation of the small-world networks become doable [3] [4] showed that the Ising model and Potts Model on a small-world (SW) network presents a phase transition well defined at a finite temperature. [2] studied the
Potts model with \( q = 3 \) and 4 states on directed small-world networks (DSWN) as a function of temperature. [5] covered the Potts model for \( q = 4 \) and had found a first-order phase transition for values of the rewiring probability \( P = 0.1 \) and \( P = 0.9 \) that agree with the [6].

In the directed small world networks, each node is randomly reconnected with \( n \) edges with probability \( P \). In this paper, we employed values of probability as \((P = 0.01, 0.2, 0.3, 0.4, 0.5)\), and the lattice size \((L = 10, 20, 30, 40, 50)\), and adopted values from both Ising and Potts model in purpose of comparison between them.

According to reference [7], Potts model in two-dimension (2D) presents a phase transition at finite temperature \( T \), for any number of states \( q > 1 \), and as in reference [8]. Ising model on non-local directed small-world lattices has a second-order phase transition with new critical exponents dependent on \( p \) \((0 < p < 1)\).

Most of all of researches in Ising and Potts models studied the magnetization, susceptibility, fourth-order Binder Cumulant, and energy. All of these parameters were studied as a function of temperature. There parameters are determined and analyzed for both models [4] [8] to study the similarities and differences between them. The behavior of these models on the directed small-world networks for networks is covered.

2. Model and Simulations

In our simulations, values and shapes are adopted as in [9] these parameters are doable. We start our simulations for different values of \( L \) and \( P \) as follows: \( L = 10, 20, \) and \( 40 \) and probability \( P = 0.1, 0.2, 0.3, 0.4, \) & \( 0.5 \), with magnetization as a function of temperature (0 to 4) in Kelvin. As a case study \( P = 0.3 \), was chosen to be plotted herein. We shall now define all parameters on the two models as defined in statistical physics as listed below:

2.1. Ising Model on Non-Local Directed Small-World Networks

For Ising model we considered here the molar magnetization \( m \), as

\[
m = \frac{\sum \sigma_i}{L},
\]

where \( \sigma_i \) is a spin variable at each node of the network, and \( L \) is the length of a linear chain where.

The susceptibility

\[
\chi(T) = \frac{L}{T} \left[ \langle m^2 \rangle - \langle m \rangle^2 \right]_{\text{av}}
\]

and the fourth-order Binder Cumulant

\[
U_4(T) = 1 - \frac{\langle m^4 \rangle}{3 \langle m^2 \rangle^2}
\]

In the above equations \( \langle ... \rangle \) stand for thermodynamic averages and \( [... \] \)_{av}.
for averages over different realizations. To calculate the exponents of these models, we apply the finite-size scaling (FSS) theory [10]. We expect, for large system sizes, an asymptotic FSS behavior of the form:

\[
m = L^{-\beta/v} f_m \left( x \right) \left[ 1 + \cdots \right],
\]

(4)

\[
\chi = L^{-\gamma/v} f_\chi \left( x \right) \left[ 1 + \cdots \right],
\]

(5)

where \( \beta \) and \( \gamma \) are the usual critical exponents, and \( f_i \left( x \right) \) are FSS functions with

\[
x = (T - T_c) L^{1/v},
\]

(6)

being the scaling variable. The dots in the brackets \( \left[ 1 + \cdots \right] \) indicate corrections to scaling terms. We calculated the error bars from the fluctuations among the different realizations. Therefore, from the size dependence of \( M \) and \( \chi \), we obtain the exponents ratios \( \beta/v \) and \( \gamma/v \), respectively. The susceptibility at its maximum also scales as \( L^{1/v} \)

Moreover, the value of \( q \) for which \( \chi \) has a maximum, \( T^{x_{max}}_c = T_c \left( L \right) \), scales with the lattice size as:

\[
T_c \left( L \right) = T_c + b L^{-1/v}
\]

(7)

where \( b \) is a non-universal constant.

The correlation length exponent \( 1/v \) can be estimated from Equation (7).

### 2.2. Potts Model on Non-Local Directed Small-World Networks

For Potts model as in reference [8], we use the energy per spin, \( e = E/N \), and the magnetization per spin, \( m = M/N \) with \( M = (q \max [n_i] - N)/(q-1) \), were evaluated. Here \( n_i \leq N \) denotes the number of spins with “orientation” \( i = 1, \cdots, q \).

From the energy measurements, we can compute the average energy, specific heat, and energetic cumulant.

\[
u(T) = \left[ \langle e \rangle \right]_{av}
\]

(8)

\[
C(T) = \frac{N}{T^2} \left[ \left[ \langle e^2 \rangle \right] - \left[ \langle e \rangle \right]_{av}^2 \right]
\]

(9)

\[
B_e \left( T \right) = 1 - \frac{\left[ \langle e^2 \rangle \right]_{av}}{3 \left[ \langle e \rangle \right]_{av}^2},
\]

(10)

Similarly, we can derive from the magnetization measurements the average magnetization, the susceptibility, and the fourth-order magnetic Cumulant,

\[
m(T) = \left[ \langle m \rangle \right]_{av}
\]

(11)

\[
\chi(T) = \frac{L}{T} \left[ \langle m^2 \rangle \right]_{av} - \left[ \langle m \rangle \right]_{av}^2
\]

(12)

\[
U_4 \left( T \right) = 1 - \frac{\left[ \langle m^4 \rangle \right]}{3 \left[ \langle m \rangle \right]_{av}^2}
\]

(13)
In order to calculate the exponents for this model, we applied the finite-size scaling (FSS) theory, and for large system sizes we got [9]

\[ C = C_{\text{reg}}(T) + L^{\nu} f_{C}(x)[1 + \cdots], \]  
\[ \left[ \langle m \rangle \right]_{\text{av}} = L^{\beta} f_{m}(x)[1 + \cdots], \]  
\[ \chi = \chi_{\text{reg}}(T) + L^{\gamma} f_{\chi}(x)[1 + \cdots], \]

where \( C_{\text{reg}}(T) \) and \( \chi_{\text{reg}}(T) \) are regular temperature dependent background terms. \( \nu, \alpha, \beta, \) and \( \gamma \) are the usual critical exponents, and \( f_{i}(x), \) with \( i = C, m, \chi, p, \) are FSS functions with

\[ x = (K - K_c) L^{\nu}, \]  

being the scaling variable, and \( K = J/k_B T. \) We calculated the error bars from the fluctuations among the different realizations. Note that these errors contain both the average thermodynamic error for a given realization, and the theoretical variance for infinitely accurate thermodynamic averages which are caused by the variation of the quenched random geometry of the networks.

For both Ising and Potts models from the magnetization, we can investigate other measures such as the average magnetization, susceptibility and the fourth-order Binder Cumulant, for probability \( p = 0.5 \) to 1. and for Potts model the phase transition is always first order. For probability \( p = 0.5 \) to 1 and for Ising, the code doesn’t work well. And for both models and \( p = 1 \) the networks are not a small world, they are a complex network.

### 3. Results and Discussion

When all plots and probability values are done, the magnetization behavior shows a very small change. On the other hand, when the size of \( L = 10, 20, \) and 40 the changes in Figure 1 show that the magnetization of Potts model (left plots in Figure 1) decayed faster than Ising model (right plot in Figure 1), the values of magnetization varies slowly with temperatures after 2.1 kelvin.

Also, we determined the “flatness” of the curves of Potts and Ising model as \((-0.76), \) and \((-2.87) \) respectively, and we find that Ising and Potts model illustrates a continuous phase transition and the decay behavior of magnetization which agrees with magnetization universality.

Moreover, we plot the susceptibility as function of temperature (0 to 4) in Klein for \( L = 10, 20, \) and 40 and probability = 0.1, 0.2, 0.3, 0.4, 0.5, a case study of is presented as \( P = 0.3, \) and we take the log scale for the y-axis because there is variation in values. Figure 2 shows a shift in the peak as the lattice size \( (L) \) increases, the peak of \( L = 40 \) is taking the maximum peak of all values, and for Ising model the average peak is at 2.5 while for Potts model is at 1.1, and the shape of the peak for Potts model more sharp than the Ising model.

Figure 3 shows the behavior of energy versus temperature (0 to 4) for \( L = 10, 20, \) and 40 and probability = 0.1, 0.2, 0.3, 0.4, 0.5, \( P = 0.3 \) is again our sample plot, we can see a typical behavior attesting for a first-order phase transition in the...
range of all $P$ values, and we see a variation at the beginning point of Ising model at $x$-axis $= 1.39363$, but Potts model at $0.575685$ and intersection at the point $(x = 2.8783, y = -0.844846$ and we noticed that the energy of Potts model after temperature of $2$ Kelvin becomes more stable than the Ising model after $T = 2.5$ Kelvin.

Figure 4 shows the Fourth-order Binder Cumulant ($U_4$) as a function of Temperature (0 to 4) for several lattice sizes $L = 10, 20, and 40$ and probability $= 0.1, 0.2, 0.3, 0.4, & 0.5$ and we plot as a sample $P = 0.5$. Also, we notice a continuous phase transition for all values. We notice that for Potts model at $L = 40$ there is a sharp peak concave down at 1.9 temperature.

The actual values of the magnetization for each L to Ising and Potts model for the probability $P = 0.3$ as shown in Table 1, from these values we get Figure 5 which shows the difference between the values of magnetization of Ising and Potts model with lattice size for sample probability $P = 0.3$. Also, we can clearly see that the exponent is, within the errors, independent of $P$, in agreement with universality ideas [8].

In Table 1 and from our simulation that used huge computing time and space, the magnetization values for both Ising and Potts Model for three Lattices sizes (10, 20 and 40) is tabulated in Table 1.

Figure 5 shows that the values of magnetization of Ising model is larger and varies with the magnetization of the Potts model, but the values of magnetization of Potts model is stable.

**Table 1.** The values of magnetization for Ising and Potts model with different lattice sizes ($L$).

<table>
<thead>
<tr>
<th>$L$</th>
<th>Magnetization for Ising</th>
<th>Magnetization for Potts</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.90832949499980631</td>
<td>0.1172197000000800</td>
</tr>
<tr>
<td>20</td>
<td>0.91423680999999635</td>
<td>5.83670812499994876E-002</td>
</tr>
<tr>
<td>40</td>
<td>0.914494221875002</td>
<td>2.924515624998562E-002</td>
</tr>
</tbody>
</table>

Figure 1. Magnetization versus temperature for sizes $L = 10, 20,$ and $40$ with rewiring probability $p = 0.3$. 

Figure 1. Magnetization versus temperature for sizes $L = 10, 20,$ and $40$ with rewiring probability $p = 0.3$. 

DOI: 10.4236/jamp.2020.86080
Figure 2. Susceptibility versus temperature for sizes $L = 10, 20,$ and $40$ with rewiring probability $p = 0.3$.

Figure 3. Shows the behavior of energy versus temperature (0 to 4) for $L = 10, 20,$ and $40$ and probability $= 0.1, 0.2, 0.3, 0.4, 0.5, P = 0.3$.

Figure 4. Fourth-order Binder Cumulant ($U_4$) temperature for sizes $L = 10, 20,$ and $40$ with rewiring probability $p = 0.5$.

Figure 5. Lattice size ($L$) versus Magnetization ($M$) for sizes $L = 10, 20,$ and $40$ with rewiring probability $p = 0.3$. 

DOI: 10.4236/jamp.2020.86080 1036 Journal of Applied Mathematics and Physics
4. Conclusion

The Potts and Ising models are studied for different Lattice sizes and different probability on non-local directed small-world networks, including long-range interaction between spins depending on the probability $P$. We found a phase transitions in both Ising and Potts models depending essentially on the probability $P$. This behavior is the influence of long-range interactions that occur in the presence of $P$ directed bonds and also the number of the Potts model states on the directed SW network. The change in the universality of these models can be described as the influence of directed non local interactions that occur with the presence of $P$ directed bonds [11].

Acknowledgements

S. Baraka thanks the IAP-CNRS, France, for compiling the Ising Model on their supercomputing facilities. And authors would like to express thanks to the University of Natural Resources and Life Sciences (BOKU) at Vienna in Austria for giving our access to their supercomputer to do our simulation of our work.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

On Two Double Inequalities (Optimal Bounds and Sharps Bounds) for Centroidal Mean in Terms of Contraharmonic and Arithmetic Means

Mohammed El Mokhtar Ould El Mokhtar¹, Hamad Alharbi²

¹Qassim University, Al-Mulida, Qassim, KSA
²Shaqra University, Shaqra, KSA
Email: med.mokhtar66@yahoo.fr, M.labdi@qu.edu.sa, halharby@su.edu.sa

Abstract

This research work considers the following inequalities:

\[ \lambda A(a,b) + (1-\lambda) C(a,b) \leq \bar{C}(a,b) \leq \mu A(a,b) + (1-\mu) C(a,b) \]

and

\[ C[\lambda a + (1-\lambda) b, \lambda b + (1-\lambda) a] \leq \bar{C}(a,b) \leq C[\mu a + (1-\mu) b, \mu b + (1-\mu) a] \]

with \( A(a,b) = \frac{a+b}{2} \); \( C(a,b) = \frac{a^2 + b^2}{a+b} \), \( \bar{C}(a,b) = \frac{2(a^2 + ab + b^2)}{3(a+b)} \). The researchers attempt to find an answer as to what are the best possible parameters \( \lambda, \mu \) that (1.1) and (1.2) can be held? The main tool is the optimization of some suitable functions that we seek to find out. By searching the best possible parameters such that (1.1) and (1.2) can be held. Firstly, we insert \( f(t) = \lambda A(a,b) + (1-\lambda) C(a,b) - \bar{C}(a,b) \) without the loss of generality. We assume that \( a > b \) and let \( t = \frac{a}{b} > 1 \) to determine the condition for \( \lambda \) and \( \mu \) to become \( f(t) \leq 0 \). Secondly, we insert \( g(t) = \mu A(a,b) + (1-\mu) C(a,b) - \bar{C}(a,b) \) without the loss of generality. We assume that \( a > b \) and let \( t = \frac{a}{b} > 1 \) to determine the condition for \( \lambda \) and \( \mu \) to become \( g(t) \geq 0 \).

Keywords

Centroidal Mean, Arithmetic Mean, Contraharmonic Mean
1. Introduction

For \( a, b > 0 \) with \( a \neq b \), the Centroidal mean \( C(a, b) \), Harmonic mean \( A(a, b) \) and Contraharmonic mean \( C(a, b) \) are defined by:

\[
C(a, b) = \frac{2(a^2 + ab + b^2)}{3(a + b)}, \quad A(a, b) = \frac{a + b}{2}; \quad C(a, b) = \frac{a^2 + b^2}{a + b}
\]

respectively.

The main objective of this research work is to present optimization of the following inequalities:

\[
\lambda A(a, b) + (1 - \lambda) C(a, b) \leq \bar{C}(a, b) \leq \mu A(a, b) + (1 - \mu) C(a, b)
\]

and

\[
C[\lambda a + (1 - \lambda) b, \lambda b + (1 - \lambda) a] \leq \bar{C}(a, b) \leq C[\mu a + (1 - \mu) b, \mu b + (1 - \mu) a]
\]

Recently, both mean values have been the subject of intensive research. In particular, many remarkable inequalities and properties for these means can be found in the literature [1] [2].

This work finds out such inequality that arises in the search for determination of a point of reference about which some function of variants would be minimum or maximum. Since very early times, people have been interested in the problem of choosing the best single quantity, which could summarize the whole information contained in a number of observations (measurements). Moreover, the theory of means has its roots in the work of the Pythagorean who introduced the harmonic, geometric, and arithmetic means. Peter et al. [3] introduced seven other means and gave the well-known elegant geometric proof of the celebrated inequalities among the harmonic, geometric, and arithmetic means. The strong relations and introduction of the theory of means with the theories of inequalities, function equations, probability and statistics add greatly to its importance. This single element is usually called a means or average. The term “means” or “average” (middle value) has for a long time been used in all branches of human activity.

The basic function of mean value is to represent a given set of many values by some single value. In [4], the authors were the first time introduced power means defined the meaning of the term “representation” as determination of appoint of reference about which some function of variants would be minimum. More recently the means were the subject of research and study whereas essential areas in several applications such as: physics, economics, electrostatics, heat conduction, medicine and even in meteorology. It can be observed that the power mean \( M_p(a, b) \) of order \( p \) can be rewritten as (see as [5])

\[
M_p(a, b) = \begin{cases} 
\left(\frac{a^p + b^p}{2}\right)^{\frac{1}{p}} & p \neq 0 \\
\sqrt[ab]{p} & p = 0
\end{cases}
\]
If we denote by
\[
A(a,b) = \frac{1}{2}(a+b), \quad G(a,b) = \sqrt{ab} \quad \text{and} \quad H(a,b) = \frac{2ab}{a+b},
\]
the arithmetic, geometric and harmonic means of two positive numbers \(a\) and \(b\), respectively. In addition, the logarithmic and identric means of two positive real numbers \(a\) and \(b\) defined by [6]
\[
L(a,b) = \begin{cases} 
\frac{b-a}{\log b - \log a} & a \neq b \\
\frac{a}{a} & a = b
\end{cases}
\]
\[
I(a,b) = \begin{cases} 
\frac{1}{e} \left( \frac{b^a}{a^b} \right)^{rac{1}{b-a}} & a \neq b \\
\frac{a}{a} & a = b
\end{cases}
\]

Several authors investigated and developed relationship of optimal inequalities between the various means.

The well-known inequality that:
\[
\min \{a, b\} \leq H(a,b) = M_{-1}(a,b) \leq G(a,b) = M_0(a,b) \\
\leq L(a,b) \leq I(a,b) \leq A(a,b) = M_1(a,b) \leq \max \{a, b\}
\]
and all inequalities are strict for \(a \neq b\).

In [7], researchers studied what are the best possible parameters \(\alpha_1, \alpha_2, \beta_1\) and \(\beta_2\) by two theorems:

Theorem (1) the double inequality:
\[
\alpha_1 A(a,b) + (1-\alpha_1) H(a,b) \leq L(a,b) \leq \beta_1 A(a,b) + (1-\beta_1) H(a,b)
\]
holds for all \(a, b > 0\) if and only if \(\alpha_1 \leq 0\) and \(\beta_1 \geq \frac{2}{3}\) when proved that the parameters \(\alpha_1 \leq 0\) and \(\beta_1 \geq \frac{2}{3}\) cannot be improved.

Theorem (2) the double inequality:
\[
\alpha_2 A(a,b) + (1-\alpha_2) H(a,b) \leq L(a,b) \leq \beta_2 A(a,b) + (1-\beta_2) H(a,b)
\]
holds for all \(a, b > 0\) if and only if \(\alpha_2 \leq \frac{2}{e}\) and \(\beta_2 \geq \frac{5}{6}\) when proved that the parameters \(\alpha_2 \leq \frac{2}{e}\) and \(\beta_2 \geq \frac{5}{6}\) cannot be improved.

Interestingly in [5] B. Long et al., proved that the following results: \(M_{\alpha_1}(a,b)\) and \(M_{\alpha_2}(a,b)\) are the best possible lower and upper power bounds for the generalized logarithmic mean \(L_t(a,b)\) for any fixed \(t > 0\) the double inequalities
\[
M_{\alpha_1}(a,b) < L_t(a,b) < M_{\alpha_2}(a,b)
\]
holds for all \(a, b > 0\) with \(a \neq b\), and they found \(L_x(a,b)\) the optimal lower generalized logarithmic means bound for the identric means \(I(a,b)\) for inequalities \(L_x(a,b) < I(a,b)\) holds for all \(a, b\) are positive numbers with \(a \neq b\).
Pursuing another line of investigation, in [8] the authors showed the sharp upper and lower bounds for the Neuman-sandor $NS(a, b)$, [9] in terms of the liner convex combination of the logarithmic means $L(a, b)$, and second seiffert means $T(a, b)$ [10] of two positive numbers $a$ and $b$, respectively for the double inequalities

$$aL(a, b)+(1-\alpha)T(a, b) \leq NS(a, b) \leq \beta L(a, b)+(1-\beta)T(a, b)$$

holds for all $a, b > 0$ with $a \neq b$ is true if and only if $\alpha \geq \frac{1}{4}$ and $\beta \leq 1-\pi \left[4\log(1+\sqrt{2})\right]$.

In [11] have improvements and refinements by HZ Xu et al., for they found several sharp upper and lower bounds for the Sandor-yang means $R_{\lambda\mu}(a, b)$ and $R_{\lambda\mu}(a, b)$ [12] [13] in terms of combinations of the arithmetic means $A(a, b)$ and the contra-harmonic mean $C(a, b)$ [4] [14].

The authors have to proven our main results several lemmas find the best possible parameters $\alpha_i, \beta_i (i = 1, 2, 3, 4)$ such that the double inequalities

$$c^{\alpha_1}(a, b)A^{\alpha_1}(a, b) < R_{\lambda\mu}(a, b) < c^{\beta_1}(a, b)A^{\beta_1}(a, b)$$

holds for all $a, b > 0$ with $a \neq b$.

In [15], Neuman proved that the double inequalities

$$\lambda C(a, b)+(1-\lambda)A(a, b) \leq M(a, b) \leq \mu C(a, b)+(1-\mu)A(a, b),$$

with $M(a, b)$ is the Neuman-S andor mean, hold for all holds for all $a, b > 0$ with $a \neq b$ if and only if $\lambda \leq \frac{1-\log(1+\sqrt{2})}{\log(1+\sqrt{2})}$ and $\mu \geq \frac{1}{6}$. In [2] Shen, the inequalities sharps bounds for Seiffert mean in terms of Contraharmonic mean

$$C\left[\lambda a+(1-\lambda)b, \lambda b+(1-\lambda)a\right] \leq T(a, b) \leq C\left[\mu a+(1-\mu)b, \mu b+(1-\mu)a\right],$$

with $T(a, b) = \frac{a-b}{2\arctan\left(\frac{a-b}{a+b}\right)}$, were proved to be valid for $\frac{1}{2} < \lambda, \mu < 1$ and

for all $a, b < 0$ with $a \neq b$ if and only if $\lambda \leq \left(1+\frac{\sqrt{4-\pi}}{\pi}\right)$ and $\mu \geq \frac{3+\sqrt{3}}{6}$.
Wen-Hui Li and Feng Qi [16], proved that the double inequality
\[ \lambda Q(a, b) + (1 - \lambda) M(a, b) \leq C(a, b) \leq \mu Q(a, b) + (1 - \mu) M(a, b), \]
with \( Q(a, b) = \sqrt{\frac{a^2 + b^2}{2}} \) is the root-square mean, holds for all \( a, b > 0 \) with \( a \neq b \) if and only if \( \lambda \leq \frac{1}{2} \) and \( \mu \geq \frac{3 - 4 \ln(1 + \sqrt{2})}{3 - 4 \ln(1 + \sqrt{2})} = 0.7107 \ldots \).

For more information on this topic, you can refer to the following references: [17] [18] [19].

2. Main Results

Motivating by results mentioned above, we naturally ask a question: what are the best possible parameters \( \lambda, \mu \) that (1.1) and (1.2) can be hold?

The aim of this paper is to answer this question. The solution to this question may be stated as the following Theorem:

**Theorem 1.** Assuming \( a > 0, b > 0 \) with \( \frac{a}{b} > 1 \) then,

1) if \( \lambda \in \left( \frac{2}{3}, +\infty \right) \) and \( \mu \in \left( -\infty, \frac{2}{3} \right) \) then, the double inequality (1.1) holds.

2) if \( \lambda \in \left( \frac{3 - \sqrt{3}}{6}, \frac{3 + \sqrt{3}}{6} \right) \) and \( \mu \in \left( -\infty, -6 \right] \cup \left[ 0, \frac{3 - \sqrt{3}}{6} \right) \cup \left( \frac{3 + \sqrt{3}}{6}, +\infty \right) \)
then the double inequality (1.2) holds.

**Proof.** 1): Assuming \( a > 0, b > 0 \) with \( \frac{a}{b} > 1 \)

\[ \lambda \left( \frac{a + b}{2} \right) + (1 - \lambda) \left( \frac{a^2 + b^2}{a + b} \right) \leq \frac{2(a^2 + ab + b^2)}{3(a + b)} \leq \mu \left( \frac{a + b}{2} \right) + (1 - \mu) \left( \frac{a^2 + b^2}{a + b} \right) \]

Set \( t = \frac{a}{b} > 1 \). Then, we obtain

\[ \lambda \left( \frac{bt(t + 1)}{2} \right) + (1 - \lambda) \left( \frac{bt(t^2 + 1)}{t + 1} \right) \]

\[ \leq \frac{2bt(t^2 + t + 1)}{3(t + 1)} \leq \mu \left( \frac{bt(t + 1)}{2} \right) + (1 - \mu) \left( \frac{bt(t^2 + 1)}{t + 1} \right) \]

We start by showing that

\[ \lambda \left( \frac{bt(t + 1)}{2} \right) + (1 - \lambda) \left( \frac{bt(t^2 + 1)}{t + 1} \right) - \frac{2bt(t^2 + t + 1)}{3(t + 1)} \leq 0, \]

\[ \Rightarrow \frac{3b(t + 1)^2}{6(t + 1)} + (1 - \lambda) \frac{6b(t^2 + 1)}{6(t + 1)} - \frac{4b(t^2 + t + 1)}{6(t + 1)} \leq 0 \]

Because \( t > 0 \) therefore the study amounts to proving that
Let 
\[ f(t) = 3\lambda b(t+1)^2 + 6(1-\lambda)b(t^2+1) - 4b(t^2+t+1) \]

We have to prove that the function \( f \) is negative under certain conditions on the parameter \( \lambda \), a.e: \( f(t) \leq 0 \). So
\[ f(t) = 3\lambda b(t+1)^2 + 6(1-\lambda)b(t^2+1) - 4b(t^2+t+1) \leq 0 \]

Because \( f(1) = 0 \), it will suffice to show that \( f \) is decreasing for all \( t > 1 \), which amounts to studying the sign of the derivative \( f' \) of \( f \). We have:
\[ f'(t) = 6\lambda b(t+1) + 12(1-\lambda)bt - 4b(2t+1) \]

Because \( f'(1) = 0 \), it will suffice to show that \( f' \) is decreasing for all \( t > 1 \), which amounts to studying the sign of the derivative \( f'' \) of \( f' \). We have:
\[ f''(t) = 2b(2-3\lambda) < 0 \Leftrightarrow \lambda > \frac{2}{3} \]
so that \( f' \) is decreasing for \( t > 1 \) and therefore, we obtain that \( f(t) < 0 \) because \( f(1) = 0 \).

Finally in this part for \( a > 0, b > 0 \) with \( \frac{a}{b} > 1 \), we obtain that
\[ \lambda \left( \frac{b(t+1)}{2} \right) + (1-\lambda) \left( \frac{b(t+1)}{t+1} \right) \leq \frac{2b(t^2+t+1)}{3(t+1)}, \text{ for all } \lambda > \frac{2}{3}. \]

To show the second inequality in this first case, we proceed by similar calculations. This is done by considering the function \( g \) defined by
\[ g(t) = g(t) = 3\mu b(t+1)^2 + 6(1-\mu)b(t^2+1) - 4b(t^2+t+1). \]

So, after all the calculations, we get that for \( a > 0, b > 0 \) with \( \frac{a}{b} > 1 \), that
\[ g(t) \geq 0, \text{ for all } \mu < \frac{2}{3}, \text{ a.e:} \]
\[ \frac{2(a^2+ab+b^2)}{3(a+b)} \leq \mu \left( \frac{a+b}{2} \right) + (1-\mu) \left( \frac{a+b}{a+b} \right) \]

2): Assuming \( a > 0, b > 0 \) with \( \frac{a}{b} > 1 \) and with similar calculations and by the same idea we obtain that for all
\[ \lambda \in \left( \frac{3-\sqrt{3}}{6}, \frac{3+\sqrt{3}}{6} \right) \text{ and } \mu \in (-\infty, -6] \cup \left[ 0, \frac{3-\sqrt{3}}{6} \right] \cup \left( \frac{3+\sqrt{3}}{6}, +\infty \right) \]
then the double inequality
\[ C_0[\lambda a + (1-\lambda)b, \lambda b + (1-\lambda)a] \leq C_0(a,b) \leq C \left[ \mu a + (1-\mu)b, \mu b + (1-\mu)a \right] \]
holds.
Conclusion 1. In our work, we studied the following double inequalities: respectively (1.1) and (1.2)

\[ \lambda A(a,b) + (1 - \lambda) C(a,b) \leq \bar{C}(a,b) \leq \mu A(a,b) + (1 - \mu) C(a,b) \]

and

\[ C[\lambda a + (1 - \lambda)b, \lambda b + (1 - \lambda)a] \leq \bar{C}(a,b) \leq C[\mu a + (1 - \mu)b, \mu b + (1 - \mu)a] \]

by searching the best possible parameters such that (1.1) and (1.2) can be hold.

Firstly, we have inserted

\[ f(t) = \lambda A(a,b) + (1 - \lambda) C(a,b) - \bar{C}(a,b) \]

and

\[ g(t) = \mu A(a,b) + (1 - \mu) C(a,b) - \bar{C}(a,b) \]

Without loss of generality, we have assumed that \( a > b \) and let \( t = \frac{a}{b} > 1 \) to determine the condition for \( \lambda \) and \( \mu \) to become \( f(t) \leq 0 \) and \( g(t) \geq 0 \).

Secondly, we have inserted Without loss of generality, we assume that \( a > b \) and let \( t = \frac{a}{b} > 1 \) to determine the condition for \( \lambda \) and \( \mu \) to become

And finally, we got that:

1) if \( \lambda \in \left( \frac{2}{3}, +\infty \right) \) and \( \mu \in \left( -\infty, \frac{2}{3} \right) \) then, the double inequality (1.1) holds.

2) if \( \lambda \in \left( \frac{3 - \sqrt{3}}{6}, \frac{3 + \sqrt{3}}{6} \right) \) and \( \mu \in \left( -\infty, -6 \right] \cup \left( 0, \frac{3 - \sqrt{3}}{6} \right] \cup \left( \frac{3 + \sqrt{3}}{6}, +\infty \right) \)

then the double inequality (1.2) holds.

Acknowledgements

The authors gratefully acknowledge Qassim University, represented by the Deanship of Scientific Research, on the material support for this research WORK under the number (1063) during the academic year 1441AH/2020AD.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References


Neuroevolution Strategy for Time Series Prediction

George Naskos¹, Konstantinos Goulianas¹, Athanasios Margaris²

¹Department of Information and Electronic Engineering, International Hellenic University, Thessaloniki, Greece
²Department of Digital Systems, University of Thessaly, Larissa, Greece

Abstract

Optimization is a concept, a process, and a method that all people use on a daily basis to solve their problems. The source of many optimization methods for many scientists has been the nature itself and the mechanisms that exist in it. Neural networks, inspired by the neurons of the human brain, have gained a great deal of recognition in recent years and provide solutions to everyday problems. Evolutionary algorithms are known for their efficiency and speed, in problems where the optimal solution is found in a huge number of possible solutions and they are also known for their simplicity, because their implementation does not require the use of complex mathematics. The combination of these two techniques is called neuroevolution. The purpose of the research is to combine and improve existing neuroevolution architectures, to solve time series problems. In this research, we propose a new improved strategy for such a system. As well as comparing the performance of our system with an already existing system, competing with it on five different datasets. Based on the final results and a combination of statistical results, we conclude that our system manages to perform much better than the existing system in all five datasets.

Keywords

Neuroevolution, Neural Networks, Evolutionary Algorithms, Time Series

1. Introduction

The great potential of neural networks has been proven repeatedly in the past years by many studies, as they are applicable to real problems, even in our daily lives [1]. Evolutionary algorithms (EA) are used in optimization problems, with the ability to cope better with problems where the optimal solution is found.
among a chaotic set of solutions [2]. Generally speaking, the larger the range of possible solutions to a problem, the more efficient the EA associated with it. The combination of the two approaches results in the neuroevolution (NE) [3] [4]. This approach describes a neural network equipped by an evolutionary optimization algorithm. Bearing in mind the benefits of evolutionary algorithms, they are proven to be useful in deep learning [5], because of the enormous size of such a network that making it ideal for EA use. This allows the achievement of the same or better results in less time than a conventional optimization algorithm. It is also possible to build an EA system to find out which neural network architecture is appropriate for the problem. Then, the proposed network can be used, using a conventional optimization method [6]. In this study we will investigate the use of NE in time series problems [7]. This is a type of problem that has an increased degree of difficulty, having an additional factor, and more specifically, the time. Five datasets will be used, with some of their data to be derived from real problems and some other data, from simulations. Five different experiments will be performed on each of the datasets and then our own system will be compared with the memetic cooperative neuroevolution (MCNE) system [8]. In both systems, the same datasets have been used for the experiments, which are split into training/validation/testing in the same way [9]. Our system is generally based on their own MCNE. At the same time, proposals will be made for different methods and parameters in evolutionary algorithms, where they can be used not only in NE but generally in optimization problems. Finally, the results of our system will be compared to the upgraded version of their own MCNE system.

2. Artificial Neural Network

Neural networks try to replicate the biological function of the human brain. Each neuron receives an excitation and, after processing, sends its own excitation to the other neurons to which it is connected. There are many types of neural networks, one of which is multilayer perceptron (MLP) [10].

It is well known and it can give very good results in a wide range of problems. This is the kind of neural network that will be used in the following experiments. An MLP is described by one or more layers of neurons. Each neuron has inputs and each input is a data characteristic of the situation. Each entry, it associated with a weight value. This value is a number that indicates how important this feature is for that particular neuron. There is an extra weight called bias. The bias moves the activation function to the x-axis so that it fits better depending on the input data. The weights are multiplied by their associated input values and the resulted products are summed each other. The resulted value goes through an activation function whose return value is the output signal of the specific neuron in Figure 1. The neuron acquires the ability to learn through an optimization method that appropriately adapts the weights of the neuron.

The main advantages of neural networks are:
The ability to learn. More specifically, through their training, they can learn to solve a problem by adjusting to their weights.

Their most powerful feature is generalization. It is defined as their ability to produce results for data they have not seen in their training.

There is a large set of optimization methods based on neural network which help to solve classification and regression problems. In addition, given an entity’s past values, they can predict its future values.

Their major disadvantages are:

There is no general rule for the most appropriate use of the activation function, as well as for the optimum number of layers and the number of neurons in each layer. Thus, it takes multiple efforts and continuous testing, to determine the best configuration and specify a network that produces satisfactory results.

Because of its nature, there is no satisfactory control over the network as well as useful information that someone can derive by looking at the weights and connections of the network. In general, the real function of a neural network is that of a black box, for which the only thing we know is the input data and its results.

3. Evolutionary Algorithms

The creation of evolutionary algorithms (EA) is the idea that the biological process of evolution could be the basis for the development of optimization algorithms [1] [2]. The theory associated with these algorithms is based on natural selection and genetic change, with a population of individuals competing with each other to choose the best. Recombining the best among them will produce better offspring. If the process continues repeatedly, the members of the resulting population will be much better with respect to the initial members of this set. In a real problem, the members of the population, are a candidate solution to the problem.

EA are divided into several stages:

1) Initially, a population is created by individuals (Algorithm 1-Step 1). Each individual contains chromosomes and the chromosomes contain genes. Genes are the values that have their characteristics, which are the ones that will determine how good the solution is to the problem. This number is the fitness value.
of the individual.

2) Once a fitness value has been found for each individual (Algorithm 1-Step 2), a new set of individuals is created, selected on the basis of a method based on their fitness (Algorithm 1-Step 3). It has to be noted, that in the general case, the best ones are selected, but the worst ones are also given the opportunity to be selected, because they may contain genetic code that is useful for the next step.

3) This stage is the recombination of the total from the previous stage. So, these people will fertilize the next generation (Algorithm 1-Step 4). Recombination is the most basic way of population growth. In most cases, from the set produced by the selection method there are two people that are selected. Therefore, by recombining these two individuals, we can produce two even better offspring. The fact that evolutionary algorithms, are targeted search algorithm, and not just a random search algorithms is based to this recombination procedure.

4) The young offspring will go through a method of random mutation of their genes. This happens in order to enable the population to surpass any local optimum as it grows (See Algorithm 1-Step 5).

5) Finally, the offspring are integrated into the population, fitness is calculated for each new individual in the population (Algorithm 1-Step 6) and the process starts from the beginning.

The algorithm terminates when the specified number of iterations or a computational limit has been reached.

The advantage of EA is that they can solve quickly difficult problems. The greater the number of possible optimal solutions, the more efficient they are in terms of both speed and result. Their main advantages are:

Most conventional methods are stiff or sometimes inappropriate for several problems, due to their need for complex mathematics. The use of such mathematics is indifferent to EA, which makes them suitable for a wide range of problems.

It is one of the few methods that simultaneously explores the search space and takes advantage of the information already processed. Thus, random searches make a good exploration of the space and the search takes advantage of the information.

1) BEGIN
2) \hspace{2em} **Step 1**: INITIALISE population with random candidate solutions;
3) \hspace{2em} **Step 2**: EVALUATE each candidate;
4) \hspace{2em} REPEAT UNTIL (TERMINATION CONDITION is satisfied) DO
5) \hspace{4em} **Step 3**: SELECT parents;
6) \hspace{4em} **Step 4**: RECOMBINE pairs of parents;
7) \hspace{4em} **Step 5**: MUTATE the resulting offspring;
8) \hspace{4em} **Step 6**: EVALUATE new candidates;
9) \hspace{2em} DONE
10) END

*Algorithm 1.* Pseudocode of evolutionary algorithm.
Another major advantage is their easy parallelism, a potential that is scarce in other competing methods. Parallelism results in better results in less time.

Finally, they can be combined with other methods. Although the power of EA is high, there are some cases of problems in which other methods produce very good results, due to their specialization in the problem. This is a result of the great flexibility of EA.

As with any method there are disadvantages:

The first thing anyone should do to solve a problem with EA is to encode the problem in such a way that EA can process it. However, this can sometimes be difficult depending on the nature of the problem. Specifically, this encoding is a process that will take time for one to understand and feel comfortable with.

The next step is to be able to set and implement a goal for EA. This goal is to guide the population of the algorithm to the optimal solution. Therefore, this stage greatly influences finding the optimal solution and in turn requires time and familiarity with it.

4. Proposed Neuroevolution Architecture

Thanks to the great flexibility that EA possess, they are able to combine with other methods. However, neural networks are needed to solve machine learning problems. This results in the combination of EA and neural networks. Neural networks alone cannot be trained, they need to use an optimization algorithm. One of the best known is backpropagation (BP). In NE [3] [4] the EA take the place of the optimization algorithm. The good news is that it is no longer necessary to use sophisticated mathematics to train and optimize the network. There are also many ways to customize and control during training, as are significantly reduced the effort and experimentation needed to find the right network. For example, in very large networks, in deep learning, training times are greatly reduced and consistently good results are produced.

4.1. Existing Study

There are several studies in the NE industry. One of the most recent that deals with time series problems can be found in [7]. An upgraded system to solve such problems is proposed in [8]. In practice, it combines EA with conventional algorithms, in order to achieve better results. Another study that uses the same system, but different strategies is described in [9], with very good results. Having in mind the best results from [9], that were compared to ours and based on an overview of the MCNE strategy described in [8], we were able to produce better results at about the same cost. We have also used the same datasets with [11] [12] [13] [14] [15] from the most recent study [9] to allow a meaningful comparison of the systems.

4.2. Our System

The encoding of the neural network differs in our method from the present study [8], since this is one of the most important pieces of EA and has a great
impact on the good results of the algorithm. Our system puts a subpopulation, that is, a set of neural networks, which implies that the subpopulation individuals are neural networks, while in [8] each neuron is also a subpopulation and its atoms are neurons with different weights. As a result, in the process of evolution, when the worst of the population is replaced by the bests, only the neurons actually change, while in proposed method the whole neural network changes. Meanwhile, the fitness of both systems is calculated in the same way and the overall rationale of the system is the same, so that the results can be compared between the two systems. Among the three methods cited in [9], MCNE [8] is closer to our system and the comparison will be performed between these two systems.

The implementation of the proposed system is not based on a library or on an available neural simulator, but everything has been created from scratch. This decision was made on the grounds that in order to do proper research there must be complete control over the system, a feature that generally is not supported by third-party software. Of course during the implementation, we used the well-known techniques for both EA and neural networks with some customizations. The advantage of this implementation is that it has many parameters that they can easily be controlled, so that the system can adapt and solve the problem in the best possible way. This system was able to give better results across all datasets [11] [12] [13] [14] [15].

In a more detailed description the following methods for EA have been used.

Regarding the selection process, the tournament method has shown far better results than roulette. It seems to help in rapid population growth, while at the same time is able to keep the population in balance in case of a high likelihood of mutation. The tournament has only one parameter that controls how many people will compete against each other. In this way it is possible to control how possible is the selection of a non-good individual. However, there may be some cases in which the use of fitness is not adequate to decide which one to choose, due to the generality of fitness. One solution is to create an extra value for each person. So, we don’t have to mess with actual fitness since it is calculated every epoch right after fitness is calculated. This feature is called the “percentage of fitness” because it is a percentage-based fitness. This enables the creation of many methods for its calculation. The selection method can be based on this value instead of fitness, since the logic of fitness calculation remains the same, but at the same time, we can select people with different strategies. In the next paragraph we propose three methods of calculating this value.

1) We called the first method, the rank method, because every person will get its value based on the worst and the best of the current population. In Equation (1) we consider the worst to be 0% and the best 100%.

\[
Percent = 100 \cdot \frac{c - \text{min}}{\text{max} - \text{min}} \tag{1}
\]

where
c is the fitness of the person;
min is the smallest fitness in the population;
max is the greatest fitness in the population.

2) The current best, calculates the percent based on the best person in the current population. Thus, the range of the percentage is \([0, \text{best fitness}]\), and therefore, in order to get a value of 0% the value of fitness should be zero. This one is less stringent with the not-so-good people of the population compared with the rank method.

3) The last method requires the maximum value calculated via the fitness, which is not always available. The range of the percentage is \([0, \text{total best fitness}]\) and it is used in problems where the fitness is not available or there is not a good way for its calculation.

The crossover is used for the recombination method with only two parents, producing two offspring. The parameter in this method determines how many genes the crossover will take. The smaller the parameter number, the better the knitting of the genes. The parameter helps to better control the growth.

In the mutation method, it has to be determined how many chances each gene has to mutate. Regarding this method, there are two adjustments that have been implemented. When the mutation is done, it doesn’t just get a random value through a range, but a value emerged via the addition or the subtraction of a value from the old one. This makes growth smoother, without killing good people due to a big value change. Also, as the epochs pass, the likelihood of mutation increases. So, the population lets to grow on its own and then when it sticks to local optimum, it is directed to move forward with the high mutation potential. We change the probability based on the epochs with the linear function. At this point, it can be used any function the user wants. However, the experiments show that there was a little change in the results with more complicated functions than linear.

Another important method for growth, as well as for the end results, is the parallel growth with subpopulation. We did this by starting different threads for each population. An interesting feature of the proposed method is the implementation of a method called the “migration”. From time to time, a set of the best person from every sub-population is created. It is this group that will replace the worst persons of every subpopulation Figure 2. As a consequence, the overall performance and growth are greatly increased, because some subpopulations may not have enough good persons, in such cases the migration method is used to fix this problem. The downside to this method is when it is given to the system to generate many subpopulations, and therefore many threads. Specifically, when the computer processor cannot support so many threads, it requires more time to complete, although it is executed in parallel due to threads creation.

The proposed system as well as the system described in [9] use hybrid EA. When a conventional algorithm is executed at certain times for better development then the EA are called hybrid [16].
In the NE, the conventional algorithm is an optimization method for neural networks. In our case it is BP [6]. At a certain number of EA epochs, there are some BP epochs. In our system this is done immediately before migration is complete. Specifically, once the group is destined for migration, these individuals in the group execute some BP epochs [Figure 2]. Without the hybrid, our system would produce good results, but it could not overcome some of the problems, with the results of [9]. With the hybrid it was able to overcome these problems and the best results found in [9].

Specifically, for BP, its simple form was used [6]. The difference is in the behavior of a BP parameter and more specifically, the learning rate. In the proposed method, an adapted learning rate was used. So, the algorithm starts with a relatively large learning rate and during the training process its value progressively decreases. This behavior is due to the fact that for small iteration numbers, namely after the beginning of the training process, the large variation in the values of the neuron weights is not a major issue. On the contrary, it helps in faster growth. However, as the number of epochs increases and the optimum solution is approached, smaller changes in the weight values are required in order for the system to reach the optimum solution as close as possible.

Another feature of our system is the ability to develop EA with a variable neural network, given two parameters. The first of these parameters controls the range of the number of layers, while the other one controls the range of number of neurons per layer. Therefore, the EA in the creation phase of the first population, construct different random neural networks, based on the two previous parameters.

This process creates individuals with a variable set of genes, which implies that a different method is needed for recombination and mutation, simply to
support individuals with a variable set of genes. There are many implementations for EA with variable number of genes. The problems we have encountered are that these implementations require the addition or removal of random N genes in the individual, as well as other actions such as the transfer of a number of genes from one position to another during the population mutation phase [17]. This task is impossible for our system: in our case, a “person” is a neural network and his genes are all the weights of the network; therefore all we can do is to add or subtract neurons only to the hidden layers of the network. So, the above mutation method can be applied to our system but with some limitations. Also, if we consider that a set of N genes is a neuron, then we cannot transfer this neuron to another layer safely, as the neurons of the layer intended to go may not have the same number of inputs as the one we transport, so that a valid network remains after the transport. The recombination we have proposed and implemented in the system consists of one function, which is a map between two one-dimensional arrays A and B, each having a different size. Each array is an individual and each array element is a gene. This map shows, given a position in A, what should be the corresponding position in Array B. Of course, all the positions of the two arrays cannot be fully matched. This function can be easily combined with previous fixed size recombination methods. Given a position in Array A, the corresponding position in Array B is given by Equation (2).

\[ P_B = \frac{P_A \cdot L_B}{L_A} \]  

where:

- \( P_B \) is the position of B we are looking for;
- \( P_A \) is the position of A that we know;
- \( L_B \) is the size of array B;
- \( L_A \) is the size of array A.

One of the major problems of conventional neural networks is overtraining. This means that the network loses the ability to generalize and specializes only in some cases. It is a destructive case of training that is usually solved by defining a set of data called a validation set. All of this is used by the system during training, not to train the network, but to be able to monitor whether it loses its generality.

Our system supports this, as it provides the ability to interrupt training or finish normally completing all epochs and selecting the best person from all epochs based on the validation set. Our experiments generally show that the system achieves the best validation error relatively close to the end of the epochs. This indicates that the system does not lose generalization easily and that the use of validation set is only necessary in a few problems (see Figure 3).

5. Experiments and Analysis

During the simulation and the experimentation stage, five different datasets were used and more specifically, the following:
1) Sunspot [12]. The magnetic field of the Sun makes a periodic motion and each period lasts about 11 years. At the end of each period the poles of the sun have changed their position with each other. It has been observed that during this time period of eleven years, dark spots are formed on the surface of the sun, due to the change in their polarity. This phenomenon greatly affects our solar system and the climate on Earth. The proposed system is configured to predict the number of spots during these periods. It has been observed that this number is difficult to predict, as there are unexpected changes affecting the number of spots.

2) Mackey-Glass [14]. The situation is associated with a set of functions that produce diverse waveforms, which apparently have limitless or “chaotic” solutions. More specifically, these results are related to dynamic breathing and hematopoietic diseases. Essentially, some chronic or even acute diseases have been observed that some of their symptoms exhibit a periodicity and attempt to correlate mathematical functions in this study to predict the periodicity of the symptoms.

3) Lazer [15]. This data set is composed of data recorded using a remote infrared laser, produced for benchmark use, which has been used in many competitions. Its data is low dimensional, non-linear and constant time series.

4) Lorenz [13]. It is made up of data from observations of various hydrodynamic flows that exhibit general periodicity, but over time it appears volatility in its data. It is another natural observation, so it contains the element of the unpredictable.

5) Taiwan Trading Index Exchange (TWI Exchange) [11]. Simple recording of the Taiwanese exchange rate within a given time, this dataset is quite different from the others, as its values range in a much smaller range.

We use the already split format of the datasets, with 60%/20%/20% data, training/validation/testing respectively, where they are taken from their own system (https://github.com/gary-wong-fiji/Meme-Collection-SEQ).
5.1. Experiments Plan

In order to be able to properly compare the two systems, the datasets [11] [12] [13] [14] [15] from the research [9] had to be used for simulation, just as they divided them into their subsets. Therefore, the error is measured using the same root mean square error method (RMSE). The same method is the one that guides the evolution of subpopulations. The best person for each run is selected based on the fitness of the individual in the validation set. The $C$ cost of each experiment is calculated from Equation (3).

$$C = \left( \sum_{i=1}^{n} m \cdot t \right) + \left( \sum \right)$$  \hspace{1cm} (3)

where:
- $n$ the epochs of the evolutionary algorithm;
- $m$ the crowd of people in the population;
- $t$ the multitude of subpopulations;
- $r$ the number of times that BP algorithm will be executed (see Equation (5));
- $z$ the number of epochs the BP algorithm will execute each time.

$$z = p \cdot n$$  \hspace{1cm} (4)

$$r = \frac{n}{z}$$  \hspace{1cm} (5)

There is a parameter $N$ that is increased one by one to take into account fitness calculation when initializing the system. Another parameter is the so-called “percentage” $p$, that defines two things:

1) The number of epochs of the evolutionary algorithm, that the BP algorithm execute.
2) The number of epochs of the BP algorithm.

The motivation for using a variable value for the $p$ parameter, was to keep a balance between how often and for how many epochs BP will executed. For example, if the $p$ parameter increases its value, the BP algorithm will execute for more epochs, but less often.

5.2. Parameters of the Experiments

The experiments were divided into four types. The formulas have some common parameters Table 1 and differ Table 2 with respect to some others, whereby the final conclusions will be drawn on the overall picture of the system as well as its efficiency. All four types were applied to all datasets and thirty system executions were performed for each type. This is due to the fact that the evolutionary algorithms by they own nature are characterized by the property of randomness to a large extent. The conclusions are produced via statistical methods applied to the results emerged from all executions.

The parameters described as common, have been selected after many repeated system tests on the datasets [11] [12] [13] [14] [15]. Observing the results of the tests as well as how the system reacts to the changes in the values of the parameters, we came to the following values Table 1.
Table 1. Common parameters, for all test types.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Epochs</td>
<td>2000</td>
</tr>
<tr>
<td>Population size</td>
<td>15</td>
</tr>
<tr>
<td>Selection method</td>
<td>Tournament $K = 3$ (Size of tournament group)</td>
</tr>
<tr>
<td>Recombination method</td>
<td>Random $K = 5$ (Crossover break size)</td>
</tr>
<tr>
<td>Mutation method</td>
<td>Simple random (Add or subtract a random value)</td>
</tr>
<tr>
<td>Value range, for variable learning rate</td>
<td>$[0.3, 0.00001]$</td>
</tr>
<tr>
<td>Range of values for the probability of mutation</td>
<td>$[200, 80]$</td>
</tr>
<tr>
<td>Range of values for the mutation value</td>
<td>$[-1, 1]$</td>
</tr>
<tr>
<td>Neural network size</td>
<td>A hidden layer with 9 neurons</td>
</tr>
<tr>
<td>Activation function, for hidden layer neurons</td>
<td>Logistic sigmoid [18]</td>
</tr>
<tr>
<td>Network Output Function</td>
<td>Ground-relu [19] (except for the Lorenz dataset [13], where it has the linear)</td>
</tr>
</tbody>
</table>

Table 2. Different parameters.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>percentage ($P$)</th>
<th>Threads</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment type</td>
<td>A</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>0.1</td>
</tr>
</tbody>
</table>

In more detail, the greater the number of epochs, the better, but the epochs are costly in time. So after a point the performance/cost ratio will be very small and unprofitable.

The size of the population helps the system to make great leaps in improvement, from the earliest epochs, due to the larger volume of genetic material in the population. And that, in turn, adds a lot of time cost. Experiments have shown that our system does not need a large population.

Due to the selection of the small population, the total selection method must be small. The break size for the recombination method is small enough to keep the knitting of the genes satisfactory because the neural networks we have chosen are small in size.

The learning rate for the conventional BP [6] method is also small because most progress must be made by the EA [2] and the conventional method has the auxiliary role.

The relatively high probability of mutation was chosen because through the experiments the system showed that it can withstand them. In an even higher probability of mutation, we would see a sharp drop in the course of the population, because it’s best individuals are killed, the high probability of mutation destroys their genetic material.
Once our data is normalized, the range we selected for the mutation values is sufficient. In their research [9], they use five neurons in the hidden layer, we chose to have nine. In our test D, where the network has a variable size, in its hidden layer it can have from five to nine neurons. However, the results show that most neurons do not help much in our system for these datasets. Even when we tried it with a hundred neurons in the hidden layer it did not produce better results and made the execution of the program very slow.

The sigmoid function in the hidden layer and the ground-relu at the output of the network, are two simple but efficient methods and very fast in execution time. For the dataset Lorenz [13] we have the linear function at the output because in this dataset there are negative values and the ground-relu method does not produce negative values.

5.3. Extra Experiments

During simulation, another type of experiment, called the D experiment, was performed. With this type of experiment, we can draw conclusions regarding the effectiveness of variable evolution. This experiment has the same parameters as the previous types, except from the parameters described at Table 3.

To be able to understand the effectiveness of using a percentage method, the test type E was created. It is essentially just like the test D, but without the use of the percentage method. Finally, we have five types of problems A, B, C, D, E.

5.4. Results

After all executions were done, all data were collected from the system output and the following tables were generated (see Table 4). In these tables, the results of our system (A, B, C, D, E) are compared with the results of the existing study (Sequential, Concurrent, MCNE) [9]. As mentioned above, EA have the element of randomness to a large extent, and because of this, an additional column, namely the Mean error bias (MEB) has been added in Table 4. MEB column values are in the interval $[-100, 100]$, indicating how close (in percentage) the Mean error is to the best error (Best error column). A value of 100% means that the average error is equal to the best error, while a value of $-100\%$ means that it is equal to the worst error. This is a very important column, since we can see if the value we got at the best error is a pretty likely value or we just got lucky. However, MEB cannot be compared from one test to another because they have different values in the maximum and minimum error. The type that MEB describes is Equation (6).

$$\text{MEB} = -\left(2 \cdot \frac{\mu_e - m_e}{M_e - m_e} - 1\right) \cdot 100$$

where:

- $\mu_e$ is the average error;
- $m_e$ is the minimum error;
- $M_e$ is the maximum error;
Table 3. Different parameters for the extra experiment D.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage (P)</td>
<td>0.1</td>
</tr>
<tr>
<td>Threads</td>
<td>5</td>
</tr>
<tr>
<td>Recombination method</td>
<td>Variable size</td>
</tr>
<tr>
<td>Percent of fitness method</td>
<td>Current best ranked</td>
</tr>
<tr>
<td>Variable size neural network</td>
<td>Hidden layer: [5, 9] neurons A neuron in the output layer</td>
</tr>
</tbody>
</table>

Table 4. All tests results.

**Santa Fe Laser** [15]

<table>
<thead>
<tr>
<th>Test type</th>
<th>Best error</th>
<th>Mean error</th>
<th>Mean error bias</th>
<th>Worst error</th>
<th>Cost</th>
<th>Cost per thread</th>
<th>Mean (best validation error epoch)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.0122094</td>
<td>0.0369875</td>
<td>5.63%</td>
<td>0.0647200</td>
<td>200,075</td>
<td>40,015</td>
<td>1619</td>
</tr>
<tr>
<td>B</td>
<td>0.0151571</td>
<td>0.0430133</td>
<td>−17.93%</td>
<td>0.0624001</td>
<td>195,075</td>
<td>39,015</td>
<td>1742</td>
</tr>
<tr>
<td>C</td>
<td>0.0138906</td>
<td>0.0323462</td>
<td>11.01%</td>
<td>0.0553674</td>
<td>500,150</td>
<td>50,015</td>
<td>1715</td>
</tr>
<tr>
<td>D</td>
<td>0.0186784</td>
<td>0.0606052</td>
<td>38.98%</td>
<td>0.1560946</td>
<td>200,075</td>
<td>40,015</td>
<td>1509</td>
</tr>
<tr>
<td>E</td>
<td>0.0203611</td>
<td>0.0656812</td>
<td>54.02%</td>
<td>0.2174987</td>
<td>200,075</td>
<td>40,015</td>
<td>1269</td>
</tr>
<tr>
<td>Sequential [9]</td>
<td>0.0571243</td>
<td>0.0695330</td>
<td>−65.25%</td>
<td>0.0721420</td>
<td>269,421</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Concurrent [9]</td>
<td>0.0634781</td>
<td>0.0768557</td>
<td>−68.66%</td>
<td>0.0793412</td>
<td>121,200</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MCNE [8] [9]</td>
<td>0.1471420</td>
<td>0.1949820</td>
<td>−33.44%</td>
<td>0.2188464</td>
<td>100,000</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

**Lorenz** [13]

<table>
<thead>
<tr>
<th>Test type</th>
<th>Best error</th>
<th>Mean error</th>
<th>Mean error bias</th>
<th>Worst error</th>
<th>Cost</th>
<th>Cost per thread</th>
<th>Mean validation error epoch</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.0027403</td>
<td>0.0052443</td>
<td>24.78%</td>
<td>0.0093983</td>
<td>200,075</td>
<td>40,015</td>
<td>1897</td>
</tr>
<tr>
<td>B</td>
<td>0.0022897</td>
<td>0.0096347</td>
<td>33.82%</td>
<td>0.0244856</td>
<td>195,075</td>
<td>39,015</td>
<td>1813</td>
</tr>
<tr>
<td>C</td>
<td>0.0019629</td>
<td>0.0038730</td>
<td>28.18%</td>
<td>0.0072823</td>
<td>500,150</td>
<td>50,015</td>
<td>1949</td>
</tr>
<tr>
<td>D</td>
<td>0.0022847</td>
<td>0.0144763</td>
<td>32.90%</td>
<td>0.0386242</td>
<td>200,075</td>
<td>40,015</td>
<td>1767</td>
</tr>
<tr>
<td>E</td>
<td>0.0037872</td>
<td>0.0146563</td>
<td>62.75%</td>
<td>0.0621459</td>
<td>200,075</td>
<td>40,015</td>
<td>1809</td>
</tr>
<tr>
<td>Sequential [9]</td>
<td>0.0713540</td>
<td>0.0731450</td>
<td>48.59%</td>
<td>0.0783210</td>
<td>260,668</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Concurrent [9]</td>
<td>0.3214887</td>
<td>0.3445700</td>
<td>22.62%</td>
<td>0.3811421</td>
<td>121,200</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MCNE [8] [9]</td>
<td>0.0747062</td>
<td>0.0753210</td>
<td>73.42%</td>
<td>0.0793321</td>
<td>100,000</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

**Mackey Glass** [14]

<table>
<thead>
<tr>
<th>Test type</th>
<th>Best error</th>
<th>Mean error</th>
<th>Mean error bias</th>
<th>Worst error</th>
<th>Cost</th>
<th>Cost per thread</th>
<th>Mean validation error epoch</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.0027723</td>
<td>0.0037160</td>
<td>45.77%</td>
<td>0.0062529</td>
<td>200,075</td>
<td>40,015</td>
<td>1880</td>
</tr>
<tr>
<td>B</td>
<td>0.00277492</td>
<td>0.0037760</td>
<td>14.47%</td>
<td>0.0051503</td>
<td>195,075</td>
<td>39,015</td>
<td>1864</td>
</tr>
<tr>
<td>C</td>
<td>0.0023960</td>
<td>0.0033838</td>
<td>11.41%</td>
<td>0.0046261</td>
<td>500,150</td>
<td>50,015</td>
<td>1966</td>
</tr>
<tr>
<td>D</td>
<td>0.0010170</td>
<td>0.0040382</td>
<td>6.24%</td>
<td>0.0074617</td>
<td>200,075</td>
<td>40,015</td>
<td>1814</td>
</tr>
<tr>
<td>E</td>
<td>0.0020943</td>
<td>0.0043915</td>
<td>36.02%</td>
<td>0.0092757</td>
<td>200,075</td>
<td>40,015</td>
<td>1873</td>
</tr>
<tr>
<td>Sequential [9]</td>
<td>0.0019264</td>
<td>0.0045463</td>
<td>−37.10%</td>
<td>0.0057482</td>
<td>271,031</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Concurrent [9]</td>
<td>0.0032004</td>
<td>0.0059527</td>
<td>−56.75%</td>
<td>0.0067121</td>
<td>121,200</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MCNE [8] [9]</td>
<td>0.0123215</td>
<td>0.0252556</td>
<td>−16.57%</td>
<td>0.0345122</td>
<td>100,000</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
### Sunspot [12]

<table>
<thead>
<tr>
<th>Test type</th>
<th>Best error</th>
<th>Mean error</th>
<th>Mean error bias</th>
<th>Worst error</th>
<th>Cost</th>
<th>Cost per thread</th>
<th>Mean validation error epoch</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.0077448</td>
<td>0.0100750</td>
<td>13.07%</td>
<td>0.0131057</td>
<td>200,075</td>
<td>40,015</td>
<td>1909</td>
</tr>
<tr>
<td>B</td>
<td>0.0083803</td>
<td>0.0109812</td>
<td>13.68%</td>
<td>0.0144062</td>
<td>195,075</td>
<td>39,015</td>
<td>1817</td>
</tr>
<tr>
<td>C</td>
<td>0.0063976</td>
<td>0.0077735</td>
<td>30.12%</td>
<td>0.0103354</td>
<td>500,150</td>
<td>50,015</td>
<td>1877</td>
</tr>
<tr>
<td>D</td>
<td>0.0078447</td>
<td>0.0128080</td>
<td>27.05%</td>
<td>0.0214516</td>
<td>200,075</td>
<td>40,015</td>
<td>1578</td>
</tr>
<tr>
<td>E</td>
<td>0.0077889</td>
<td>0.0134720</td>
<td>38.76%</td>
<td>0.0263481</td>
<td>200,075</td>
<td>40,015</td>
<td>1485</td>
</tr>
<tr>
<td>Sequential [9]</td>
<td>0.0107341</td>
<td>0.0127696</td>
<td>53.78%</td>
<td>0.0195412</td>
<td>205,039</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Concurrent [9]</td>
<td>0.0146470</td>
<td>0.0193530</td>
<td>10.14%</td>
<td>0.0251210</td>
<td>121,200</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MCNE [8] [9]</td>
<td>0.0246412</td>
<td>0.0478444</td>
<td>-9.27%</td>
<td>0.0671124</td>
<td>100,000</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### TWI Exchange [11]

<table>
<thead>
<tr>
<th>Test type</th>
<th>Best error</th>
<th>Mean error</th>
<th>Mean error bias</th>
<th>Worst error</th>
<th>Cost</th>
<th>Cost per thread</th>
<th>Mean validation error epoch</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.0118479</td>
<td>0.0127522</td>
<td>48.92%</td>
<td>0.0153886</td>
<td>200,075</td>
<td>40,015</td>
<td>1658</td>
</tr>
<tr>
<td>B</td>
<td>0.0120645</td>
<td>0.0130102</td>
<td>42.49%</td>
<td>0.0153533</td>
<td>195,075</td>
<td>39,015</td>
<td>1576</td>
</tr>
<tr>
<td>C</td>
<td>0.0115597</td>
<td>0.0123720</td>
<td>46.37%</td>
<td>0.0145887</td>
<td>500,150</td>
<td>50,015</td>
<td>1701</td>
</tr>
<tr>
<td>D</td>
<td>0.0121878</td>
<td>0.0134124</td>
<td>53.20%</td>
<td>0.0174216</td>
<td>200,075</td>
<td>40,015</td>
<td>1465</td>
</tr>
<tr>
<td>E</td>
<td>0.0113766</td>
<td>0.0133122</td>
<td>40.42%</td>
<td>0.0178735</td>
<td>200,075</td>
<td>40,015</td>
<td>1604</td>
</tr>
<tr>
<td>Sequential [9]</td>
<td>0.0354120</td>
<td>0.0394227</td>
<td>−38.23%</td>
<td>0.0412148</td>
<td>272,318</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Concurrent [9]</td>
<td>0.0363142</td>
<td>0.0397674</td>
<td>0.59%</td>
<td>0.0432614</td>
<td>121,200</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MCNE [8] [9]</td>
<td>0.0745214</td>
<td>0.0852743</td>
<td>−28.59%</td>
<td>0.0912457</td>
<td>100,000</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

The Cost column is calculated by the type of cost mentioned above. Also, there is an additional “Cost per thread” column, because as the initial cost is calculated, it’s like implying that the execution of the program is serial, while in fact it is a parallel task, because of the threads. The Mean (best validation error epoch) (MBVE) column is the average of the epochs, from which the best validation error occurred. It follows that the larger the MBVE, the less the system dependence on the validation error, in order to find the optimal network. All of the above explains, whether the validation dataset is useful for the system and therefore the validation error (Figure 4).

### 5.5. Initial Conclusions

In their most recent study [9] the authors have shown that their system performs better with their Sequential strategy than the other two strategies (Concurrent, MCNE), considering MCNE worse. Although our system is generally based on MCNE, it achieves and produces better results than all of the strategies used in [9], both in the average error and the smallest error, at approximately the same cost. The dominant tests are C and A.
5.6. Additional Analysis

An additional statistical analysis was performed to determine the effectiveness of our own tests with each other. One-way ANOVA [20] was used for statistical analysis, using the PSPP graphic program [21]. In order for the ANOVA method to work, it requires at least two parameters, one factor and one dependent variable. In our problem, the factor is the tests which are a qualitative variable and the dependent variable is the evaluation error, which is a quantitative variable. The first conclusion from the statistical analysis is that the test with the best results in all datasets is C. Having the best average performance, it also has the smallest dispersion, which means that it produces often and consistently the best results. The next best result comes from test A, which is not statistically very different from C, regarding its performance. The remaining tests are ranked in the order B, D and E. From the performance of D relative to E, we can conclude that the method of Percent of fitness helps, but not to a great extent, because statistically the performance of the two tests is not much different. One of the major decision problems, analyzed in [9], is how often and how many times the BP algorithm needs to be executed. By carefully examining the low performance of B, we
can see that many successive BP epochs with low frequency does not help the system, whereas the small number of BP epochs with high frequency of executions increase the system performance.

The general conclusion we get is that the larger the number of subpopulations, the better and more efficient the system is. This implies a larger population sample during migration, which helps the weaker populations, which in turn means more people for evolution than BP. Of course, this comes at a great cost only if multiple threads are not used and at the same time there is no such support from the computer hardware, in particular from its processor. Consequently, if the conditions are met, then the subpopulations will develop in real time, each in its own thread.

6. Future Work

A feature of the problem that deserves further development is that of the variable size of the chromosomes of persons, since it is associated with a proper imple-
mentation of the recombination and mutation methods, based on existing techniques [17]. This feature has to be customized so that it can work harmoniously with the neural networks. Combining EA with different networks such as recurrent neural network (RNN) [22] or long short-term memory (LSTM) [23] will also be very useful to further make the system suitable for dealing with a larger variety of problems. It would also be wise to upgrade BP to a potential learning rate, using a different method from a simple function. For example, one such effective method has been proposed by Adam [24], which has proven its effectiveness, through the multitude of libraries, which generally uses it in machine learning. Another major area for research is deep learning [5] [25] as it appears to be used in a wide range of real problems with great success.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References


A Hybrid Backward Euler Control Volume Method to Solve the Concentration-Dependent Solid-State Diffusion Problem in Battery Modeling

Kudakwashe Chayambuka1,2,3, Grietus Mulder1,2, Dmitri L. Danilov3,4, Peter H. L. Notten3,4,5*

1VITO, Mol, Belgium
2EnergyVille, Genk, Belgium
3Eindhoven University of Technology, Eindhoven, Netherlands
4Forschungszentrum Jülich, Fundamental Electrochemistry (IEK-9), Jülich, Germany
5University of Technology Sydney, Broadway, Sydney, NSW, Australia

Email: *p.h.l.notten@tue.nl

Abstract

Several efficient analytical methods have been developed to solve the solid-state diffusion problem, for constant diffusion coefficient problems. However, these methods cannot be applied for concentration-dependent diffusion coefficient problems and numerical methods are used instead. Herein, grid-based numerical methods derived from the control volume discretization are presented to resolve the characteristic nonlinear system of partial differential equations. A novel hybrid backward Euler control volume (HBECV) method is presented which requires only one iteration to reach an implicit solution. The HBECV results are shown to be stable and accurate for a moderate number of grid points. The computational speed and accuracy of the HBECV, justify its use in battery simulations, in which the solid-state diffusion coefficient is a strong function of the concentration.

Keywords

Solid-State Diffusion, Implicit Methods, Backward Euler

1. Introduction

Since the discovery of the electrochemical intercalation reaction of lithium in layered titanium disulfide (TiS2) by Whittingham in 1976, lithium-ion batteries (LIB) have grown in popularity and application to surpass all other electrical
energy storage devices [1] [2]. This groundbreaking discovery demonstrated the reversible shuttling of lithium ions in solid layered materials and sparked the development of intercalation electrode materials for room temperature secondary batteries. However, due to the low voltage of TiS$_2$ as a cathode material (approximately 2 V vs. Li$^+$/Li) and the lack of a safe anode, the Li/TiS$_2$ battery was not commercialized.

The intercalation mechanism would later be revisited in the search of a safe anode, which completely avoids the danger of lithium dendrites. These needle-like structures protrude from the metallic anode surface during plating, causing internal short circuits and battery-related fires. The coupling of two intercalation electrode materials is therefore, the fundamental basis of the ubiquitous LIB which are used in most consumer electronics and have recently captured public attention due to their application in electric vehicles. It is therefore meritorious, for the 2019 Nobel Prize in chemistry, to be awarded to John Goodenough, M. Stanley Whittingham and Akira Yoshino, for developing the LIB [3]. We should also mention the nickel metal hydride batteries (NiMH), the emerging sodium-ion batteries (SIB) and potassium-ion batteries, which likewise make use of intercalation active materials.

Inasmuch as intercalation of energy carrying species inside electrodes is fundamental to the operation of the aforementioned battery chemistries, the mathematical modeling of this mechanism is equally critical to the success of physics-based battery models. The time-dependent radial transport of species inside spherical active particles is governed by Fick’s second law

\[
\frac{\partial c}{\partial t} = \frac{1}{r^2} \frac{\partial}{\partial r} \left( D r^2 \frac{\partial c}{\partial r} \right),
\]

where $c = c(r,t)$ is the concentration of the intercalated species [mol·m$^{-3}$], $D = D(c)$ is the solid-state diffusion coefficient [m$^2$·s$^{-1}$], $r$ is the radial distance from the center of the particle [m] and $t$ is time [s]. Equation (1) does not assume a constant diffusion coefficient, which indeed, may vary as a function of the concentration. This has consequences on the complexity of the numerical solution method as we shall later explore in detail. For a physics-based battery model, we should highlight that the ultimate goal is to derive the particle surface concentration for a given surface flux boundary condition (Neumann boundary condition) using a numerical solution method that is both robust and computationally inexpensive. The surface concentration is the most important parameter which governs reaction kinetics and the state-of-charge in electrochemical battery models.

Several efficient methods for solving Equation (1) in the case of a constant diffusion coefficient exist and have been applied in many battery models [4] [5] [6]. A constant diffusion coefficient is the customary assumption in battery models, a simplification which allows analytical and numerical solutions from heat transfer theory to be used [7]. Such numerical methods are detailed in the classic reviews by Subramanian et al. [8] and Zeng et al. [9]. In our assessment, Liu’s pseudo steady-state method (PSS) resolves the seemingly antagonistic re-
quirements of computational speed and numerical accuracy [4]. A recent modification to the PSS method, provides faster and stable results for long time simulations [5].

A challenging problem which has remained relatively unscrutinized for many years, is how to efficiently simulate a variable diffusion coefficient [10] [11] [12]. This problem deserves increased attention because of the need to address stress effects in particles, [13] phase separation in two-phase materials and temperature effects occurring at high discharge rates. Overwhelming evidence corroborates that a constant diffusion coefficient is a rare exception in intercalation particles. For example, the diffusion coefficient in nickel hydroxide particles used in NiMH batteries, varies by 3 orders of magnitude over the span of the state-of-charge [14] [15]. In addition, the diffusion coefficient in Na$_3$V$_2$(PO$_4$)$_2$F$_3$ active particles in SIB varies by 2 orders of magnitude, even though ex situ XRD patterns reveal negligible variation in lattice parameters, a characteristic of solid-solution processes [16] [17].

While incorporating a concentration-dependent diffusion coefficient in electrochemical models leads to more accurate simulation results, the lack of an analytical exact solution method for Equation (1), for a general, nonlinear form of the diffusion coefficient, means one has to rely on numerical methods. [18] Traditionally, this is solved by the explicit and implicit finite difference methods (FDM) [6] [19]. However, explicit schemes are conditionally stable and therefore, computationally expensive. For a grid spacing $\Delta x$, the von Neumann stability condition for the explicit scheme

$$\Delta t \leq \frac{\Delta x^2}{2D},$$

imposes a strong time step restriction for particle sizes of the order of 0.5 - 5 $\mu$m in radius. In general, FDM do not conserve a perfect mass balance and this error is propagated to the overall battery simulation at long times [20]. Other grid-based methods, with a perfect mass conservation include, the finite element method (FEM) and the finite volume method (FVM). While both methods are renowned for their robustness, they have the inherent disadvantage of not calculating concentrations at specific node points, in particular, at the surface boundary [20] [21]. Instead, one obtains volume averaged concentrations within discrete volume elements. Despite the additional computations and approximation errors arising from approximating the surface concentration, these methods have been successfully applied to battery simulations, with variable diffusion coefficients.

To address the shortcomings of the FEM and FVM, Zeng et al. [21] proposed the control volume method (CVM). The CVM is a class of finite volume discretization, which computes concentrations at node points. Therefore, surface concentrations are obtained directly in the CVM. Compared to the FVM, the CVM has a higher accuracy for a given number of mesh points and is more suited for battery modeling. [21] While these authors focused on the Crank-Nicolson time domain discretization, which sometimes produces oscillatory solutions, we
herein present a backward Euler control volume method (BECV) to resolve the spherical diffusion problem with a variable diffusion coefficient. This method incorporates all the advantages of the CVM, with the added advantage of being stable and easier to implement. Because obtaining a fully-implicit solution involves a series of iterative steps, a hybrid backward Euler control volume method (HBECV) is herein introduced for the first time. The HBECV method is based on the linearization of the functional form of the diffusion coefficient and obtains the implicit solution in a single iteration. Using the HBECV, computationally efficient, accurate and unconditionally stable results are obtained for the surface concentration. While this work focuses on the single particle diffusion problem, the solution extension to a full, physics-based battery model is trivial.

The challenge with all grid based methods is that, the number of states in a system of coupled differential equations dramatically increases, when the number of spatially distributed grid points increases [6] [21]. In order to reduce the number of grid points, while maintaining a high accuracy, it is recommended to implement a non-regular grid spacing, which allocates more points towards the surface boundary, where the concentration gradients will be steep initially. Therefore, we present the CVM for non-uniform grid spacing and further perform a mesh optimization study in which factors affecting optimum mesh spacing are identified. It is shown that the diffusion length is a primary factor, which determines the optimum grid spacing.

2. Model

The second order, time-dependent, partial differential equation for species diffusion inside a spherical particle, is governed by Fick’s second law which is expressed in Equation (1). Two Neumann-type boundary conditions are relevant for a battery simulation, at the surface and at the center of the particle. At the surface, due to interfacial electrochemical reactions, the rate of species transport across the surface is expressed as a flux. This boundary condition is written as

$$-D \frac{\partial C}{\partial r} = J, \text{ at } r = R, \forall t,$$

where $J$ is the interfacial flux of species [mol·m$^{-2}$·s$^{-1}$] and $R$ is the radius of the particle [m].

We define $J$ as positive for species diffusing out of the particle. $J$ can also be expressed in terms of the current density and its magnitude is the same at all points of the surface, i.e. it is uniform. In turn, this implies spherical symmetry.

At the center of the particle, due to the flux symmetry, the flux is zero

$$-D \frac{\partial C}{\partial r} = 0, \text{ at } r = 0, \forall t.$$

To obtain the concentration profile inside the particle, we apply the CVM. For Neumann-type boundary conditions, the CVM performs better compared to the FDM because mass conservation breaks down in the FDM.

Since the diffusion problem of Equations (1), (3)-(4) is spherically symmetric,
the magnitude of flux depends only on \( r \). Consider a set of \( N \) discretization points on \( r \), such that

\[
\{r_i\}^N, \text{ for } 1 \leq i \leq N
\]  

where the index \( i \) defines node positions of CVM, and \( N \) is a non-zero natural number of grid points. Accordingly, \( r_0 = 0 \) and \( r_N = R \) are the particle center and surface, respectively.

Each \( i^{th} \) node point in Equation (5) can be assigned a control volume element to it. For a spherical geometry, and for \( 2 \leq i \leq N - 1 \), such a control volume element is a shell whose external faces (boundaries) are located halfway between adjacent nodes.

Let \( IB_i \) and \( OB_i \) define the \( i^{th} \) inner control volume boundary and outer control volume boundary, respectively, according to

\[
IB_i = r_{i-1} + \frac{\Delta r_i}{2}, \text{ for } 2 \leq i \leq N - 1
\]  

\[
OB_i = r_i + \frac{\Delta r_i}{2}, \text{ for } 2 \leq i \leq N - 1
\]

where \( \Delta r_i = r_{i+1} - r_i \) is the spacing between two adjacent node points. In this way, the \( i^{th} \) inner control volume element is a shell imbedding node point \( i \).

At the boundaries, exceptions arise. Both \( IB_1 \) and \( OB_N \) are not located between node points but right at node points \( r_1 \) and \( r_N \), respectively. This implies, \( IB_1 = 0 \) and \( OB_N = R \). On the other hand, \( OB_1 \) and \( IB_N \) are located between adjacent node points and can be expressed as

\[
OB_i = \frac{\Delta r_i}{2},
\]

\[
IB_N = r_{N-1} + \frac{\Delta r_{N-1}}{2}.
\]

**Figure 1** illustrates the CVM discretization. Black dots and solid black lines represent node points while white dots and dotted black lines represent control volume boundaries. A magnified view of the discretized particle illustrates the 3D nature of spherical shells arising from inner control volume discretization. Several important features of the control volume discretization should be noted:

1) Concentrations are calculated at the node points only. No concentrations are calculated at the boundaries between control volume elements.

2) The concentration profile between nodes is assumed to be linear.

3) Concentration gradients are calculated at spherical shell boundaries using concentration values from adjacent nodes.

4) Interior boundaries of the control volume shells are located halfway between adjacent nodes.

Let \( v_i \) denote the control volume element at node point \( i \). Assume \( m_i(t) \), the amount of electrochemically active species [mol] inside \( v_i \) at arbitrary time \( t \), can be expressed as a product of the concentration at node point \( i \) and volume of corresponding spherical shell. This means
Figure 1. Diffusion in a spherical particle illustrating the control volume discretization along the particle radius. The solid black lines and solid black dots represent grid point $i$, while the dotted black line and white dots represent control volume boundaries. In the magnified view, $i$ is surrounded by an imaginary control volume between the outer boundary $OB_i$ and inner boundary $IB_i$. The flux is defined positive for species diffusing out of the particle.

$$m_i(t) = \int c(r) dV(r) \approx c(r) \int dV(r) = c(r) U_i,$$

where $U_i$ is the total volume of $V_i$ [m$^3$].

Mass conservation law in the absence of source term(s) dictates that any change in $m_i$ corresponds to the net-flux via control of the volume inner and outer boundaries, i.e.

$$\Delta m_i(t) = \Delta c_i U_i = (J_i A_i - J_{i+1} A_{i+1}) \Delta t, \quad \text{for } 1 \leq i \leq N$$

where $J_i$ represents mass flux through control volume boundaries [mol·m$^{-2}$·s$^{-1}$], $\Delta t$ is the time step [s] and the symbol $\Delta$ denotes a change of a variable in time. $A_i$ is the surface area of the $i^{th}$ control volume boundary [m$^2$] which is defined as

$$A_i = 4\pi \left( r_i + \frac{\Delta r_i}{2} \right)^2, \quad \text{for } 1 \leq i \leq N - 1,$$

at the surface boundary $OB_N$ as

$$A_N = 4\pi R^3,$$

and at the center boundary $IB_1$ as

$$A_0 = 0.$$

Note that, the absence of source term(s) in Equation (11) is due to the lack of internal species production or consumption within the active particles.

Now, according to remarks ii and iii, the fluxes at the control volume boundaries can be derived as follows

$$-D_i \left( \frac{\partial c_i}{\partial r} \right)_{r = r_i, \Delta r_i} \approx -D_i \frac{c_{i+1} - c_i}{r_{i+1} - r_i} = -D_i \frac{c_{i+1} - c_i}{\Delta r_i},$$

where $D_i \frac{c_{i+1} + c_i}{2}$ is the concentration-dependent diffusion coefficient at the control volume boundary [m$^2$·s$^{-1}$]. A half-sum is applied because of remarks ii and iv. Substituting Equation (12) and (15) into Equation (11), gives
\[-D_t \left( \frac{c_t - c_{t+1}}{\Delta r_{t+1}} \right)^2 \left( r_{t+1} + \frac{\Delta r_{t+1}}{2} \right) \Delta t + D_t \left( \frac{c_{t+1} - c_t}{\Delta r_t} \right)^2 \left( r_t + \frac{\Delta r_t}{2} \right) \Delta t = \frac{U_t}{4\pi} \Delta c_t, \quad \text{for } 2 \leq i \leq N - 1. \]  

(16)

In order to eliminate the factor \(4\pi\) from subsequent derivations, let a normalized volume \(V_i\) be defined as

\[V_i = \frac{1}{3} \left[ \left( r_t + \frac{\Delta r_t}{2} \right)^3 - \left( r_{t-1} + \frac{\Delta r_{t-1}}{2} \right)^3 \right], \quad \text{for } 2 \leq i \leq N - 1, \]  

(17)

at the center \((i = 1)\) and at the surface boundary \((i = N)\), \(V_i\) and \(V_N\) are defined as

\[V_i = \frac{1}{24} \Delta r_t^3, \]  

(18)

\[V_N = \frac{1}{3} \left[ R^3 - \left( R - \frac{\Delta r_{N-1}}{2} \right)^3 \right]. \]  

(19)

To further economize notations, let variable \(K_i \quad [m^3]\) be introduced

\[K_i = D_{i+\frac{1}{2}} \frac{\Delta t}{\Delta r_t} \left( r_t + \frac{\Delta r_t}{2} \right)^2, \quad \text{for } 1 \leq i \leq N - 1. \]  

(20)

Therefore, taking Equations (17) and (20) into account, Equation (16) can be expressed in terms of \(K_i\) and \(V_i\) as

\[-K_{i-1} (c_{i-1} - c_{i+1}) + K_i (c_{i+1} - c_t) = V_i \Delta c_t, \quad \text{for } 2 \leq i \leq N - 1. \]  

(21)

Rearranging Equation (21) we finally obtain

\[K_{i-1}c_{i-1} - (K_{i-1} + K_i)c_i + K_i c_{i+1} = V_i \Delta c_t, \quad \text{for } 2 \leq i \leq N - 1. \]  

(22)

Equation (22) is defined at interior node points. It is possible, starting with the general mass balance expression of Equation (11) and following the steps shown in Equations (15)-(22), to obtain expressions for the two remaining boundary cases, \(i = 1\) and \(i = N\).

At the center (at \(i = 1\)), there is zero flux through \(IB_1\) according to Equation (4). Furthermore, the surface area at \(IB_1\) is zero according to Equation (14). Applying the general mass balance on \(V_i\) gives

\[D_{i+\frac{1}{2}} \frac{c_t - c_{i+1}}{\Delta r_t} \left( r_t + \frac{\Delta r_t}{2} \right)^2 \Delta t = V_i \Delta c_t, \]  

(23)

Finally, Equation (23) expressed in terms of the variable \(K_i\), becomes

\[K_i (c_{i+1} - c_t) = V_i \Delta c_t. \]  

(24)

At the surface (at \(i = N\)), there is a uniform interfacial flux \(J\). Applying the boundary condition of Equation (3) and the general mass balance on \(V_N\), we obtain

\[-D_{N-\frac{1}{2}} \frac{c_N - c_{N-1}}{\Delta r_{N-1}} \left( r_{N-1} + \frac{\Delta r_{N-1}}{2} \right)^2 \Delta t - JR^2 \Delta t = V_N \Delta c_N, \]  

(25)
Finally, introducing the variable \( K_i \) into Equation (25) gives

\[
-K_{N-1} (c_N - c_{N-1}) - JR^2 \Delta t = V_N \Delta c_N.
\]  

(26)

Up to this point, the temporal discretization is intentionally omitted because the system of equations, Equations (22), (24) and (26) can be solved either by the forward Euler or the backward Euler method.

Let the superscript \( j \) represent the current time step and \( j - 1 \) represent the previous time step. Therefore, \( \Delta c_i = c_i (t) - c_i (t - \Delta t) = c_i^j - c_i^{j-1} \). The system of equations, Equations (22), (24) and (26) is thus expressed in the backward Euler scheme as

\[
-K_i c_i^{j-1} + (K_i^j + V_i)c_i^j - K_{i+1} c_{i+1}^j = V_i c_i^{j-1}, \quad \text{for } 2 \leq i \leq N - 1
\]

(27)

at the center

\[
(K_i^j + V_i)c_i^j - K_{i+1} c_{i+1}^j = V_i c_i^{j-1},
\]

(28)

and at the surface

\[
-K_{N-1} c_N^{j-1} + (K_{N-1}^j + V_N)c_N^j = V_N c_N^{j-1} - JR^2 \Delta t.
\]

(29)

Equations (27)-(29) represent a coupled system of equations since all values at time step \( j \) are unknown while values at time step \( j - 1 \) are unknown.

### 2.1. Solving the Coupled System of Equations

As a first step to finding the solution, the coupled system of Equations (27)-(29) is expressed in matrix form. This is expressed as

\[
M c^j = V c^{j-1} - JR^2 \Delta t,
\]

(30)

where \( c^j \) is a column vector containing concentrations at all node points at time index \( j \), i.e. \( c^j = (c_1^j, c_2^j, \ldots, c_N^j) \), \( M \) is an \( N \)-by-\( N \) matrix

\[
\begin{pmatrix}
K_1^j + V_1 & -K_1^j & 0 & \cdots & 0 & 0 \\
-K_1^j & K_1^j + V_2 & -K_2^j & \cdots & 0 & 0 \\
0 & -K_2^j & K_2^j + V_3 & -K_3^j & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & 0 & \cdots & -K_{N-2}^j - K_{N-1}^j + V_{N-1} \\
0 & 0 & 0 & 0 & \cdots & -K_{N-1}^j + V_N \\
\end{pmatrix}
\]

\( V \) is expressed as

\[
V = \begin{pmatrix}
V_1 & 0 & \cdots & 0 \\
0 & V_2 & \cdots & \vdots \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & V_N
\end{pmatrix}
\]

and \( J \) is a (column) vector, \( (0, 0, \ldots, 0, J)' \), whose only non-zero entry is at the \( N \)-th point, corresponding to the surface.

\( M \) is a tridiagonal matrix. If \( V_i > 0 \), a condition which is trivially satisfied by the construction of sequence \( r \) as shown in Equation (5), \( M \) is strictly (row) diagonally dominant. By the Levy-Desplanques theorem, \( M \) is non-singular.
and therefore invertible [22]. The tridiagonal matrix algorithm (TDMA) can thus be applied to solve Equation (30) as a stable and fast solution method [23]. The TDMA, also known as the Thomas algorithm, is a variant of Gaussian elimination, which is applicable to a diagonally dominant tridiagonal system of $N$ unknowns. Compared to the standard Gaussian elimination or matrix inversion, which require $O(N^3)$ operations to solve, the TDMA only requires $O(N)$ operations [24].

For a constant diffusion coefficient, Equation (30) is linear. The solution $c^j$ is rapidly obtained in a single run of the TDMA. However, for the problem posed above, the diffusion coefficient at time step $j$ is not known a priori since $D$ is an implicit function of $c^j$. Equation (23) therefore takes form

$$M(c^j)c^j = Vc^{j-1} - JR^2\Delta t,$$  \hspace{1cm} (31)

and represents a non-linear system of equations. Equation (31) can be iteratively solved by various fixed point methods such as Newton’s method, Jacobi, line-by-line and Gauss-Seidel. These iterative methods can be used in combination with the TDMA to obtain convergent solutions [23] [25]. The Newton’s method in particular, attains quadratic convergence if the initial guess is sufficiently close to the solution [21] [25]. Nevertheless, great care must be taken to correctly construct and solve the Newton equation and a poor initial guess may even result in a lack of convergence [25]. In this work, the following iterative scheme (Jacobi) is used

$$M(c_{j,k}^j)c_k^j = Vc^{j-1} - JR^2\Delta t, \quad \text{for } 1 \leq k \leq k_{tot}$$  \hspace{1cm} (32)

where subscript $k$ denotes the iteration number in a total of $k_{tot}$ iterations. $c_k^j$ is therefore the implicit solution obtained after $k$ iterations. The initial value needed for the first iteration, is defined as

$$c_{0}^j = c^{j-1}.$$  \hspace{1cm} (33)

For $k_{tot} \approx 20$, the fully-implicit, BECV solution is obtained. It shall be demonstrated that, due to the initial condition Equation (33), the first iteration of Equation (32) achieves a stable and approximately accurate solution, which is acceptable in many cases. This solution is herein referred to as the HBECV. The HBECV is therefore a linearization of the concentration-dependent $M$, in order to obtain implicit solutions in a single iteration.

### 2.2. Grid Spacing

In order to accurately determine concentration profiles, many grid points are required. However, more grid points come at considerable computational costs. The accuracy of the CVM, for an economical number of grid points, depends on the spatial distribution of the same. The choice of grid spacing, however, depends on the nature of the problem and boundary conditions. More points are required at the regions where the concentration profile has steep gradients and this holds true at the particle surface boundary.
While the scheme of equations presented in this work, allows for variable spacing of grid points, the literature around grid/mesh optimization is very sparse, this in turn makes it difficult to comprehend the principle factors affecting the optimum grid spacing. In order to evaluate the different grid-point locations, the following geometric spacing equation is applied

\[ r_i = R \left( 1 - \frac{Y_{i-1}}{Y_{i-1} - 1} \right), \quad 1 \leq i \leq N, \]  

(34)

where \( Y \) is the common factor of the geometric series. In this study, \( Y \) varies between 2 and 20. If \( Y = 10 \), the logarithmic spacing is obtained, a notable choice in the preceding publication [21]. To evaluate the error in each value of \( Y \), as a function of \( D \) and \( R \), a solution obtained from a linear spaced grid of 501 points and \( dt = 5 \) s is used as reference solution.

3. Results and Discussion

To investigate the accuracy of HBECV, the set of parameters from Zeng et al. [21] is used. Table 1 shows the parameters for Li(Ni\(_{1/3}\)Mn\(_{1/3}\)Co\(_{1/3}\))O\(_2\) particles used in the referenced work and also in this work.

The concentration-dependent diffusion coefficient for Li(Ni\(_{1/3}\)Mn\(_{1/3}\)Co\(_{1/3}\))O\(_2\) used by Wu et al. (see ref. [26]) is

\[ D(c) = D^{\text{ref}} \left[ 1 + 100 \left( \frac{C_{\text{theo}}(c_{\max} - c)}{C_{\text{prac}}c_{\max}} \right)^{3/7} \right], \]  

(35)

where \( D^{\text{ref}} \) is defined as the reference diffusion coefficient of \( 2 \times 10^{-16} \) m\(^2\)·s\(^{-1}\), \( C_{\text{theo}} \) is the theoretical capacity of the electrode material (277.84 mAh·g\(^{-1}\)) and \( C_{\text{prac}} \) is the practical capacity of the electrode material of 160 mAh·g\(^{-1}\).

Figure 2 illustrates the agreement between our results and literature. After a discharge time \( t = 400 \) s, the surface concentration nearly reaches \( c_{\max} \) and thus the end of discharge. For the parameters in Table 1, a dense mesh of 501 grid points is needed to eliminate errors due to spatial discretization. These results validate the BECV method whose solutions are obtained after \( k_{\text{tot}} = 20 \) iterations of the Jacobi method.

It is interesting to evaluate the effect of reducing \( k_{\text{tot}} \) because the computation speed increases with less iteration. Figure 3 shows that the relative error in surface concentration progressively increases when \( k_{\text{tot}} \) decreases. Furthermore, approximately 10 iterations of the Jacobi method are necessary to obtain a fully-implicit BECV solution. With regards to fast simulations, the HBECV solution obtained when \( k_{\text{tot}} = 1 \) is interesting. From the results, the relative error on the surface concentration of the HBECV method is approximately 0.1%, which is good enough for practical purposes and moreover the results are stable. This validates the HBECV and justifies its use for fast, stable and practically accurate results.
Table 1. Parameters for modeling diffusion in spherical particle with variable diffusion coefficient.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Value</th>
<th>Units</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>R</td>
<td>$5 \times 10^{-6}$</td>
<td>m</td>
<td>Particle Radius</td>
</tr>
<tr>
<td>$J$</td>
<td>$-5.35 \times 10^{-5}$</td>
<td>mol·m$^{-2}$·s$^{-1}$</td>
<td>Interfacial flux</td>
</tr>
<tr>
<td>$c_0$</td>
<td>$2 \times 10^4$</td>
<td>mol·m$^{-3}$</td>
<td>Initial concentration</td>
</tr>
<tr>
<td>$c_{\text{max}}$</td>
<td>$4.665 \times 10^4$</td>
<td>mol·m$^{-3}$</td>
<td>Maximum concentration</td>
</tr>
<tr>
<td>$\Delta t$</td>
<td>5</td>
<td>s</td>
<td>Time step size</td>
</tr>
</tbody>
</table>

Figure 2. Comparison of the results of the simulated concentration gradients obtained in this work, using the BECV method and results from Zeng et al. [21] Results are obtained using 501 uniformly spaced grid points, using the parameters of Table 1 and the concentration-dependent diffusion coefficient expression, Equation (35).

Figure 3. Relative error in surface concentration over 400 s simulation as a function of the number of iterations. The relative error of the HBECV method is compared to the iterative implicit BECV method. 501 uniformly spaced grid points and parameters in Table 1 are used to calculate the surface concentration. As $k_{\text{tot}}$, the total number of iterations per time-step increases, the solution converges to the reference solution of 20 iterations.

A uniform grid of 501 points is nevertheless impractical for use in battery simulations where the number of grid points is limited. The first step in reducing the number of grid points is to determine optimal grid spacing. This is herein determined by changing the value of $Y$ in a model with 301 grid points. The optimization of $Y$ is performed using the uniform grid of 501 points as the refer-
ence solution, while the cost function is the normalized root of squared deviations between these solutions. 

Figure 4 illustrates the dependence of optimum $Y(Y_{\text{opt}})$ on the dimensionless parameter $\sqrt[4]{D_{\text{ref}} T / R}$. This parameter represents a scaled diffusion length since the mass diffusion length is known to be equal to $\sqrt[4]{D_{\text{ref}} T}$. Two regions can be easily identified in Figure 4.

1) high diffusion length region where $\sqrt[4]{D_{\text{ref}} T / R} > 0.1$ and $3 \leq Y_{\text{opt}} \leq 6$,
2) low diffusion length region where $\sqrt[4]{D_{\text{ref}} T / R} \leq 0.1$ and $Y_{\text{opt}} > 6$.

For a high diffusion length, which is more than 10% of $R$, an evenly spaced grid defined by low $Y$ values should be used. On the other hand, for a low diffusion length, which is less than 10% of $R$, a logarithmic grid spacing, defined by high $Y$ values, is more appropriate. The immediate conclusion is that more points are needed close to the surface only when the diffusion length is low. This highlights the importance of carefully selecting the grid spacing for a given $D_{\text{ref}}$ and $R$ values and not rely on an intuitive feeling of the grid spacing.

For the parameters listed in Table 1, $Y_{\text{opt}} = 12$ and the number of grid points distributed by $Y_{\text{opt}}$ is defined as $N_{\text{opt}}$. It is important to further reduce $N_{\text{opt}}$ to a practical number, relevant to full-cell battery modeling. Figure 5 shows the relative error on surface concentration at $t = 400$ s and $dt = 0.1$ s, as a function of $N_{\text{opt}}$. As expected, the error relative to the fine grid mesh increases as $N_{\text{opt}}$ decreases. However, the BECV and the HBECV remarkably converge to the same error when $N_{\text{opt}}$ decreases. This implies is that, the spatial discretization error exceeds the linearization error of the HBECV when the number of optimally spaced grid points is below 21. Therefore, based on the results shown in Figure 5, the HBECV should be used in battery simulations, instead of the more computationally expensive BECV.

**Figure 4.** Grid optimization, effect of the geometric factor $Y$ and the dimensionless factor $\sqrt[4]{D_{\text{ref}} T / R}$. The red squares show the optimum geometric factor, $Y_{\text{opt}}$ for selected $D_{\text{ref}}$ and $R$ values. The bars represent variance within 1% of $Y_{\text{opt}}$. Inset showing the distribution of 6 grid points for $Y = 2$, and $Y = 10$. 

DOI: 10.4236/jamp.2020.86083
Figure 5. Relative error on surface concentration at 400 s of the BECV and the HBECV as a function of $N_{\text{opt}}$, the number of optimized grid points. The optimum geometric factor, $Y_{\text{opt}} = 12$ and time step, $dt = 0.1$ s are used. The reference solution is obtained from 501 uniform grid points and $dt = 0.1$ s.

4. Conclusions

In this paper, the backward Euler control volume (BECV) and the hybrid backward Euler control volume (HBECV) methods are presented as efficient numerical methods to resolve the solid-state spherical diffusion problem for a variable diffusion coefficient. The implicit scheme of nonlinear equations is herein shown to be strictly diagonally dominant and efficiently solved by the tridiagonal matrix algorithm (TDMA).

The fully-implicit BECV is shown to require more computations, approximately 10 iterations of the TDMA, in order to reach a convergent solution; however, the HBECV only requires one iteration. The HBECV achieves this by linearization of the implicit scheme of equations. Although, in comparison to the BECV method, the error in surface concentration in the HBECV method is around 0.1%, for a fine grid of 501 points, the error difference decreases when the number of grid points decreases. For a course grid of 21 optimally spaced grid points, the error in surface concentrations converges to the same order of magnitude, which demonstrates that the error due to spatial discretization outweighs the error due to the linearization, as introduced by the HBECV method.

This work further explores the parameters governing optimal grid spacing. The diffusion length emerged as a guiding parameter for selecting optimum grid spacing. It is found that low diffusion length problems require more points distributed near the surface, compared to high diffusion length problems wherein an evenly spaced grid more appropriate.

As demonstrated in this work, the HBECV is accurate, stable and easy to implement. In practical battery simulations, where the number of grid points is minimized, the HBECV performs as well as the BECV method. Therefore, given the aforementioned advantages of the HBECV, this method is a justified choice for modeling the second order partial differential equation of solid-state diffusion with a concentration-dependent diffusion coefficient.
Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References


Ballistic Principle of the Property Balance in Space and Its Application to Modeling of Fluid Dynamics Problems

Nikolai Kislov
Nano CVD Company, Tampa, Florida, USA
Email: nikolai95@verizon.net

Abstract
A newly discovered Ballistic Principle of the Property Balance in the Space (BPPBS) occupied by the gas is introduced to simplify and reduce computations in applications dealing with modeling of fluid dynamics problems. The integro-differential balance equations for mass, momentum, and energy, which were formulated by applying the BPPBS, are derived. The integro-differential balance equations for mass and momentum were further approximated for the collision-dominated flow regime. Then they were reduced to the corresponding vector differential equations by the method of vector differentiation with subsequent elimination of the terms belonging to the original equation. It was shown that in the collision-dominated flow regime, the derived vector differential equations of mass and momentum balance are identical to the corresponding Navier-Stokes equations. This finding validates the BPPBS and suggests that, in the collision-dominated flow regime, the formulated integro-differential forms of the balance are exact implicit solutions for corresponding Navier-Stokes equations. Six additional tests demonstrating the feasibility of the proposed method and validity of the BPPBS are presented here. The BPPBS and the methodology discussed here will be highly useful not only as the basis to solve the fluid dynamics problems, but also to model any dynamic system composed of presumably chaotically moving particles/elements, each carrying a specific amount of property/information.

Keywords
Navier-Stokes, Fluid Dynamics, Fluid Flow, Rarefied, Gas

1. Introduction
Computational Fluid Dynamics (CFD) is widely used in many practical applica-
tions ranging from basic hydrodynamic and kinematic to fundamental cosmological applications. The fundamental basis of any CFD tool is a solver, which solves the Navier-Stokes equations that are a set of second-order partial differential equations. Navier-Stokes equations are based on the assumption that the fluid is a continuum, and they are formulated by considering the mass, momentum, and energy conservations for a control volume of any size. The flow is considered continuous and differentiable, allowing the mass, momentum, and energy balances to be expressed as partial differential equations. Scientists made further approximations and simplifications to the Navier-Stokes equation set until it can be solved [1]. However, this intentional simplification of a fluid model may diminish the usefulness of the results of the computations. Also, the theoretical understanding of the solutions to these equations is still inadequate. Specifically, for three-dimensional Navier-Stokes equations and given initial conditions, mathematicians have not yet proved that smooth solutions always exist, and the solutions have limited energy per unit mass.

In the most general form, the Navier-Stokes equations of mass and momentum conservation for compressible fluid are expressed as:

\[
\frac{\partial}{\partial t} \rho + \nabla \cdot (\rho \vec{u}) = 0
\]

and

\[
\frac{\partial}{\partial t} \vec{u} + (\vec{u} \cdot \nabla) \vec{u} + \frac{1}{\rho} \nabla p - \vec{g} = \frac{1}{\rho} \nabla \cdot \mathbf{T},
\]

respectively, for \( t > 0 \) and \( \vec{r} \in \mathbb{R}^3 \). In the equations above, \( \rho \) is the density of the fluid, \( \vec{u} \) is mass flow velocity, \( \nabla p \) is the pressure gradient, \( \nabla \cdot \mathbf{T} \) represents stresses inside the fluid, and \( \vec{g} \) is the external force per unit mass. The right-hand term \( \nabla \cdot \mathbf{T} \) has generally unknown functionality and contains too many unknowns and up to date does not apply to practical problems.

The general major disadvantage of any existing "mathematical" approach in solving the fluid dynamics problem is based on using an infinitesimal fluid element viewed as a continuous medium, to which fundamental physical principles are applied. This approach contradicts the molecular or particle nature, thus providing a source of significant uncertainty in interpreting the results of modeling and calculations. In the flows at small-scale, when the characteristic hydrodynamic length scale approaches the fluid characteristic length scale, the Navier-Stokes description is expected to fail (see [2] and [3]). In rarefied gases, the mean free path or the average distance traveled between intermolecular collisions is also considered a fail factor for the Navier-Stokes description [4]. Here we shall enumerate the most important properties of gases that often behave very close to ideal [5]:

1) The ideal gas composed of a large number of identical particles (molecules or inert atoms).

2) The volume occupied by gas particles themselves is negligible compared to the volume occupied by the gas.
3) The particles obey Newton’s laws of motion, and they move in random motion.
4) The particles experience forces only during collisions; any collisions are wholly elastic and take a negligible amount of time.

Mathematicians and physicists consider the recent advancement of the method of the lattice Boltzmann (LB) equation as a significant alternative to standard computational fluid dynamics [6]. This approach consists in modeling “the physical reality at a mesoscopic level: the generic features of microscopic processes can be expressed through simple rules, from which the desired macroscopic behavior emerges as a collective effect of the interactions between the many elementary components [7].” Typical hydrodynamic quantities, such as mass, density, fluid velocity, and temperature, are not associated with individual particle movement and are quantified by simple moments of the particle distribution function [8]. However, solving the LB equation represents a significant challenge because it involves a numerical evaluation of an integral-differential equation in position and velocity phase space. Fluid simulation of such complexity can be implemented only by massively parallel data processors equipped with combinational logic for processing collision rules. This method is extremely complicated and restricted because of the limitation of the available computational approaches for modeling real physical statistical systems. Besides, it is challenging for schemes developed for solving LB equations to consider the interaction between particles and boundary conditions [9]. On the other hand, molecular dynamics (MD) simulation can be a unique tool based on the first principles involving the analysis of the physical movements of atoms and molecules. However, because model systems for engineering applications require hundreds of thousands of particles, it is considered impossible to determine the properties of such systems analytically [10]; the problem is solved by using numerical methods. The downside is that the MD simulation is hugely computationally intensive [11], and long MD simulations are mathematically ill-conditioned, generating cumulative errors in numerical integration [10].

In this paper, we discuss a physical approach to solving fluid dynamics problems by a novel analytical molecular dynamics technique (NAMDT), which was initially presented in our patent [12] and PCT publication [13]. Our method consists of approximating fluid flow as a flow of a model gas, assigning to the model gas unique properties that differ from the properties of the ideal gas, and forming integro-differential balance equations for mass, momentum, and energy transport in any non-moving point in space occupied by the model gas. In the patent publications [12] and [13], we disclosed detailed steps of the one-dimensional simulation of the model gas flow in space between parallel plates, which support our method of modeling flows by the NAMDT. This paper aims the theoretical validation of the method.

In the following, Section 2 describes the physical principles of the fluid model and the general physical approach of forming balancing equations. Here we also
introduce newly discovered Ballistic Principle of the Property Balance in the Space occupied by the gas, applying of which is expected to simplify and reduce computations in applications dealing with modeling of fluid dynamics problems. The section also provides an analytical representation of the general integro-differential forms of mass balance, momentum balance, and energy balance equations.

In Section 3, we provide integro-differential forms of mass balance and momentum balance equations, which were adapted to the collision-dominated flow regime.

In Section 4, we demonstrate seven validation tests aiming to prove the feasibility of the proposed method. The first two validation tests, 4.1 and 4.2, illustrate that interaction of the model gas and the ideal gas with the gas-solid interface are identical (compare the rate of collision per unit area and the pressure exerted on the surface); nevertheless, some the most essential properties characterizing the model gas and the ideal gas are different. In the third validation test, 4.3, we supported our approach by formulating the mass-balance and the momentum-balance in the three-dimensional unlimited incompressible gas space at the uniform temperature. Analogously, in the fourth validation test, 4.4.2, we supported our approach by formulating the mass-balance and the momentum-balance in one-dimensional incompressible gas space confined between two parallel plates at the uniform temperature.

Also, in the validation tests 4.4 and 4.5, we explain a method for obtaining an analytical solution describing the incompressible model gas flow at the uniform temperature in the channel, which is confined in the space between two infinite parallel plates. Validation test 4.4 demonstrates the model gas flow with diffuse particle scattering from the plates. Validation test 4.5 demonstrates the model gas flow in a case of mixed diffuse and specular particle scatterings from the plates being at rest.

Lastly, in the validation test 4.6, we supported the proposed approach by revealing that, in the collision-dominated flow regime, the differential equations, which were converted from the integro-differential mass and momentum balance equations, are identical to the corresponding Navier-Stokes equations.

Finally, in Section 5, we present the conclusions and highlight the prospectsiveness of the proposed physical approach for developing a new generation of the CFD software based on the NAMDT.

2. Physical Principles of the Fluid Model

We suggest an approach in which fluid flow in a fluid system is model gas flow in a model gas system, which is equivalent to the fluid system. The transport processes involve the exchange of properties such as mass, momentum, and energy between interacting particles. In a more abstract sense, all particle interactions and property randomizing are an exchange of property/information [14]. We have also realized that 1) transport phenomena include any situation
that involves a net transfer of property/information between particles, which equals randomizing properties between interacting particles; 2) a randomized by property exchange between interacting particles physical/statistical property can be taken out into the surrounding gas [14]. However, we exclude interaction resulted in chemical reactions between colliding particles and fragmentation or fusion.

In the following, all references are made to absolute time, which is measured equally within the model gas system. Besides, when referencing an appropriate law of motion, Newton’s Second Law of motion is considered. Also, in the interests of simplicity, we analyzed the model gas flow at a uniform body force such as a gravitational field of force or the acceleration field if dealing with a particle of a unit mass.

2.1. Properties and Features of the Model Gas

We assign these unique properties to the model gas [12]:

1) The model gas enables a distant transport of one or more properties, including one or more of mass, momentum, and energy by particles being in a constant state of mostly random motion and interaction by collisions.

2) Each of the particles of the model gas is assigned to travel by obeying a ballistic trajectory that is governed by a law of motion in free space. It overcomes a distance between any of the two points of the ballistic trajectory with certain survival probability.

3) Each of the particles is adapted to transport a combination of one or more properties, comprising mass, momentum, and energy between a point of initial collision and a point of ending collision.

4) Each point within the space occupied by the model gas is treated as a point of collisions for converging particles, each following a ballistic trajectory with the same ending point simultaneously.

5) Each point of collisions is treated as either a point source for diverging ballistic particles or a point sink for converging ballistic particles.

6) Each of the particles moving from the point source to the point sink is treated as a property carrier. The property carrier is created in the point source during the initial collision by obtaining one or more properties of specific values being intrinsic to the model gas surrounding the point source. It is ended in the point sink during ending collision by transferring one or more properties of specific values in the point sink.

7) The value of the property, which is delivered in the point sink, or the value of the property, which is taken away from the point source, is evaluated regarding whether the value of property carried by each of the particles is modified because of interaction with an external field.

8) The velocity of a point source equals the mass flow velocity of the model gas flow in a corresponding point of the initial collisions at the time of the initial collision.
One can note from the above that the model gas properties differ from the properties typically assigned to the ideal gas (see above in Introduction).

In this paper, we investigate the transport of properties that are conserved during the ballistic traveling time.

**Figure 1** illustrates a schematic of the ballistic trajectories of a particle between two subsequent collisions in a model gas system.

The schematic diagram above shows the model gas composed of identical randomly moving particles and positioned in the observer’s Cartesian coordinate system 100. Note that in the paper, the observer’s coordinate system is designated by index “100.” Here we consider an isotropic model, which requires that the coordinate system needs to be at rest. To detect the position of the event, the observer reads a space coordinate at the location of an event. Also, the clocks at any location within a system are synchronized. The observer allocates space coordinates and time by recording both the space coordinates and time at the clock nearest the event position [12]. For clarity, the observer’s Cartesian coordinate system 100 is orientated, so that y-axis is directed in the opposite direction of an applied field of external force 107, which provides, for each of particles, acceleration $g_r$. Each of the particles (shown as black disks), particularly a particle 101 (shown as white disks), travels between its two consecutive collisions: an initial collision 102 and an ending collision 103 by following a trajectory 104 or 105 governed by applicable law of motion including Newton’s laws of motion. Note that in a lack of external force, all ballistic trajectories between consecutive collisions will be just straight lines, respectively, as indicated by trajectory 106. Specifically, referring to **Figure 1**, particle 101 obtains a set of properties of certain values in a point source 102 at the time of the initial collision. The property may include a scalar property of value $\Psi$ or a vector property of value $\Psi^r$, which is inherent to the model gas near the point source 102 at the time of a collision $t'_i$.

In the point source 102 positioned in point $r'_i$ at time $t'_i$ and moving with velocity $(\mathbf{u}(t'_i, r'_i))$, particle 101, as a properties carrier, obtains one or more of properties being intrinsic to the model gas surrounding the point of the initial collision at the time of the initial collision. In the point source, the particle acquires a thermal velocity component $\mathbf{v}_T(t'_i, r'_i)$ of an arbitrary direction relatively to the point source. For certainty, the magnitude of the average thermal velocity component in three-dimensional configuration can be defined as:

$$v_T = \sqrt{\frac{3k_B T}{m}},$$

where $k_B$ is Boltzmann constant, $T$ is the temperature, and $m$ is the mass of a particle.

**Figure 1** shows a property carrier 101 converging in a point sink 103, which may result in delivering in point 103 of property $\Psi_{in}$, carried by the particles at the time of the ending collision.
2.2. Principles of Construction of the Property Balance in the Model Gas

In the microscopic scale, the model gas flow is characterized by the group of particles of mass $m$, which move randomly and interact by collisions with effective collision cross-section $\sigma_c$. In each of the points in space at a given time, the particle density $n$, the magnitude of thermal velocity $v_t$, and the vector of mass flow velocity $\bar{u}$ quantify the model gas. In the interests of simplicity, unless otherwise stated, the particles are considered to have a unit mass, which, in the presence of external force, are accelerated during ballistic traveling with acceleration $\bar{g}$. We have recognized that each point in space occupied by the model gas may serve as both a sink and a collector of property delivered by converging ballistic particles from the entire model gas system and a source or a disperser into the surrounding of the property taken away by diverging ballistic particles.

Here, we reasonably may expect maintenance of a general property balance in each of the points of collisions within the model gas system. We formulate the balance, illustrated as a word equation in Figure 2 as follows [12]. In a given non-moving point $\bar{r}$ at a given time $t$, the net rate of property influx per unit volume, $B_{in}^{FS}(\bar{r}, t)$, formed the converging ballistic particles (each traveling along a ballistic trajectory with certain survival probability) from the model gas system is equated to the temporal rate of property change per unit volume $\frac{\partial}{\partial t}[nt]$ and the net rate of property efflux per unit volume, $B_{out}^{FS}$, formed the diverging ballistic particles. This statement is expressed symbolically as

$$B_{in}^{FS}(\bar{r}, t) = B_{out}^{FS}(\bar{r}, t) + \frac{\partial}{\partial t}[n(t, \bar{r})\Psi(t, \bar{r})].$$

(4)

For identification, we call the quantitative relationship above as the Ballistic Principle of the Property Balance in the Space (BPPBS) occupied by the particles in presumably chaotic motion. The BPPBS applies in general to any gas system, including the model gas systems containing heterogeneous gas-solid interfaces. Also, for clarity, we call our model as the Ballistic Model (BM).

This conceptual relationship can be expanded to the infinite space, for example, in a hypothetical system with no gravitational force. Straight-line trajectories of the particles may start from the infinity.
In the interests of simplicity, we concentrate our further analysis on the homogeneous model gas flow in the three-dimensional space having uniform gas properties on its periphery. Therefore, Equation (4) is reduced to:

\[
\Psi_{\text{in}} (\mathbf{r}, t) = \Psi_{\text{out}} (\mathbf{r}, t) + \frac{\partial}{\partial t} \left[ n(t, \mathbf{F}) \Psi (t, \mathbf{F}) \right],
\]

where \( \Psi_{\text{in}} \) is the net rate of property influx per unit volume, which is formed by the converging ballistic particles from the surrounding model gas in the given non-moving point \( \mathbf{r} \) at the given time \( t \). Still, the space of the model gas system may be separated from the infinite space by defining, for example, non-uniform gas pressure over the surface confining the system. This situation is further discussed when analyzing the momentum balance.

Here we admit the virtual nature of the balance described by Equation (4) or Equation (5). We consider that the value of property/information carried by a particle ejected from a given point in space at a given time is not a result of preceding physical interactions by collisions of all virtual converging ballistic particles capable of targeting with a certain probability the given point in space at the given time, but the result of the expectation of that value because of the cumulative effect from the surrounding space, in which each point of space complies with the BPPBS. Each given point in space at a given time is a point of reality (present) or a pivot point of consuming the results of events from the gas space that occurred in the past and sending the result of consumption and balancing from the present into the future. The value of the property/information at the given point in space at the given time, which is to be sent in the future, can be determined by solving the balance equations shown above. The analytical tools needed to formulate the balance according to Equation (5) are described in more detail below.

### 2.3. Defining a Net Rate of Total Property Influx Per Unit Volume in a General Non-Moving Point at a Given Time from the Surrounding Model Gas

We have recognized and explained afterward that there exists a combination of a specific direction of an initial instant vector of thermal velocity \( \mathbf{v}_i \) and a
vector of mass flow velocity \( \mathbf{u}(\mathbf{r}', \mathbf{r}) \), which allows each of the selected particles to arrive in the given non-moving point \( \mathbf{r} \) at the given time, \( t \) \cite{12}. These particles, which originate from the initial collisions within the whole model gas system, form the converging flux in the given non-moving point at the given time. The table of the model parameters associated with defining the net rate of total property influx per unit volume is presented in Table 1.

**Table 1.** List of the model parameters associated with defining the net rate of total property influx per unit volume.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Short description</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \nabla = \frac{\partial}{\partial x} \mathbf{i} + \frac{\partial}{\partial y} \mathbf{j} + \frac{\partial}{\partial z} \mathbf{k} )</td>
<td>the operator of vector differentiation</td>
</tr>
<tr>
<td>( t )</td>
<td>given time</td>
</tr>
<tr>
<td>( t' )</td>
<td>the time of the initial collision of the converging particle</td>
</tr>
<tr>
<td>( \mathbf{r} )</td>
<td>position of the starting point of the converging particle</td>
</tr>
<tr>
<td>( \mathbf{r}' )</td>
<td>position of the ending point of the converging particle</td>
</tr>
<tr>
<td>( \mathbf{u}(\mathbf{r}', \mathbf{r}) )</td>
<td>mass flow velocity in the point ( \mathbf{r}' ) at time ( t' )</td>
</tr>
<tr>
<td>( v_v(\mathbf{r}', \mathbf{r}) )</td>
<td>the average magnitude of the thermal velocity of converging particle in point ( \mathbf{r}' ) at time ( t' )</td>
</tr>
<tr>
<td>( Z_c(\mathbf{r}', \mathbf{r}) )</td>
<td>the rate of collisions per unit volume in the point of the collision ( \mathbf{r}' ) at the time ( t' ) of the initial collision</td>
</tr>
<tr>
<td>( \mathbf{v}(\mathbf{r}', t, \mathbf{r}) )</td>
<td>velocity vector in the ending point ( \mathbf{r}' ) at the given time ( t )</td>
</tr>
<tr>
<td>( Q(t, t') )</td>
<td>the probability of free path traveling along the ballistic trajectory of the converging ballistic trajectory starting at time ( t' ) and ending at time ( t )</td>
</tr>
<tr>
<td>( \Psi_c(\mathbf{r}', \mathbf{r}, t, \mathbf{r}) )</td>
<td>property content delivered by the converging ballistic particle in the ending point ( \mathbf{r}' ) at the given time ( t )</td>
</tr>
<tr>
<td>( \phi = t - t' )</td>
<td>traveling time between an initial and ending consecutive collisions or the ballistic traveling time</td>
</tr>
<tr>
<td>( n )</td>
<td>particles density</td>
</tr>
<tr>
<td>( m )</td>
<td>particle mass</td>
</tr>
<tr>
<td>( \sigma )</td>
<td>the cross-section of collisions</td>
</tr>
<tr>
<td>( P = \sigma \cdot n )</td>
<td>the number of particles placed within a collision tube of a unit length</td>
</tr>
<tr>
<td>( \sigma )</td>
<td>the cross-section of collisions</td>
</tr>
<tr>
<td>( V )</td>
<td>the volume of integration over space occupied by the model gas</td>
</tr>
<tr>
<td>( \mathbf{r}'(\mathbf{r}', t', \mathbf{i}) )</td>
<td>the position of a virtual ballistic particle at a time ( t' ), which has zero magnitude of thermal velocity in the starting point ( \mathbf{r}' ) at the time of the initial collision ( t' )</td>
</tr>
<tr>
<td>( \mathbf{v}'(\mathbf{r}', t', \mathbf{i}) )</td>
<td>the velocity vector of the virtual ballistic particle having a zero component of the thermal velocity at a time ( t' )</td>
</tr>
<tr>
<td>( \mathbf{h} = \frac{\mathbf{r} - \mathbf{r}'(\mathbf{r}', t, \mathbf{i})}{V - \mathbf{r}'(\mathbf{r}', t, \mathbf{i})} )</td>
<td>instant unit vector directing thermal velocity component, so a traveling particle targets point ( \mathbf{r} ) at time ( t )</td>
</tr>
<tr>
<td>( \mathbf{f} = \mathbf{r}'(\mathbf{r}', t, \mathbf{i}) )</td>
<td>the position vector of a ballistic particle at time ( t )</td>
</tr>
<tr>
<td>( \mathbf{v} = \mathbf{v}(\mathbf{r}', t, \mathbf{i}) )</td>
<td>the velocity vector of the ballistic particle at time ( t )</td>
</tr>
<tr>
<td>( \mathbf{g} )</td>
<td>an external force that applies to a particle of a unit mass</td>
</tr>
<tr>
<td>( v_{rel} )</td>
<td>the average magnitude of the velocity of the traveling particle with respect to a nearby passed particle</td>
</tr>
</tbody>
</table>
Figure 3 shows a perspective view of the model gas system for explaining a ballistic movement of the converging ballistic particle after the initial collision, which is affected by the external field of force. For clarity, the observer’s Cartesian coordinate system 100 is oriented, so y-axis is along the negative direction of the applied acceleration field 306, which provides, for each particle, acceleration $g$.

We define the net rate of property influx from the model gas in the general non-moving point $\vec{r}$ at the given time $t$ by these six steps:

Step 1: Identifying the converging ballistic trajectory and trajectory characteristics

Step 1 includes:

1) Formulating position vector $\vec{r}(\vec{r}',t',\vec{i})$ of particle 304 on trajectory 301 and velocity vector $\vec{v} = \vec{v}(\vec{i})$ (not shown) at time $\vec{i}$ by applying Equations (6) and (7), respectively, given below:

$$\vec{r} = \vec{r}(\vec{r}',t',\vec{i}) = v_r(t',\vec{r}') (\vec{i} - t') \hat{n}_i + \vec{v}_i(\vec{r}',t',\vec{i})$$

and

$$\vec{v} = \vec{v}(\vec{r}',t',\vec{i}) = v_r(t',\vec{r}') \hat{n}_i + \vec{v}_i(\vec{r}',t',\vec{i})$$

where $t \geq t' \geq t$, $\hat{n}_i$ is a unit vector defined in Table 1, and $t'$ is defined in (2) below. In Equations (6) and (7) above, $\vec{r}_i(\vec{r}',t',\vec{i})$ and $\vec{v}_i(\vec{r}',t',\vec{i})$ are defined by an appropriate law of motion.

Specifically, when Newton’s Laws of Motion govern the motion of the particles, then $\vec{r}_i(\vec{r}',t',\vec{i})$ and $\vec{v}_i(\vec{r}',t',\vec{i})$ are defined by Equations (8) and (9) below:

$$\vec{r}_i(\vec{r}',t',\vec{i}) = \vec{r}' + u(t',\vec{r}') (\vec{i} - t') + \frac{1}{2} g (\vec{i} - t')^2$$

and

$$\vec{v}_i(\vec{r}',t',\vec{i}) = u(t',\vec{r}') + g (\vec{i} - t')$$

2) Determining the time of the initial collision $t'_i$ in point $\vec{r}'$. It can be done by solving Equation (6) in which $\vec{i} = t$. Where a model gas system is governed by Newton’s Law of Motion, it can be done by resolving, for each of the ballistic particles, the equation of projectile motion, given by Equation (10), with respect to the ballistic traveling time, $\varphi_i$:

$$\frac{1}{2} g \varphi_i^2 + [v_r(t',\vec{r}') \hat{n}_i + u(t',\vec{r}')] \varphi_i + \vec{r}' - \vec{r} = 0,$$

which is obtained by substitution of Equation (8) in Equation (6) followed by the assignment of $\vec{i} = t$ and substitution of $\varphi_i$ defined in Table 1 in Equation (6) and rearrangement of the terms. Upon resolving Equation (10) and selecting meaningful values for $\varphi_i$, $t'_i$ is computed as follows

$$t'_i = t - \varphi_i$$

for each of the converging ballistic particles.
3) Defining an instant unit vector directing thermal velocity component of each particle in point \( \mathbf{r}' \) at time \( t'_i \) by presenting Equation (7) in the following form:

\[
\varphi_i \mathbf{v}_i \mathbf{n}_i = \mathbf{r} - \mathbf{r}'_i
\]  

(12)

where \( \mathbf{r}'_i \) is defined as:

\[
\mathbf{r}'_i(t, t', t) = \mathbf{r}' + \mathbf{u}(t', \mathbf{r}') \varphi_i + \frac{1}{2} \mathbf{g} \varphi_i^2
\]  

(13)

In this, vector \( \mathbf{r}'_i \) is interpreted as the location of the center of the expansion zone at time \( t \) or the location of a particle having zero magnitude of an arbitrary or thermal velocity in point \( \mathbf{r}' \) at the time \( t'_i \) of the divergence, which is observed at time \( t \).

4) Defining the size of the expansion zone \( R'^{np}_i \), by executing scalar multiplication of Equation (12) on itself resulted and averaging as:

\[
\varphi_i^2 \mathbf{v}_i^2 = (\mathbf{r} - \mathbf{r}'_i)^2
\]  

(14)

and computing \( R'^{np}_i \) from Equation (14) as

\[
R'^{np}_i = \varphi_i \mathbf{v}_i = |\mathbf{r} - \mathbf{r}'_i|
\]  

(15)

The velocity vector \( \mathbf{v}'_i \) of the center of the expansion zone 302 at time \( t \) is computed as

\[
\mathbf{v}'_i(t, t', t) = \mathbf{u}(t', \mathbf{r}') + \mathbf{g} \varphi_i.
\]  

(16)

The velocity \( \mathbf{v} \) of a particle reaching the general non-moving point \( \mathbf{r} \) at time \( t \) on any point of the control surface 302 is computed as

\[
\mathbf{v}(\mathbf{r}, t', t) = \mathbf{v}_i(t', \mathbf{r}') \mathbf{n}_i + \mathbf{u}(t', \mathbf{r}') + \mathbf{g} \varphi_i
\]  

(17)

which is obtained by assigning \( t = t' \) and substitution of \( \varphi_i = t - t'_i \) in Equation (7) given above and rearrangement of the terms.

Step 2: Defining the probability of free path traveling along the ballistic trajectory from the starting point to the ending point.
Step 2 includes:

1) Defining the average magnitude of the velocity at a particular point of a trajectory of the ballistic particle with respect to nearby passed particles at a particular point of a trajectory.

2) Expressing the probability of traveling along the ballistic trajectory in the three-dimensional configuration by Equation (18) given below:

$$Q_i(t,t') = \exp \left( -\int_{t'}^{t} \frac{1}{2} \rho \left( \vec{r}(\tilde{t}) \right) v_{rel} \left( \vec{r}(\tilde{t}) \right) d\tilde{t} \right),$$

where $\tilde{t}$ is a parametric time $t' < \tilde{t} \leq t$, $\vec{r}(\tilde{t})$ is a point on the ballistic trajectory at the parametric time $\tilde{t}$, $v_{rel} \left( \vec{r}(\tilde{t}) \right)$ is the average magnitude of the velocity with respect to a nearby passed particle in the trajectory point $\vec{r}(\tilde{t})$, and $P_i \left( \vec{r}(\tilde{t}) \right)$ is the number of particles placed within a collision tube of a unit length in the trajectory point $\vec{r}(\tilde{t})$.

**Figure 4** shows a perspective view for explaining a method for determining the average magnitude of the instantaneous velocity of the ballistic particle with respect to a nearby passed particle in a three-dimensional configuration. For clarity, the observer’s Cartesian coordinate system 100 is oriented, so $y$-axis is along the negative direction of the applied acceleration field, 401, which provides, for each particle, acceleration $\vec{g}$. Here, at time $\tilde{t}$, particle P1 having a velocity $\vec{v}_1 \left( t', r', \tilde{t} \right)$ and originated from a collision in point A, passes in point B, which is positioned on ballistic trajectory 402, particle P2 having a velocity $\vec{v}_2 \left( \vec{r}(\tilde{t}) \right)$.

Following sub-steps calculate the average magnitude of the velocity $v_{rel} \left( \vec{r}(\tilde{t}) \right)$ of the ballistic particle with respect to nearby passed particles at a particular point of a trajectory B at a specified time $\tilde{t}$:

1) by defining an instant magnitude of the velocity of the converging ballistic particle in the trajectory point with respect to nearby particles in the trajectory point $\vec{r}$ at time $\tilde{t}$ as

$$\vec{v}_{rel} \left( \vec{r}(\tilde{t}) \right) = \vec{v}_1 \left( t', r', \tilde{t} \right) - \vec{v}_2 \left( \vec{r}(\tilde{t}) \right).$$

**Figure 4.** Perspective view for explaining a method of defining the magnitude of the instant velocity of the ballistic particle with respect to a nearby passed particle.
which is formed by connecting the end of the instant velocity vector \( \mathbf{v}_r(t', \mathbf{r}^\prime) \) with any point \( S \) on spherical surface \( 403 \) of radius \( \mathbf{r}_\mathbf{\hat{n}}(t', \mathbf{r}^\prime) \), where

\[
\mathbf{v}_r(t', \mathbf{r}^\prime, \mathbf{i}) = \mathbf{v}_r(t', \mathbf{r}^\prime) \mathbf{\hat{n}}_i + \mathbf{u}(t', \mathbf{r}^\prime) + \mathbf{g}(t' - t'),
\]

(20)

where \( \mathbf{\hat{n}}_i \) is a unit vector having the point of origin \( \mathbf{r}^\prime \), \( \mathbf{v}_r(t', \mathbf{r}^\prime) \) and \( \mathbf{u}(t', \mathbf{r}^\prime) \) are thermal velocity and mass flow velocity components in the rest frame of the model gas in point \( \mathbf{r}^\prime \) at time \( t' \), which are acquired by particle \( P_1 \) because of a collision in this point, and where

\[
\mathbf{v}_r(\mathbf{r}(\mathbf{i})) = \mathbf{v}_r(\mathbf{r}(\mathbf{i})) \mathbf{\hat{n}}_i + \mathbf{u}(\mathbf{r}(\mathbf{i})),
\]

(21)

where \( \mathbf{v}_r(\mathbf{r}(\mathbf{i})) \) and \( \mathbf{u}(\mathbf{r}(\mathbf{i})) \) are thermal velocity and mass flow velocity components in the rest frame of the model gas in point \( \mathbf{r} \) at time \( \mathbf{i} \), which are acquired by particle \( P_2 \) because of a collision in this point, and

2) by averaging the instant magnitude of the velocity overall directions of the thermal velocity component of one of the nearby particles in the trajectory point \( \mathbf{r} \) at time \( \mathbf{i} \). This is done by integrating the instant magnitude of the relative velocity of Equation (19) over the angle of \( \vartheta \) from 0 to \( \pi \) and \( \varphi \), which is the angle of rotation around axis \( OP \) from 0 to \( 2\pi \), and by normalizing by the solid angle of \( 4\pi \), which results in:

\[
v_{\text{av}}(\mathbf{r}(\mathbf{i})) = \frac{1}{2} \int_0^{2\pi} \int_0^\pi \left[ \mathbf{v}_r(\mathbf{r}(\mathbf{i})) \right]^2 + [v_r(\mathbf{r}(\mathbf{i}))]^2 - 2v_r(\mathbf{r}(\mathbf{i})) |\mathbf{v}_r(\mathbf{r}(\mathbf{i}))| \cos(\vartheta) \sin(\vartheta) \cos(\vartheta) \sin(\vartheta) d\vartheta d\varphi,
\]

(22)

where

\[
\mathbf{v}_r(\mathbf{r}(\mathbf{i})) = \mathbf{v}_i(t', \mathbf{r}', \mathbf{i}) - \mathbf{u}(\mathbf{r}(\mathbf{i})).
\]

(23)

Note that, typically, the magnitude of the relative mass flow velocity or the mass flow velocity component of the passing particle \( P_1 \) with respect to nearby passed particle \( P_2 \) is insignificant compared to the magnitude of the thermal velocity of either passing particle \( P_1 \) or nearby passed particle \( P_2 \) or of both.

The average magnitude of the velocity of the traveling particle with respect to a nearby passed particle is calculated from Equation (24) given below, which is obtained by substitution of \( \mathbf{v}_r(\mathbf{r}(\mathbf{i})) = v_r(\mathbf{r}(\mathbf{i})) \) in Equation (22) and executing the integration of the resulted equation:

\[
v_{\text{av}}(\mathbf{r}(\mathbf{i})) = \frac{1}{6v_r(t', \mathbf{r}')} \left[ \left( v_r(t', \mathbf{r}') + v_r(\mathbf{r}(\mathbf{i})) \right)^2 - \left| v_r(t', \mathbf{r}') - v_r(\mathbf{r}(\mathbf{i})) \right| \right],
\]

(24)

where \( v_{\text{av}}(\mathbf{r}(\mathbf{i})) \) is the average magnitude of the velocity with respect to a nearby passed particle in the trajectory point, \( v_r(t', \mathbf{r}’) \) is the magnitude of the thermal velocity in the starting point of the ballistic trajectory, and \( v_r(\mathbf{r}(\mathbf{i})) \) is the thermal velocity of a nearby passed particle in the trajectory point.

Also, typically, the magnitudes of the thermal velocity of nearby particles are approximately identical. For non-relativistic particles, the average magnitude of
the velocity with respect to each particle moving in an arbitrary direction is calculated from Equation (25) given below, which is obtained by substitution of \( v_r \left( \mathbf{\tilde{r}}(t) \right) = v_r \left( t'_r, \mathbf{\tilde{r}}' \right) = v_r \) in Equation (24):

\[
v_{rel} = \frac{4}{3} v_r.
\]  

(25)

Analogously, for relativistic particles, the average magnitude of the velocity with respect to each particle moving in an arbitrary direction is calculated from Equation (25) given below:

\[
v_{rel} = \frac{v_r}{2} \int_0^\infty \frac{2 - 2 \cos(\theta) - \left[ 1 - \cos^2(\theta) \right] v_r^2 / c^2}{1 - v_r^2 \cos(\theta) / c^2} \sin(\theta) \, d\theta
\]  

(26)

where \( c \) is the speed of light. For \( \frac{v_r}{c} \ll 1 \), integrating Equation (26) will yield Equation (25). For \( \frac{v_r}{c} \approx 1 \), integrating Equation (26) will yield:

\[
v_{rel} \approx c.
\]  

(27)

**Step 3: Defining the net rate of particle efflux per unit volume from a point source positioned in a point of the initial collisions and moving with the mass flow velocity of the model gas in that point.**

Step 3 includes the following sub-steps:

1) Defining the particle flux \( \mathbf{J}_n^N \) along the ballistic trajectory in a point of the space \( \mathbf{\tilde{r}} \) surrounding the point of the initial collision, the step that includes representing \( \mathbf{J}_n^N \) by applying Equation (28) as given below:

\[
\mathbf{J}_n^N = \frac{1}{2} n(t'_r, \mathbf{\tilde{r}}') Q_i(t, t'_r) \mathbf{v}(t'_r, \mathbf{\tilde{r}}', t, \mathbf{\tilde{r}}),
\]  

(28)

where \( \mathbf{v} \) is defined by Equation (17), \( Q_i \) is a survival probability defined by Equation (18), and \( n(t'_r, \mathbf{\tilde{r}}') \) is particle density at a specific point \( \mathbf{\tilde{r}}' \) at time \( t'_r \).

2) Representing, in a coordinate system associated with the point of origin \( \mathbf{\tilde{r}}_i \) that moves with velocity \( \mathbf{\tilde{v}}_i \), the vector field of the particle flux \( \mathbf{J}_n^N \) through the in control surface 302 of Figure 3 in the following form:

\[
\mathbf{J}_n^N = \frac{1}{2} n(t'_r, \mathbf{\tilde{r}}') Q_i(t, t'_r) \left[ \mathbf{v}(t'_r, \mathbf{\tilde{r}}', t, \mathbf{\tilde{r}}) - \mathbf{\tilde{v}}_i \right] = \frac{1}{2} n(t'_r, \mathbf{\tilde{r}}') Q_i(t, t'_r) v_r(t'_r, \mathbf{\tilde{r}}') \mathbf{\hat{n}}_i.
\]  

(29)

3) Applying and executing the divergence operator \( \nabla \cdot \) to the vector field of Equation (29) followed by shrinking the volume of the auxiliary control volume to infinitely small volume, i.e., \( \mathbf{\tilde{r}} \to \mathbf{\tilde{r}}' \), which, in formula form, is expressed as:

\[
Z_r(t'_r, \mathbf{\tilde{r}}') = \left[ \nabla \cdot \left( \mathbf{J}_n^N \right) \right]_{\mathbf{\tilde{r}} \to \mathbf{\tilde{r}}'} = \frac{1}{2} \left[ \mathbf{\nabla}' \cdot \left[ n(t'_r, \mathbf{\tilde{r}}') Q_i(t, t'_r) v_r(t'_r, \mathbf{\tilde{r}}') \mathbf{\hat{n}}_i \right] \right]_{\mathbf{\tilde{r}} \to \mathbf{\tilde{r}}'},
\]  

(30)

where

\[
\mathbf{\nabla}' = \mathbf{i} \frac{\partial}{\partial x'} + \mathbf{j} \frac{\partial}{\partial y'} + \mathbf{k} \frac{\partial}{\partial z'},
\]  

(31)
and includes representing the particle flux production rate, or the net rate of particle efflux per unit volume, or the rate of collisions per unit volume \( Z_p(t', \vec{r}') \), in a point of the initial collision moving with the mass flow velocity \( \vec{u}(t', \vec{r}') \) at time \( t' \) by following Equation (32) given below:

\[
Z_p(t', \vec{r}') = \frac{1}{2} n(t', \vec{r}') P_c(t', \vec{r}') v_{rel}(t', \vec{r}'),
\]

(32)

where \( n(t', \vec{r}') \) is particle density, \( P_c(t', \vec{r}') \) is the number of particles placed within a collision tube of a unit length in the corresponding point of the initial collisions at the time of the initial collision, and \( v_{rel}(t', \vec{r}') \) is the average magnitude of the velocity with respect to a nearby passed particle in the corresponding point of initial collisions at the time of the initial collision. The above is obtained upon acknowledgment that the control volume 305, which is confined by inflated control surface 302, is isolated (see Figure 3), and then the total particle efflux through the surface 302 at time \( t \) should be equal to the total particle efflux from the point \( \vec{r}' \) at time \( t' \) through the surface 303 closely surrounding point \( \vec{r}' \). One should recognize that, after a collision in point \( \vec{r}' \) at time \( t' \), the particle may move in a random direction because of the arbitrary nature of the unit vector \( \vec{n} \), thus making an expansion zone around the point source in point \( \vec{r}' \) at time \( t' \), which is shown as 302.

Step 4: Defining property flux in a given non-moving point at a given time from one of the point sources of the model gas.

Step 4 includes representing the property vector flux \( \vec{J}_{p \rightarrow Fs}(t, \vec{r}) \) originated from the point source of the initial collisions in point \( \vec{r}' \) at time \( t' \), which moves in the space of the model gas with a mass-flow velocity \( \vec{u}(t', \vec{r}') \), and being detected by a point sink positioned in a being at the rest point \( \vec{r} \) at the given time \( t \):

\[
\vec{J}_{p \rightarrow Fs}(t, \vec{r}) = \frac{1}{4\pi \left| \vec{r} - \vec{r}' \right|} Q_p(t, t') Z_p(t, \vec{r}') \frac{\vec{u}(t', \vec{r}', t, \vec{r})}{v_{rel}(t', \vec{r}') \Psi_{in}(t', \vec{r}', t, \vec{r})} dV'.
\]

(33)

Step 5: Defining the rate of the property vector flux \( \vec{J}_{FS \rightarrow B}(t, \vec{r}) \) in point \( \vec{r} \) at the given time \( t \), which is originated from initial collisions within entire space occupied by the model gas.

Step 5 includes applying Equation (34) given below, which is obtained by integrating Equation (33) over the volume of the model gas system:

\[
\vec{J}_{FS \rightarrow B}(t, \vec{r}) = \iiint_{V} \frac{1}{4\pi \left| \vec{r} - \vec{r}' \right|} Q_p(t, t') Z_p(t, \vec{r}') \frac{\vec{u}(t', \vec{r}', t, \vec{r})}{v_{rel}(t', \vec{r}') \Psi_{in}(t', \vec{r}', t, \vec{r})} dV'.
\]

(34)

here point \( \vec{r} \) is excluded from integration in the equation above because we are interested in calculating the total rate of the property flux in the point sink at \( \vec{r} \), which is originated from the surrounding point sources of the initial collisions at \( \vec{r}' \).

Step 6: Defining the net rate of property influx per unit volume \( B_{in}^{w,FS} \) formed by the flow of ballistic particles and converging from the gas space in the
general non-moving point at the given time.

Step 6 further includes a step of representing \( B_{m,FS}^{\Psi} \) by applying Equation (35) given below:

\[
B_{m,FS}^{\Psi}(\vec{r},t) = -\nabla \cdot \int_{\mathbb{R}^3} \frac{1}{4\pi}\frac{1}{|\vec{r} - \vec{r}'|} Q_i(t,t') Z_i(t',\vec{r}',\vec{r}) \frac{\vec{v}(t',\vec{r}',\vec{r})}{v_r(t',\vec{r}')} \Psi_m(t',\vec{r}',\vec{r},\rho) dV'.
\] (35)

From the equation above, one can conclude that the net rate of property influx in the general non-moving point \( \vec{r} \) at the given time \( t \) is resulted from the impact of the flow of ballistic particles converging from the gas space. The equation above calculates the impact in point \( \vec{r} \) at the given time \( t \) from all initial collisions of converging ballistic particles having trajectories allowing them to target point \( \vec{r} \) at the given time \( t \), the initial collisions taking place during all preceding dynamic history of the system preceding the given time, i.e. \( t' < t \).

2.4. Defining a Net Rate of Total Property Efflux Per Unit Volume from the General Non-Moving Point at the Given Time

To define the net rate of property efflux per unit volume into surroundings from the general non-moving point \( \vec{r} \) at the given time, \( t \), the linear dimensions of the main control volume surrounding point \( \vec{r} \) are selected to be sufficiently small for preventing two and more consecutive collisions of the same particle within the main control volume. We anticipate that the net rate of property efflux per unit volume is formed by diverging particles. Each of the diverging particles is selected from all available particles by the ballistic trajectory having the starting point in the given non-moving point at the given time. The table of the model parameters associated with defining the net rate of total property efflux per unit volume is presented in Table 2.

**Figure 5** is a schematic shown to illustrate a method for analytical representation of the net rate of property efflux from a non-moving point in three-dimensional space occupied by the model gas. For clarity, the observer’s Cartesian coordinate system 100 is oriented, so the \( y \)-axis is along the negative direction of the applied acceleration field 504. In **Figure 5**, particle P1 in point A, which is indicated by the position vector \( \vec{r} \) (501) at the given time \( t \), diverges from the point source in this point and, in point B, which is indicated by the position vector \( \vec{r}' \) (502) at the time \( t' \), has velocity \( \vec{v} \) while crossing control surface 503 enclosing the point source in point A. We assume that property \( \Psi(t,\vec{r},t',\vec{r}') \) was initially acquired by the particle at the moment of the initial collision in point \( \vec{r} \) at time \( t \). Specifically, in the point source in point \( \vec{r} \) at time \( t \) (point A), each of the diverging particles obtains the thermal velocity of the magnitude of \( v_r \) and mass flow velocity \( \vec{u} \) (not shown).

We define the net rate of property efflux from the general non-moving point \( \vec{r} \) at the given time \( t \) by these four steps:

Step 1: Identifying a trajectory and trajectory characteristics, for each particle diverging from the general non-moving point at the given time.
Figure 5. Perspective view of the model gas system for explaining the ballistic movement of the diverging particle after a collision.

### Table 2. List of the model parameters associated with defining the net rate of total property efflux per unit volume.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Short description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t$</td>
<td>given time</td>
</tr>
<tr>
<td>$\mathbf{r}$</td>
<td>position of the starting point of a ballistic trajectory of the diverging particle</td>
</tr>
<tr>
<td>$\mathbf{r}'$</td>
<td>position of the ending point of a ballistic trajectory of the diverging particle</td>
</tr>
<tr>
<td>$t_\epsilon$</td>
<td>time of positioning the ending point of the diverging particle</td>
</tr>
<tr>
<td>$n(t,\mathbf{r})$</td>
<td>particle density in the starting point $\mathbf{r}$ at the given time $t$</td>
</tr>
<tr>
<td>$\mathbf{u}(\mathbf{r},t)$</td>
<td>mass flow velocity in point $\mathbf{r}$ at time $t$</td>
</tr>
<tr>
<td>$v_t(\mathbf{r},t)$</td>
<td>the average magnitude of the thermal velocity of a diverging particle in point $\mathbf{r}$ at time $t$</td>
</tr>
<tr>
<td>$\mathbf{v}_t(t,\mathbf{r},\mathbf{r}')$</td>
<td>the velocity vector of the diverging particle at the time of positioning in point $\mathbf{r}'$</td>
</tr>
<tr>
<td>$\mathbf{v}_c(t,\mathbf{r},\mathbf{r}')$</td>
<td>the position of a virtual ballistic particle at time $t'$, which has zero magnitude of thermal velocity in the starting point $\mathbf{r}$ at the time of initial collision $t$</td>
</tr>
<tr>
<td>$\mathbf{v}_{cav}(t,\mathbf{r},\mathbf{r}')$</td>
<td>the velocity vector of the virtual ballistic particle at time $t'$</td>
</tr>
<tr>
<td>$Q(t_\epsilon,t)$</td>
<td>the probability of free path traveling along the ballistic trajectory starting at time $t$ and ending at time $t_\epsilon$</td>
</tr>
<tr>
<td>$\Psi(t_\epsilon,t_\epsilon',\mathbf{r}')$</td>
<td>property content carried by the diverging particle at the time $t_\epsilon'$ of crossing in point $\mathbf{r}'$ enclosing control surface 503</td>
</tr>
</tbody>
</table>

Step 1 includes:

1) Formulating position vector $\mathbf{r}'(506)$ of the particle on trajectory 505 and velocity vector $\mathbf{v}'(507)$ (not shown) at time $t'$ by applying Equations (36) and (37), respectively, given below:

$$\mathbf{r}' = \mathbf{r}'(t,\mathbf{r},\mathbf{r}') = v_t(t,\mathbf{r})\mathbf{n}_{ce}(t'-t) + v_{cav}(t,\mathbf{r},\mathbf{r}')$$  \hspace{1cm} (36)

and

$$\mathbf{v}' = \mathbf{v}'(t,\mathbf{r},\mathbf{r}') = v_t(t,\mathbf{r})\mathbf{n}_{ce} + v_{cav}(t,\mathbf{r},\mathbf{r}')$$  \hspace{1cm} (37)

where $\mathbf{n}_{ce}$ is a unit vector with the initial point of origin $\mathbf{r}$ at time $t$ and
In Equations (36) and (37) above, \( \vec{r}_c^* (t, \vec{r}, \vec{t}) \) and \( \vec{v}_c^* (t, \vec{r}, \vec{t}) \) are defined by an appropriate law of motion.

Specifically, when Newton’s Law of Motion governs the model gas system, \( \vec{r}_c^* (t, \vec{r}, \vec{t}) \) and \( \vec{v}_c^* (t, \vec{r}, \vec{t}) \) are defined by Equations (38) and (39), respectively, given below:

\[
\vec{r}_c^* (t, \vec{r}, \vec{t}) = \vec{r} + \vec{u}(t, \vec{r})(\vec{t} - t) + \frac{1}{2} \vec{g}(\vec{t} - t)^2 \tag{38}
\]

and

\[
\vec{v}_c^* (t, \vec{r}, \vec{t}) = \vec{u}(t, \vec{r}) + \vec{g}(\vec{t} - t). \tag{39}
\]

2) Determining the time needed, for each particle diverging from a non-moving point \( \vec{r} \) at the given time \( t \) to cross a control surface enclosing point \( \vec{r} \) in point \( \vec{r}' \) of the control surface. It can be done by solving Equation (36) in which \( \vec{t} = \vec{t}' \). Specifically, when Newton’s Laws of Motion govern the motion of the particles, then it can be done by resolving the equation of projectile motion, Equation (40) given below, with respect to the ballistic traveling time \( \varphi \).

\[
\vec{r}' = \vec{r} + [v_2(t, \vec{r})\vec{n}_{c_\varphi} + \vec{u}(t, \vec{r})]\varphi + \frac{1}{2} \vec{g}\varphi^2 \tag{40}
\]

which is obtained by assigning \( \vec{t} = \vec{t}' \) and substitution of \( \varphi = \vec{t}' - \vec{t} \) in Equation (36) and rearrangement of the terms. Traveling time \( \varphi \) is needed to indicate a time \( \vec{t}' \) at which the departing particle has reached a point \( \vec{r}' \) on control surface 503 enclosing point \( \vec{r} \).

3) Defining an instant unit vector directing thermal velocity component along the diverging ballistic trajectory includes:

presenting Equation (40) in the following form:

\[
\varphi, v_2\vec{n}_{c_\varphi} = \vec{r}' - \vec{r}_c^*, \tag{41}
\]

where \( \vec{n}_{c_\varphi} \) is a unit vector and \( \vec{r}_c^* \) (not shown) is the location of a virtual ballistic particle leaving point \( \vec{r} \) at time \( t \), which would have zero magnitude of the thermal velocity and is observed at time \( \vec{t}' \). In a case, where a model gas system is governed by Newton’s Laws of Motion, \( \vec{r}_c^* \) is expressed as:

\[
\vec{r}_c^* = \vec{r} + \vec{u}(t, \vec{r})\varphi + \frac{1}{2} \vec{g}\varphi^2 \tag{42}
\]

and deriving \( \vec{n}_{c_\varphi} \) from Equation (41), which is given as

\[
\vec{n}_{c_\varphi} = \frac{\vec{r}' - \vec{r}_c^*}{|\vec{r}' - \vec{r}_c^*|}, \tag{43}
\]

where \( \vec{n}_{c_\varphi} \) is a unit vector of origin \( \vec{r}_c^* \) at time \( \vec{t}' \), directing the vector of the thermal velocity component and where \( \vec{t}' \) is the time of positioning in point \( \vec{r}' \), which is on the control surface 503.

4) Defining the velocity vector \( \vec{v}_c^* \) (507) of each particle at the moment of crossing the control surface in point \( \vec{r}' \) includes, where a model gas system is
governed by Newton’s Laws of Motion, representing \( \vec{v}_s \) by applying Equation (44) as given below:

\[
\vec{v}_s(t, \vec{r}, t'_s, \vec{r}') = v_y(t, \vec{r})\vec{n}_s + \vec{u}(t, \vec{r}) + \vec{g}\varphi_s,
\]

which is obtained by assigning \( t'_s = t \) and substitution of \( t'_s - t = \varphi_s \) in Equation (37).

**Step 2: Defining the probability of traveling along the ballistic trajectory from the general non-moving point \( \vec{r} \) at time \( t \) to one of the points in space surrounding the general non-moving point.**

Step 2 includes representing the probability \( Q_s(t'_s, t) \) by applying Equation (45) given below:

\[
Q_s(t'_s, t) = Q_s(0, \varphi_s) = \exp\left(-\int_{t}^{t'} P_s\left(\vec{r}'(\tilde{t})\right)\vec{v}_{rel}(\vec{r}'(\tilde{t}))d\tilde{r}\right),
\]

where \( t'_s = t + \varphi_s \) is the time of the particle positioning in point \( \vec{r}' \), \( P_s\left(\vec{r}'(\tilde{t})\right) \) is the number of particles placed within a collision tube of a unit length in the trajectory point \( \vec{r}'(\tilde{t}) \), \( \vec{v}_{rel}(\vec{r}'(\tilde{t})) \) is an average magnitude of the velocity with respect to a nearby passed particle in the trajectory point \( \vec{r}'(\tilde{t}) \).

**Step 3: Defining the vector/tensor field of property flux \( J^\Psi \) along with one of the ballistic trajectories of a diverging particle in point \( \vec{r}' \).**

Step 3 includes representing the property vector flux \( J^\Psi \) by applying Equation (46) as given below:

\[
J^\Psi = \frac{1}{2} n(t, \vec{r})Q_s(t'_s, t)\vec{v}_s(t, \vec{r}, t'_s, \vec{r}')\Psi(t, \vec{r}, t'_s, \vec{r}').
\]

**Step 4: Defining the net rate of property efflux per unit volume \( B^\Psi_{out} \) from the general non-moving point \( \vec{r} \) at the given time \( t \).**

Step 4 includes representing \( B^\Psi_{out} \) by applying Equation (47) as given below:

\[
B^\Psi_{out}(\vec{r}, t) = \frac{1}{2}\left[\nabla \left[n(t, \vec{r})Q_s(t'_s, t)\vec{v}_s(t, \vec{r}, t'_s, \vec{r}')\Psi(t, \vec{r}, t'_s, \vec{r}')\right]\right]_{\vec{r} \rightarrow \vec{r}'},
\]

which is obtained by executing the divergence operator \( \nabla \cdot \) to the vector field of Equation (46) and by shrinking the volume of control volume 508 confined by the control surface 503 to an infinitely small volume, \( \vec{r}' \rightarrow \vec{r} \), which also leads to the limit \( \vec{r}' \rightarrow \vec{r} \).

### 2.5. Analytical Representation of a General Integro-Differential form of Property Balance Equation in the Three-Dimensional Model Gas System

The integro-differential form of property balance equation is formulated by Equation (48) given below, which is obtained by substitution of Equations (35) and (47) in Equation (5):

\[
\frac{\partial}{\partial t}\left[\frac{1}{2}n(t, \vec{r})\Psi(t, \vec{r})\right] + \frac{1}{2}\left[\nabla \left[\frac{1}{4\pi}Q_s(t'_s, t)\vec{v}_s(t, \vec{r}, t'_s, \vec{r}')\Psi(t, \vec{r}, t'_s, \vec{r}')]\right]\right]_{\vec{r} \rightarrow \vec{r}'} = -\nabla \cdot \left[\frac{1}{4\pi}Q_s(t'_s, t)Z_s(t'_s, \vec{r}')\frac{\vec{v}_s(t'_s, \vec{r}', \vec{r})}{\sqrt{\varphi_s(t'_s, \vec{r}')}}\Psi_m(t'_s, \vec{r}', t, \vec{r})dV',
\]

DOI: 10.4236/jamp.2020.86084 1099
here point \( \vec{r} \) is not included in integration for converging ballistic particles. It implies that any singularity in the right-hand of the equation above is excluded, so Equation (48) defines \( \Psi \) at \( t \geq 0 \) as an implicit function of \( \vec{r} \) on \( \mathbb{R}^3 \), i.e. \( \vec{r} \in \mathbb{R}^3 \).

The analytical representation of the general integro-differential form of property balance equation shown by Equation (48) is generally valid for any homogeneous fluid system with any configuration of the external field of force.

Remark that the integro-differential property balance equation needs to be formed for each unknown property/variable so that the number of equations in a system of balance equations is sufficient to determine each of the unknown properties characterizing the model gas flow. In the following, we provide general governing integro-differential forms of mass balance, momentum balance, and energy balance equations.

In continuation of the discussion at the end of Section 2.2, we would like to highlight that the equation above is a general symbolic representation of the rule that shall be obeyed at any given time \( t \) in any given point \( \vec{r} \) of space occupied by presumably chaotically moving particles experiencing random collisions. Equation (48) also shows that the balance at time \( t \) in any given point \( \vec{r} \) is formed by the exhaustive combination of converging ballistic particles from the surrounding, which can target with a certain probability the given point in space \( \vec{r} \) at the given time \( t \). The converging in point \( \vec{r} \) at time \( t \) ballistic particles are originated from preceding collisions at times \( t' < t \) (past). Whereas the diverging ballistic particles originated from collisions at time \( t \) (present) transport the balanced property/information into surrounding toward the future \( t'_i > t \).

### 2.5.1. Analytical Representation of a General Integro-Differential Form of Mass Balance Equation in the Three-Dimensional Model Gas System

To formulate a general integro-differential form of mass balance equation in a given non-moving point of space occupied by the model at a given time, we will modify Equation (48) by assigning:

\[
\Psi = \Psi_{in} = 1. \tag{49}
\]

Then, we obtain the following general integro-differential form of the mass balance equation

\[
\frac{\partial}{\partial t} \left[ n(t, \vec{r}) \right] + \frac{1}{2} \left[ \nabla \cdot \left[ n(t, \vec{r}) Q_s(t', t) \vec{v}_s(t, \vec{r}, t', \vec{r'}) \right] \right]_{t,t'} \rightarrow \vec{r}.
\]

\[
= -\nabla \cdot \left[ \int_{V'} \frac{1}{4\pi |\vec{r} - \vec{r}'|} Q_s(t', t) Z_{tt'}(t', \vec{r}) \vec{v}(t', \vec{r}, \vec{r'}) dV'. \tag{50}\right.
\]

### 2.5.2. Analytical Representation of a General Integro-Differential Form of Momentum Balance Equation in the Three-Dimensional Model Gas System

To formulate a general integro-differential form of momentum balance equation in a given non-moving point of space occupied by the model at a given time, we will modify Equation (48) by assigning:
in the left-hand of the equation and
\[ \Psi_{in}(t', \vec{r}', t, \vec{r}) = \vec{v}(t', \vec{r}', t, \vec{r}) \] (52)
in the right-hand of the equation. Then we obtain:
\[ \left\{ \frac{m}{c(t)\rho} \left[ n(t, \vec{r}) \vec{v}_i(t, \vec{r}, t_0', \vec{r}') \right] \right\}_{\vec{r} \to \vec{r}'} + \frac{1}{2} \left\{ \nabla \cdot \left[ mn(t, \vec{r}) Q_l(t', t) \vec{v}_i(t, \vec{r}, t_0', \vec{r}') \vec{v}_i(t, \vec{r}, t_0', \vec{r}') \right] \right\}_{\vec{r} \to \vec{r}'} + \nabla p(t, \vec{r}) \] (53)
\[ = -m \nabla \cdot \int \frac{1}{4\pi |r - r'|} Q_l(t, t') Z_v(t', \vec{r}') \frac{\vec{v}(t', t', t, \vec{r}) \vec{v}_i(t, \vec{r}, t_0', \vec{r}')}{v_{t'}} dV' \] (56)

In the equation above, we also introduce a term of the pressure force exerted on its surroundings in point \( \vec{r} \) at time \( t \), which may appear because of a non-uniform pressure applied to a bounded system.

2.5.3. Analytical Representation of a General Integro-Differential Form of Energy Balance Equation in the Three-Dimensional Model Gas System
To formulate a general integro-differential form of energy balance equation in a given non-moving point of space occupied by the model at a given time, we will modify Equation (48) by assigning:
\[ \Psi(t, \vec{r}, t_0', \vec{r}') = \frac{1}{2} \vec{v}_i(t, \vec{r}, t_0', \vec{r}') \cdot \vec{v}_i(t, \vec{r}, t_0', \vec{r}') \] (54)
in the left-hand of the equation and
\[ \Psi_{in}(t', \vec{r}', t, \vec{r}) = \frac{1}{2} \vec{v}(t', \vec{r}', t, \vec{r}) \cdot \vec{v}(t', \vec{r}', t, \vec{r}) \] (55)
in the right-hand of the equation. Then we obtain:
\[ \left\{ \frac{\partial}{\partial t} \left[ n(t, \vec{r}) \vec{v}_i \cdot \vec{v}_i \right] \right\}_{\vec{r} \to \vec{r}'} + \frac{1}{2} \left\{ \nabla \cdot \left[ n(t, \vec{r}) Q_l(t', t) \vec{v}_i(t, \vec{r}, t_0', \vec{r}') \vec{v}_i(t, \vec{r}, t_0', \vec{r}') \right] \right\}_{\vec{r} \to \vec{r}'} + \frac{1}{m} \rho(t, \vec{r}) \vec{v}_i \] \[ = -\nabla \cdot \int \frac{1}{4\pi |r - r'|} Q_l(t, t') Z_v(t', \vec{r}') \frac{\vec{v}(t', \vec{r}', t, \vec{r}) \vec{v}_i(t, \vec{r}, t_0', \vec{r}')}{v_{t'}} dV'. \] (56)

In the second left-term of the equation above, we also introduced a term of the pressure force work on its surroundings in point \( \vec{r} \) at time \( t \), which may appear because of a non-uniform pressure applied to the system.

3. Integro-Differential Forms of Mass and Momentum Balance Equations in Collision-Dominated Flow Regime
We have recognized that, in the majority of real conditions on Earth and near-Earth space, first, the acceleration field \( \vec{g} \) applied to each of the model gas particles is approximately a constant, namely:
\[ \vec{g} = \text{constant} ; \] (57)
second, for each of the model gas particles, the average magnitude of the thermal
velocity component, \( v_r \), is much higher than the magnitude of a mass-flow velocity component \( |\mathbf{v}| \), namely:

\[
|\mathbf{v}| \ll v_r; \tag{58}
\]

third, in the collision-dominated flow regime, the magnitude of velocity gained or lost because of interaction with the acceleration field during the ballistic traveling time \( \varphi_{r0} \) is insignificant in comparison with the thermal velocity component \( v_r \), namely:

\[
|\mathbf{\Phi}_{r0}| \ll v_r; \tag{59}
\]

fourth, in the collision-dominated flow regime, the relative change of any property value or any parameter characterizing the model gas \( \Sigma \) is insignificant during the period between collisions \( \frac{1}{P_c\varphi_{r0}} \) in a given point of the model gas, which is expressed:

\[
\frac{1}{P_c\varphi_{r0}} \left| \frac{d}{dr} \frac{\Sigma}{|\mathbf{\Sigma}|} \right| \ll 1
\]

(60)

and fifth, in the collision-dominated flow regime, the relative change of any property value or any parameter characterizing the model gas is insignificant on the length scale of the average distance between the model gas particles \( \frac{1}{P_c} \), which is expressed:

\[
\frac{1}{P_c} \left| \nabla \Sigma \right| \ll 1
\]

(61)

We also note here that since point \( \mathbf{r} \) is excluded from integration in the domain of integration \( \Omega \) in which \( \mathbf{r}' \neq \mathbf{r} \), the operation of differentiation regarding a parameter \( \mathbf{r} \) is interchangeable with the operation of integration over some other variable \( \mathbf{r}' \). Also, these functions or approximations for functions involving in calculations according to Equations (50), (53), and (56) are formulated:

1) Approximating \( \mathbf{v}_s(t, t'_s, t''; \mathbf{r}'') \) as

\[
\mathbf{v}_s(t, t'_s, t''; \mathbf{r}'') \approx \mathbf{v}_s(t, t'_s, t'_{s0}; \mathbf{r}'') = v_y(t, \mathbf{r}) \mathbf{n}_{s0} + \mathbf{u}(t, \mathbf{r}) + \mathbf{\Phi}_{s0},
\]

(62)

where \( \mathbf{n}_{s0} \) is the unit vector of arbitrary direction from the point \( \mathbf{r} \), which is approximated from \( \mathbf{n}_s \) as

\[
\mathbf{n}_s \approx \mathbf{n}_{s0} = \frac{\mathbf{r}' - \mathbf{r}}{\left| \mathbf{r}' - \mathbf{r} \right|}
\]

(63)

and \( \varphi_{s0} \) is the ballistic traveling time, which is approximated from \( \varphi_s \) as

\[
\varphi_s \approx \varphi_{s0} = \frac{\left| \mathbf{r}' - \mathbf{r} \right|}{v_y(\mathbf{r}, t)}.
\]

(64)

2) Approximating \( Q_s(t_s, t) \) as

\[
Q_s(t_s, t) \approx Q_{s0}(t_{s0}, t) = \exp \left( -\int_{t_{s0}}^{t} P_c \left( \mathbf{r}(\tilde{t}) \right) v_{r0} \left( \mathbf{r}(\tilde{t}) \right) d\tilde{t} \right),
\]

(65)
where \( t_{\phi_0}' \) is the time of the particle positioning in point \( \vec{r}' \), which is approximated from \( t' \) as

\[
t' \equiv t_{\phi_0}' = t + \phi_{\phi_0}
\]

(66)

3) For \( \vec{r}' \rightarrow \vec{r} \), executing \( \nabla Q_{\phi_0}\left(t_{\phi_0}', t\right) \) as

4) Approximating \( \vec{v}\left(t_{\phi_0}', \vec{r}', t, \vec{r}\right) \) as

\[
\vec{v}\left(t_{\phi_0}', \vec{r}', t, \vec{r}\right) \equiv \vec{v}\left(t_{\phi_0}', \vec{r}', t, \vec{r}\right) = v_r\left(t_{\phi_0}', \vec{r}'\right)\vec{n}_{\phi_0} + \vec{u}\left(t_{\phi_0}', \vec{r}'\right) + \vec{g}\phi_{\phi_0},
\]

(67)

where \( \vec{n}_{\phi_0} \) is the unit vector of arbitrary direction from the point \( \vec{r}' \), which is approximated from \( \vec{n}_r \) as

\[
\vec{n}_{r} \equiv \vec{n}_{\phi_0} = \frac{\vec{r} - \vec{r}'}{|\vec{r} - \vec{r}'|}
\]

(68)

and \( \phi_{\phi_0} \) is the ballistic traveling time, which is approximated from \( \phi_r \) as

\[
\phi_r \equiv \phi_{\phi_0} = \frac{|\vec{r} - \vec{r}'|}{v_r\left(t_{\phi_0}', \vec{r}'\right)}.
\]

(69)

5) Approximating \( Q\left(t, t_{\phi_0}'\right) \) as

\[
Q\left(t, t_{\phi_0}'\right) \equiv Q_{\phi_0}(t, t_{\phi_0}' = \exp\left(-\int_{t_{\phi_0}'}^{t} P_r\left(\vec{r}(\tilde{t})\right)v_{rel}\left(\vec{r}(\tilde{t})\right)d\tilde{t}\right),
\]

(70)

where \( t_{\phi_0}' \) is the time of the particle positioning in point \( \vec{r}' \), which is approximated from \( t' \) as

\[
t' \equiv t_{\phi_0}' = t - \phi_{\phi_0}.
\]

(71)

6) Approximating \( \nabla Q_{\phi_0}\left(t, t_{\phi_0}'\right) \) as

\[
\nabla Q_{\phi_0}\left(t, t_{\phi_0}'\right) \equiv -Q_{\phi_0}(t, t_{\phi_0}')P_r(t, \vec{r})v_{rel}(t, \vec{r})\frac{\vec{n}_{\phi_0}}{v_r\left(t_{\phi_0}', \vec{r}'\right)}.
\]

(72)

7) Approximating \( \vec{r}_{t_{\phi_0}'} \) as

\[
\vec{r}_{t_{\phi_0}'} \equiv \vec{r}.
\]

(73)

8) Representing \( G\left(\vec{r}, \vec{r}'\right) \), a first vector derivative of the Green function with no boundary conditions, as

\[
G\left(\vec{r}, \vec{r}'\right) = \frac{1}{4\pi |\vec{r} - \vec{r}'|},
\]

(74)

which has the following property:

\[
\nabla \cdot G\left(\vec{r}, \vec{r}'\right) = \delta\left(\vec{r} - \vec{r}'\right).
\]

(75)

Considering that governing equations require satisfaction of the balance of any model gas property at any time, these analytical representations of approximations for mass balance, momentum balance, and energy balance are provided.

3.1. Analytical Representation of an Integro-Differential Form of the Mass Balance Equation in the Collision-Dominated Flow Regime

Here the mass balance equation is formed by considering a unique combination
of ballistic particles converging from the entire model gas system in a given point at a given time and the diverging ballistic particles from the given point at the given time. Besides, each of the converging ballistic particles can target point \( \mathbf{r} \) at given time \( t \) and originates from a preceding collision within the model gas system at a time earlier than time \( t \). Such a combination of converging and diverging ballistic particles capable of targeting or escaping point \( \mathbf{r} \) at given time \( t \) is treated as an exhaustive combination. Therefore, from a physical viewpoint, the solution \( \mathbf{u}(t, \mathbf{r}) \) is unique.

To formulate an integro-differential form of mass balance equation in the collision-dominated flow regime, which applies to a given non-moving point of space occupied by the model at a given time, we will modify Equation (50) by following steps of:

1) substituting approximations from Equation (62) to Equation (73) in Equation (50) and executing vector differentiation;
2) executing limit \( \mathbf{r}' \to \mathbf{r} \) in the second left-hand term of the resulting integro-differential mass balance equation;
3) remembering that point \( \mathbf{r} \) is not included in integration for converging ballistic particle; and
4) neglecting terms containing \( \frac{1}{\rho P_{rel}} \).

Upon executing the above, Equation (50) is reduced to:

\[
\frac{\partial}{\partial t} n(t, \mathbf{r}) + Z_{\mathbf{r}}(t, \mathbf{r}) + \frac{1}{2} \mathbf{V} \cdot \left[ n(t, \mathbf{r}) \mathbf{u}(t, \mathbf{r}) \right] \\
\approx P_{rel}(t, \mathbf{r})v_{rel}(t, \mathbf{r}) \int \int \int Z_{\mathbf{r}}(t', \mathbf{r}') \frac{1}{v_{\mathbf{r}'}(t', \mathbf{r}')} Q_{00}(t, t') \frac{1}{4\pi |\mathbf{r} - \mathbf{r}'|} dV' \\
+ P_{\mathbf{r}}(t, \mathbf{r})v_{rel}(t, \mathbf{r}) \int \int \int Z_{\mathbf{r}}(t', \mathbf{r}') \frac{1}{v_{\mathbf{r}'}(t', \mathbf{r}')} Q_{00}(t, t') G(\mathbf{r}, \mathbf{r}') \cdot \mathbf{u}(t', \mathbf{r}') dV' \\
+ P_{\mathbf{r}}(t, \mathbf{r})v_{rel}(t, \mathbf{r}) \int \int \int Z_{\mathbf{r}}(t', \mathbf{r}') \frac{1}{v_{\mathbf{r}'}(t', \mathbf{r}')} Q_{00}(t, t') G(\mathbf{r}, \mathbf{r}') \cdot \mathbf{g}(t', \mathbf{r}') dV',
\]

where \( Z_{\mathbf{r}}(t, \mathbf{r}) \) the rate of collisions per unit volume in the given non-moving point \( \mathbf{r} \) at time \( t \), which is expressed as

\[
Z_{\mathbf{r}}(t, \mathbf{r}) = \frac{1}{2} n(t, \mathbf{r}) P_{\mathbf{r}}(t, \mathbf{r}) v_{rel}(t, \mathbf{r}).
\]

Equation (76) defines function \( \mathbf{u} \) as a function of \( \mathbf{r} \) on \( \mathbb{R}^3 \), i.e. \( \mathbf{r} \in \mathbb{R}^3 \), at \( t \geq 0 \).

3.2. Analytical Representation of an Integro-Differential Form of the Momentum Balance Equation in the Collision-Dominated Flow Regime

Here we again should note that the momentum balance equation is formed by considering a unique combination of ballistic particles converging from the entire model gas system in a given point at a given time and the diverging ballistic particles from the given point at the given time. Besides, each of the converging
ballistic particles can target point \( \vec{r} \) at given time \( t \) and originates from a preceding collision within the model gas system at a time earlier than time \( t \). Such a combination of converging and diverging ballistic particles capable of targeting or escaping point \( \vec{r} \) at given time \( t \) is treated as an exhaustive combination. Therefore, from a physical viewpoint, the solution \( \vec{u}(t, \vec{r}) \) is unique.

To formulate an integro-differential form of momentum balance equation in a given non-moving point of space occupied by the model at a given time, we will modify Equation (53) as follows:

1) assigning
\[
\vec{v}_{\text{r}}(t, \vec{r}, t', \vec{r}') \equiv v_{\text{r}}(t, \vec{r}) \vec{n}_{\text{r}, 0} + \vec{u}(t, \vec{r}) + \vec{g} \varphi_{0, 0} \quad (78)
\]

and
\[
\vec{v}(t', \vec{r}', t, \vec{r}) \equiv v_{\text{r}}(t', \vec{r}') \vec{n}_{\text{r}', 0} + \vec{u}(t', \vec{r}') + \vec{g} \varphi_{0, 0}; \quad (79)
\]

2) substituting approximations for functions involving in Equation (48) by equations from Equation (62) to Equation (73) and executing vector differentiation;

3) executing limit \( \vec{r}' \rightarrow \vec{r} \) in the second left-hand term of the resulting integro-differential momentum balance equation;

4) remembering that point \( \vec{r} \) is not included in integration for converging ballistic particle; and

5) neglecting terms containing \( \frac{1}{P_{\text{r}} v_{\text{rel}}} \).

Upon executing the above, Equation (56) is reduced to:
\[
\frac{\partial}{\partial t} \nabla \overrightarrow{u} + 2Z_{\text{r}} \overrightarrow{u} + \frac{1}{2} \nabla \left( m \frac{\partial \overrightarrow{u}}{\partial t} \right) + \frac{1}{2} \nabla \left[ n \overrightarrow{u} \right] - 2n \vec{g} + \frac{1}{m} \nabla p
\]
\[
\equiv P_{\text{r}} v_{\text{rel}} \int \int \int_{V_{\text{r}}} Z_{\text{r}} Q_{\text{r}, 0} \overrightarrow{G} \overrightarrow{dV} + 2P_{\text{r}} v_{\text{rel}} \int \int \int_{V_{\text{r}}} Z_{\text{r}} \frac{1}{v_{\text{r}}} Q_{\text{r}, 0} \frac{1}{4\pi \left| \vec{r} - \vec{r}' \right|} \overrightarrow{u} \overrightarrow{dV}'
\]
\[
+ 2P_{\text{r}} v_{\text{rel}} \int \int \int_{V_{\text{r}}} Z_{\text{r}} \frac{1}{v_{\text{r}}} Q_{\text{r}, 0} \frac{1}{4\pi \left| \vec{r} - \vec{r}' \right|} \varphi_{0, 0} \overrightarrow{g} \overrightarrow{dV}'
\]
\[
+ P_{\text{r}} v_{\text{rel}} \int \int \int_{V_{\text{r}}} Z_{\text{r}} \frac{1}{v_{\text{r}}} Q_{\text{r}, 0} \overrightarrow{G} \cdot \overrightarrow{u} \overrightarrow{dV}'
\]
\[
+ P_{\text{r}} v_{\text{rel}} \int \int \int_{V_{\text{r}}} Z_{\text{r}} \frac{1}{v_{\text{r}}} Q_{\text{r}, 0} \left[ \varphi_{0, 0} \left( \overrightarrow{u} \overrightarrow{g} + \overrightarrow{g} \overrightarrow{u} \right) + \overrightarrow{g} \varphi_{0, 0} \right] \overrightarrow{dV}'. \quad (80)
\]

It implies that Equation (80) defines \( \overrightarrow{u} \) at \( t \geq 0 \) as an implicit function of \( \vec{r} \) on \( \mathbb{R}^3 \), i.e. \( \vec{r} \in \mathbb{R}^3 \).

4. Validation Tests

The following validation tests are aimed to demonstrate the feasibility of the proposed method.

4.1. Determining the Total Rate of Collisions Per Unit Area on a Surface Being in Contact with the Gas

Figure 6 is a perspective view for explaining a method for determining the total
rate of collisions per unit area on a surface being in contact with the gas. In a system with no gravitational force, the ballistic particles have straight-line trajectories. Here we limit further our consideration to an incompressible model gas at the uniform temperature in a steady-state condition. In the semi-sphere filled with the model gas over the being at rest surface \( A_s \) having a directional vector \( \vec{n} \), each particle having an instantaneous randomly directed vector of the thermal velocity of magnitude, \( v_r \) may have the instant vector-velocity component directing a particle toward the surface \( A_s \). In Figure 6 the particle 601 positioned at \( \vec{r}' \) is shown at distance \( y \) from the surface \( A_s \) in point \( \vec{r} \). The angle between the instant vector-velocity \( \vec{v} \) and the directional vector \( \vec{n} \) is labeled as \( \theta \). The ends of the directionally random vector-velocity of magnitude \( v_r \) form spherical surface 602.

Adopting Equation (34) to the conditions above and assigning \( \Psi_{in} = 1 \), the total rate of collisions per unit area on the surface \( A_s \) is given as:

\[
Z = -Z_y \frac{1}{4\pi} \int \int \int \frac{1}{|\vec{r} - \vec{r}'|} Q_{02}(\vec{r}, \vec{r} ') \vec{n}_{in} \cdot \vec{n} dV',
\]

where

\[
\vec{n}_{in} = \frac{\vec{r} - \vec{r}'}{|\vec{r} - \vec{r}'|},
\]

\[
Q_{02}(\vec{r}, \vec{r}') = \exp\left(-\frac{4}{3} P_v |\vec{r} - \vec{r}'|\right),
\]

and \( Z_v \) is the rate of collisions per unit volume, which, upon substitution of Equation (25) in Equation (77) is expressed as

\[
Z_v = \frac{2}{3} n P_v v_r,
\]

where \( n \) is particle density, \( v_r \) is the magnitude of the thermal velocity, and \( P_v \) is the number of particles placed within a collision tube of a unit length.

Finally, using the geometry illustrated in Figure 6, considering that

\[
|\vec{r} - \vec{r}'| = r' = y/\cos(\theta)
\]

and

Figure 6. The perspective view of the geometry and coordinate system for determining the total rate of collisions per unit area and the pressure exerted on the surface from the gas volume.
\[ \vec{n}_{io} \cdot \vec{n} = -\cos(\theta) \]  \hspace{1cm} (86)

and substituting Equations (82), (83), (84), (85), and (86) in Equation (81), in which \( dV' = \rho^2 \sin(\theta) \, d\theta d\phi d\rho' \), we obtain:

\[
Z = \frac{2}{3} nT v_T \frac{1}{4\pi} \int_0^{2\pi} \int_0^\pi \exp \left( -\frac{4}{3} \frac{P_{zy}}{\cos(\theta)} \right) \sin(\theta) \, d\theta d\phi dy = n \frac{v_T}{4}. \hspace{1cm} (87)
\]

That the result of derivation above according to our method applied to the model gas is identical to the result of the derivation of the rate of collisions per unit area of an ideal gas, which one can find in any course of the kinetic theory of gases, supports the Ballistic Model.

**4.2. Determining the Pressure Exerted on the Surface from the Entire Gas Volume Being in Contact with the Surface**

Referring to the previous section and Figure 6 we recognize that each particle of the model gas carries the momentum

\[ \vec{p} = m v_T \vec{n}_{io} \]  \hspace{1cm} (88)

and the momentum component delivered by the particle to the surface \( A_s \) is

\[ \Delta \rho_{v} = m v_T \vec{n}_{io} \cdot \vec{n}, \]  \hspace{1cm} (89)

where \( m \) is mass of the particle/molecule.

If the particle undergoes an elastic collision with the surface, in such a collision, the momentum passed on the surface is steady-state. Adopting Equation (34) to the conditions listed initially referring to Figure 7 and assigning \( \Psi_{in} = 2\Delta \rho_{v} \), the pressure \( P \) exerted on the surface \( A_s \) is given as:

\[
P = Z \frac{2}{4\pi} \frac{mv_T}{nT v_T} \int_0^{2\pi} \int_0^\pi \int_0^\infty Q_{io}(\vec{r}, \vec{r}') (\vec{n}_{io} \cdot \vec{n}) (\vec{n}_{io} \cdot \vec{n}) \, dV'. \hspace{1cm} (90)
\]

Now, using the geometry illustrated in Figure 6 and substituting Equations (82), (83), (84), (85), and (86) in Equation (90), in which \( dV' = \rho^2 \sin(\theta) \, d\theta d\phi d\rho' \), we obtain:

\[
P = \frac{2}{3} nT v_T \frac{1}{\pi} \int_0^{2\pi} \int_0^\pi \int_0^\infty \exp \left( -\frac{4}{3} \frac{P_{zy}}{\cos(\theta)} \right) \cos^2(\theta) \sin(\theta) \, d\theta d\phi dy = \frac{mnv_T^2}{3}. \hspace{1cm} (91)
\]

Again, that the result of derivation above according to our method, which considers an impact on the surface of unlimited number particles of the model gas (integration along \( y \) direction from zero to infinity), is identical to the result of the derivation of the pressure of an ideal gas according to the kinetic theory of gases, corroborates with the Ballistic Model.

Besides, analyzing validation tests 4.1 and 4.2 above, one may note that the proposed method provides the possibility of quantifying impacts of the limited number of particles on the gas-solid interface. This, from the practical viewpoint, is important in many applications dealing with MEMS technology (capacitive sensing, electrostatic actuation mechanisms, a block of sensing mass in micro gyroscopes, accelerometers, switches, mirrors, pressure sensors and so on) [15].

---

DOI: 10.4236/jamp.2020.86084

1107 Journal of Applied Mathematics and Physics
4.3. Direct Validation of the BPPBS in the Three-Dimensional Unlimited Incompressible Gas Space at the Uniform Temperature

Recognizing that in a steady-state model incompressible gas flow at the uniform temperature with no gravitational force, all variables describing flow depend on the position in space and are not dependent on time, so that the particle density \( n \), the mass \( m \), and the magnitude of the thermal velocity of the particle/molecule, \( v_t \), are constant. In the model gas system being at rest and characterized by the above conditions, when the applied to the model gas system external pressure is uniform, the mass-flow velocity, \( \bar{u} \), is expected to be zero. Now we may formulate the mass-balance equation by adopting Equations (76) as

\[
mZ_v = P_v v_m Z_r \frac{1}{v_r} m \frac{1}{4\pi} \int_0^\infty \int_0^{2\pi} \int_0^1 Q_{00}(\bar{r}, \bar{r}^\prime) \frac{1}{\bar{r} - \bar{r}^\prime} dV^\prime
\]  

(92)

and the momentum-balance equation by adopting Equations (80) as

\[
0 = P_v v_m Z_r m \int_0^\infty \int_0^{2\pi} \int_0^1 Q_{00}(\bar{r}, \bar{r}^\prime) G(\bar{r}, \bar{r}^\prime) dV^\prime.
\]  

(93)

We need to verify these two equalities above by integration of the right-hands of Equations (92) and (93).

Placing the basis of the coordinate system in point \( \bar{r} \) and transforming the Cartesian coordinate system into the spherical coordinate system, then, after defining

\[
\rho^\prime = |\bar{r} - \bar{r}^\prime|,
\]  

(94)

and substituting Equations (83) and (94) in Equation (92), in which \( dV^\prime = \rho^\prime \sin(\theta) d\theta d\phi d\rho^\prime \), we calculate the right-hand of Equations (92) as follows:

\[
P_v \frac{4}{3} Z_r m \frac{1}{4\pi} \int_0^\infty \int_0^{\pi} \int_0^{2\pi} \exp\left(-\frac{4}{3} P_v \rho^\prime \right) \sin(\theta) d\theta d\phi d\rho^\prime
\]  

\[
= P_v \frac{4}{3} Z_r m \int_0^{\pi} \exp\left(-\frac{4}{3} P_v \rho^\prime\right) d\rho^\prime = mZ_v.
\]  

(95)
Analogously, after substitution of Equations (74), (83), and (94) in Equation (93) and taking into account that point $\vec{\rho}' = 0$ is not included into integration, we calculate the right-hand of Equations (93) as follows:

$$
P_c \frac{4}{3} V_Z \int_0^\infty \int_0^\pi \int_0^{2\pi} \exp\left(-\frac{4}{3} P_c \rho'\right) \sin(\theta) d\theta d\rho \frac{\vec{\rho}'}{\rho'} d\rho' = 0.
$$

(96)

Here $P_c$ is the number of particles placed within a collision tube of a unit length, $Z_v$ is the rate of collisions per unit volume, which is defined by Equation (84), and $\vec{\rho}'/\rho'$ is a unit vector with the point of origin at $\vec{\rho}' = 0$.

That the results of integration provided by Equations (95) and (96) are identical to the left-hands of Equations (92) and (93), respectively, evidently supports the validity of the BPPBS in three-dimensional infinite space.

Note: A short communication about these two validation tests below is present on the website and can be accessed on [16]. This communication is not published nor is under publication elsewhere.

4.4. Incompressible Model Gas Flow between Two Infinite Parallel Plates at the Uniform Temperature in a Case of Diffuse Particle Scatterings

The velocity profile generated in the model gas due to the pressure gradient along the channel is analyzed by an analytical method based on the proposed model gas flow described above. Figure 7 shows schematically a one-dimensional model gas system, in which the main control volume (CV) of the unit length volume $dV = \Delta x \Delta y$ is at $y$ within the model gas flow confined by two parallel plates at $y = 0$ and $y = H$. Plate 1 and Plate 2 confine across $y$-axis a model gas flow along the $x$-axis and surfaces 701 and 702 positioned at distance $\Delta x$ bound a portion of the model gas system along the $x$-axis. The known method of obtaining analytical solutions for isothermal gaseous flow with slip boundary conditions is based on the locally fully developed flow assumption and applying the second-order velocity slip boundary conditions (Maxwell-type assumption) in the following form [17]

$$
u - u_w = \pm C_s \lambda_f \frac{\partial u}{\partial y} - C_s \lambda_f^2 \frac{\partial^2 u}{\partial y^2},
$$

(97)

where $u$ is the gas slip velocity near the wall, $u_w$ is the tangential velocity of the wall, and $\lambda_f$ is the mean free path.

The method described below uses none of the Maxwell-type assumptions.

4.4.1. Properties and Features of the Gas-Solid Interface

Here we assign the following additional unique properties to the model gas being in contact with a gas-solid interface revealing mixed diffuse and specular scattering of particles [13]:

1) Each collision on a gas-solid interface of the model gas system, which has
resulted in the scattering of the diffuse particles from the gas-solid interface, is treated as an act of interaction involving a property transport from the gas-solid interface to the scattered particle.

2) Each point of the diffuse particle scattering on the gas-solid interface is treated as a heterogeneous point source for each of the scattered particles.

3) The velocity of each of the heterogeneous point sources on the gas-solid interface equals the velocity of the gas-solid interface of corresponding points of diffuse particle scattering.

4) The point source strength of the heterogeneous point sources on the gas-solid interface is directly proportional to a property accommodation coefficient $\sigma$ in a corresponding point of diffuse particle.

Note: Diffuse scattering is an act of interaction involving property transfer from a gas-solid interface to a scattered particle. Specular scattering does not involve property exchange between the gas-solid interface and a scattered particle.

In the model gas system confined by gas-solid interfaces with the purely diffuse scattering of particles, each particle initiated from the preceding diffuse scatterings from the interfaces delivers in the CV some property obtained from the location of the initial diffuse scattering (ballistic trajectory 703 from plate 2 and ballistic trajectory 704 from plate 1 of Figure 7. Ballistic trajectories 705 and 706 show movement of ballistic particles, which are initiated from the preceding collisions in the gas space, into targeting point $y$. Finally, ballistic trajectories 707 and 708 show the movement of diverging ballistic particles, which are originated from the preceding collisions in space surrounding point $y$. The fluid flow, with the velocity distribution $u_x(y)$, is forced by both the pressure gradient $\frac{dP}{dx}$ and the plates' movement with velocity $u_{x0}$ and $u_{xH}$ along the x-direction.

4.4.2. Direct Validation of the BPPBS in the Gas Space Confined between Two Parallel Plates

Recognizing that in a steady-state model incompressible gas flow at the uniform temperature, all variables describing flow depend on the position in space and are not dependent on time so that the particle density $n$, the mass $m$ and the magnitude of the thermal velocity of the particle/molecule, $v_T$, are constant. In the model gas system confined by the infinite parallel plates being at rest along $y$ direction and characterized by the above conditions, the mass-flow velocity along $y$ direction is expected to be zero. Now, considering that point $y$ is not included in integration, we may formulate the system of integral equations in these forms:

$$Z_e m = \frac{1}{2} P_e Z_e m \int_0^H \exp(-P_e |y - y'|) dy'$$

$$+ Z_e P_e m \exp(-P_e (H - y)) + Z_e P_e m \exp(-P_e y)$$

(98)

for mass-balance, and
for $y$-momentum-balance. Here $P_v$ is the number of particles placed within a collision tube of a unit length, $Z_b$ is the rate of collisions per unit area on plate 1 or plate 2, and $Z_v$ is the rate of collisions per unit volume, which is defined as

$$Z_v = \frac{1}{2} n P_v,$$

and $n_o$ is the unit vector of arbitrary direction from the point $y$, which is defined as

$$n_o = \frac{y - y'}{|y - y'|}.$$

The left-hand of Equation (99) has zero value because of the average momentum of the diverging particle, which is measured by the instant momentum $m v_n n_o$, is zero.

Solving Equation (98) will cause finding that

$$Z_b = \frac{Z_v}{2 P_v}.$$

Substituting Equation (102) in Equation (99), we may verify that the mass and momentum balance in any point of space between plate 1 and plate 2 along $y$ axis are conserved, which suggests the validity of the BPPBS in the one-dimensional configuration.

### 4.4.3. Analytical Derivation of the Velocity Profile Induced in the Model Gas Due to the Pressure Gradient along the Channel between Two Parallel Plates

Step 1: Formulating an integral form of the $u_x$-momentum balance equation

Considering the above, we obtain the following integral form of the $u_x$ momentum balance equation in a steady-state model gas flow [12]:

$$Z_v u_x(y) m = -\frac{dP}{dx} + \frac{1}{2} Z_v P_v m \int_0^y \exp(-P_v |y - y'|) u_x(y') dy' + \frac{1}{2} Z_v u_{y n} m \exp(-P_v (H - y)) + \frac{1}{2} Z_v u_{x n} m \exp(-P_v y),$$

(103)

Step 2: Obtaining a differential form matching to the corresponding integral form of the $u_x$ momentum balance equation.

Here we use the method of differentiation (twice). Each step of the differentiation is followed by the step of subsequent elimination of the integral terms by using the original equation, which is given as:

$$m Z_v \frac{d^2}{dy^2} u_x = \frac{dP}{dx}.$$

(104)

Step 3: Obtaining a general solution having arbitrary coefficients.
Integrating twice the equation above, we obtain:

\[ u_x(y) = Ay^2 + By + C, \tag{105} \]

where

\[ A = \frac{P^2}{2m} \int \frac{1}{Z_v} \frac{dP}{dx} = \frac{P}{mnv_v} \frac{dP}{dx} \tag{106} \]

and \( B \) and \( C \) are the arbitrary coefficients.

Step 4: Determining the values of each of the arbitrary coefficients.

Introducing Equation (105) in Equation (103) (for a specific number of points (two) within the model gas system), we determine the values of each of the arbitrary coefficients. For certainty, we selected points \( y = 0 \) and \( y = H \) for the balance establishment.

Step 5: Solving a system of the obtained in Step 4 algebraic equations.

We obtained the following functional relationship of fluid velocity \( u_x \) from other properties and geometry parameters characterizing the model gas system:

\[ B = -AH + \frac{P}{P_H} \left( \frac{u_{slip} - u_{so}}{P_H + 2} \right) \tag{107} \]

and

\[ C = -A \left( \frac{2}{P^2} \frac{H}{P_v} + \frac{u_{slip}}{P_H + 2} + \frac{u_{so}}{P_H + 2} [P_H + 1] \right). \tag{108} \]

where Plate 1 and Plate 2 are at rest, i.e. \( u_{slip} = u_{so} = 0 \), substitution in Equation (105) of Equations (106), (107), and (108), in which \( P_H \) is replaced by \( Kn^{-1} \), will yield:

\[ u_x(y) = -\frac{H}{mnv_v} \frac{dP}{dx} Kn \left( \frac{1}{Kn} \frac{1}{6} + Kn + \frac{2}{2n} \right), \tag{109} \]

where \( Kn \) is the Knudsen number defined as the ratio of the mean free path \( \lambda_f \sim 1/P_v \) and the representative length scale \( H, i.e., Kn = \lambda_f/H \). Further integration of \( u_x(y) \) in \( y \)-direction followed by normalization with \( -\frac{H}{mnv_v} \frac{dP}{dx} \) results in an expression for a non-dimensional flow rate \( Q_x \):

\[ Q_y = \frac{1}{Kn} \left( \frac{1}{6} + Kn + \frac{2}{2n} \right), \tag{110} \]

The equation above reveals there is a minimum in the normalized mass flow rate (at about \( Kn \equiv 0.3 \)), which is called the Knudsen paradox in the literature \cite{18}. It is explained by the fact that with increasing Knudsen number, the additional diffusive flux becomes significant and deviates from the no-slip solution. Beyond a critical limit (\( Kn > 0.3 \)), the diffusive flux dominates the convective flux \cite{18}.

**Figure 8** compares the non-dimensional flow rates calculated by Equation (110) for the proposed model (Ballistic Model) and a selection of slip models proposed by various authors. The non-dimensional flow rate was proposed in
the form [19]

\[ Q_N = \sqrt{\pi} \left( 1 + 6C_1 Kn + 12C_2 Kn^2 \right) / (12 Kn) \]  \hspace{1cm} (111)

where corresponding slip coefficients \( C_1 \) and \( C_2 \) are taken from Table 6 in [20]. The results of the linearized Boltzmann solution of [21] are plotted from Figure 2 of [17].

In Figure 8, the flow rate according to Equations (110) is rescaled by \( \sqrt{\pi}/2 \) factor for the comparison purposes.

We compared the non-dimensional flow rates calculated by Equation (110) for the Ballistic Model and a selection of slip models proposed by various authors [21] [22] [23] [24], and [25]. The comparison shows that the discrepancy among various slip models, including the model presented here, is small for \( Kn < 0.1 \). However, the significant discrepancy occurs for \( Kn > 0.1 \). Still, the results obtained according to our Ballistic Model are within the reasonable range of discrepancies of variety of the literature data (see also Table 6 in [20]).

4.5. Incompressible Model Gas Flow at the Uniform Temperature in the Space between Two Being at Rest Infinite Parallel Plates in a Case of Mixed Diffuse and Specular Particle Scatterings

Here we assume the symmetricity of the expected solution (because of the zero velocity of the confining plates). We also consider that the influx of \( u_x \) momentum in a given point \( y \) is formed by converging ballistic particles originated from initial collisions within the model gas volume and having at least the last previous specular scattering either from the gas-solid interface of Plate 2 (709) or from the gas-solid interface of Plate 1 (710). With these assumptions, we may obtain the following integral form of the \( u_x \) momentum balance equation in a steady-state model gas flow:

![Figure 8. Comparison of the non-dimensional flow rate \( Q_N \) as a function of the Knudsen number for a selection of slip models and the Ballistic Model.](image)
where $\sigma$ is the momentum accommodation coefficient, which is the probability, for an incident particle, to accommodate momentum from the gas-solid interface and to scatter back in the model gas as a diffuse particle. Comparative analysis of the integral equations describing model gas flow confined between the parallel plates with purely diffuse scattering (see Equation (103)) and mixed diffuse and specular scattering (see Equation (112)) results in finding they have similar forms if

$$ u_{sl} = u_{so} = P_e \exp(-P_e H) \frac{(1-\sigma)}{1-(1-\sigma)\exp(-P_e H)} \int_0^H u_s(y') \exp(P_e y') dy', \quad (113) $$

where $u_{sl}$ and $u_{so}$ represent coefficients of Equation (103), and the right-hand of the equation above represents the similar coefficient in Equation (112). The analysis also leads to the conclusion that the velocity profile can be described by Equation (105), where coefficient $A$ is expressed by Equation (106). Substituting Equation (113) in Equation (107), we obtain:

$$ B = -AH. \quad (114) $$

Finally, substituting $u_s(y') = A(y')^2 + By' + C$ in Equation (113) and substituting the resulting equation in Equation (108), then executing corresponding integrations and algebra operations, we obtain:

$$ C = -AH^2 \left[ \frac{2-\sigma}{\sigma} Kn + 2Kn^2 \right]. \quad (115) $$

Remarkably, the derived tangential slip velocity coefficient $C$ does contain the term being proportional to $\frac{2-\sigma}{\sigma}$. Its appearance is the result of the application of the BM but not the result of usage of the semi-empirical Maxwell-type assumptions as of Equation (97).

Substitution in Equation (105) of Equations (106), (114), and (115) will yield:

$$ u_s(y) = -\frac{H}{mvy} \frac{dP}{dx} \frac{1}{Kn} \left( -\frac{y^2}{H^2} + \frac{y}{H} + \frac{2-\sigma}{\sigma} Kn + 2Kn^2 \right). \quad (116) $$

That the tangential slip velocity coefficient $C$ is analytically defined from the continuum through the slip and transition to free-molecule flow regimes gave us confidence that our approach is valid. Again, our method uses none of the Maxwell-type assumptions. Still, coefficient $C$ contains the terms proportional to $Kn$ and $Kn^2$ in Equation (109) or Equation (116).

### 4.6. Reducing Integro-Differential Forms of Mass and Momentum Balance Equations into Corresponding Differential Forms

According to Tenenbaum and Pollard [26], to test whether an implicit function
defined by the relation \( f(x,y)=0 \) is a solution of a given differential equation; there is a need to show that the function does satisfy a given differential equation on an interval \( a < x < b \). Then the relation \( f(x,y)=0 \) is called an implicit solution of the differential equation.

Definition 3.6 [26]: A relation \( f(x,y)=0 \) will be called an implicit solution of the differential equation \( F(x,y,y',\ldots,y^{(n)})=0 \) on an interval \( a < x < b \), if

1) it defines \( y \) as an implicit function of \( x \) on \( I \), i.e., if there exists a function \( g(x) \) defined on \( I \) such that \( f(x,g(x))=0 \) for every \( x \) on \( I \), and if

2) \( g(x) \) satisfies (117), i.e., if

\[
F(x,g(x),g'(x),\ldots,g^{(n)}(x)) = 0
\]  

(118)

for every \( x \) on \( I \).

The standard procedure in calculus to prove that an implicit function defined by relationship \( f(x,y)=0 \) (a) is a solution of a given differential equation \( F(x,y,y',\ldots,y^{(n)})=0 \) (b) on an interval \( a < x < b \) is the following [26]: “Differentiate (a) implicitly. If it yields (b), then (a) is said to be an implicit solution of (b),” if \( f(x,y) \) defines \( y \) as an implicit function of \( x \) on the same interval \( a < x < b \).

To reduce an equation to an ordinary differential equation, we apply the method of differentiation for integral equations (ones, twice, and so on) with subsequent elimination of the terms belonging to the original equation [27]. Specifically, we will use this technique for reducing vector integro-differential balance equations into corresponding vector differential balance equations. Here we need also note that the point \( \tilde{r} \) should be included when we apply the operator of vector differentiation at this point. In the domain of integration \( \Omega \) in which \( \tilde{r}' \neq \tilde{r} \), operation of differentiation regarding a parameter \( \tilde{r} \) is interchangeable with the operation of the integration over some other variable \( \tilde{r}' \). Even if domain \( \Gamma \) includes singularity point \( \tilde{r} \), one can easily determine that the integrals containing the term \( \nabla \left[ \frac{1}{4\pi} \frac{1}{|\tilde{r} - \tilde{r}'|^2} \right] \) are zeroed. The “fate” of other integrals containing singularity point \( \tilde{r} \) because of the Green function, \( G(\tilde{r},\tilde{r}') \), will be examined afterward.

4.6.1. Reducing the Integro-Differential Form of Mass Balance Equation into a Corresponding Vector Differential Form

Deriving a vector differential form of mass balance equation is shown:

Step 1. Normalizing Equation (76) by \( P_0 \nu_{ad} \), then applying differential operator \( \nabla \) to the left- and right-hands of the equation and executing procedures of neglecting insignificant terms, which are similar to the described in Section III. It leads to establishing the following equality:
Step 2. Applying divergence operator \( \nabla \cdot \) to the left- and right-hands of the equation above and neglecting insignificant terms by the steps, which are similar to the described in Section III:

\[
\nabla \left\{ \frac{1}{P_v \rho_{ref}} \nabla \left[ \frac{1}{P_v \rho_{ref}} \left( \frac{\partial n + Z_v}{\partial t} + \frac{1}{2} \nabla \cdot [n\bar{u}] \right) \right] \right\} \\
\approx - \int_{V_f} Z_f (t', \bar{r}') \frac{1}{v_f (t', \bar{r}') v_f (t, \bar{r})} Q_{10} (t, t') G (\bar{r}, \bar{r}') dV' \ \\
- \int_{V_f} Z_f (t', \bar{r}') \frac{1}{v_f (t', \bar{r}') v_f (t, \bar{r})} Q_{10} (t, t') \frac{1}{4\pi |\bar{r} - \bar{r}'|} \bar{u} (t', \bar{r}') dV' \ \\
- \int_{V_f} Z_f (t', \bar{r}') \frac{1}{v_f (t', \bar{r}') v_f (t, \bar{r})} Q_{10} (t, t') \frac{1}{4\pi |\bar{r} - \bar{r}'|} \tilde{g} \varphi_0 dV' \ \\
+ \frac{1}{P_v \rho_{ref}} \int_{V_f} Z_f (t', \bar{r}') Q_{10} (t, t') (\bar{u} \cdot \nabla) GdV' \ \\
+ \frac{1}{P_v \rho_{ref}} \int_{V_f} Z_f (t', \bar{r}') Q_{10} (t, t') \frac{1}{4\pi |\bar{r} - \bar{r}'|} \tilde{g} dV'. \tag{119}
\]

Step 3. Comparing the first three right-hand terms in the equation above and Equation (76), we may suggest a modification of Equation (76) by introducing a coefficient of proportionality to the mass, \( \frac{1}{v_f^2} \). This results in obtaining of the following equation:

\[
\frac{\partial}{\partial t} \left[ \frac{n}{v_f^2} \right] + \frac{Z_f}{v_f} \left[ \frac{1}{2} \nabla \cdot \left[ \frac{n \bar{u}}{v_f} \right] \right] \\
\approx + P_v \rho_{ref} \int_{V_f} Z_f (t', \bar{r}') \frac{1}{v_f (t', \bar{r}') v_f (t, \bar{r})} Q_{10} (t, t') \frac{1}{4\pi |\bar{r} - \bar{r}'|} dV' \ \\
+ P_v \rho_{ref} \int_{V_f} Z_f (t', \bar{r}') \frac{1}{v_f (t', \bar{r}') v_f (t, \bar{r})} Q_{10} G (\bar{r}, \bar{r}') \cdot \bar{u} (t', \bar{r}') dV' \ \\
+ P_v \rho_{ref} \int_{V_f} Z_f (t', \bar{r}') \frac{1}{v_f (t', \bar{r}') v_f (t, \bar{r})} Q_{10} G (\bar{r}, \bar{r}') \cdot \tilde{g} \varphi_0 dV' \ \\
- \frac{Z_f}{v_f} + \frac{1}{P_v \rho_{ref}} \nabla \cdot \left( \frac{\bar{u} Z_f}{v_f^2} \right). \tag{120}
\]

Substituting the first three right-hand terms of Equation (120) by the
left-hand terms of Equation (121) and rearranging terms, we finally obtain:

$$\frac{\partial}{\partial t} \left( \frac{n}{v_r^2} \right) + \nabla \cdot \left( \frac{\bar{u} n}{v_r^2} \right) = \frac{9}{16} \nabla \cdot \left( \frac{1}{P_v r} \nabla \left[ \frac{1}{P_v r} \left( \frac{\partial}{\partial t} n + \frac{2}{3} n P_v r + \frac{1}{2} \nabla \cdot [\bar{n}\bar{u}] \right) \right] - \frac{1}{2} \frac{n u}{v_r} \cdot \nabla \left( \frac{n v_r}{v_r} \right) \right). \tag{122}$$

Rearranging terms in the equation above, we derived the following reduced vector-differential form of the mass balance equation:

$$\frac{\partial}{\partial t} n + \nabla \cdot (n \bar{u}) = \frac{9}{16} v_r^2 \nabla \cdot \left( \frac{1}{P_v r} \nabla \left[ \frac{1}{P_v r} \left( \frac{\partial}{\partial t} n + \frac{2}{3} n P_v r + \frac{1}{2} \nabla \cdot [\bar{n}\bar{u}] \right) \right] + 2n \frac{\partial}{\partial t} \frac{n v_r}{v_r} + \bar{n} \bar{u} \left( \frac{2 \nabla v_r}{v_r} - \frac{1}{2} \frac{n}{v_r} \right) \right) \tag{123}$$

In the collision-dominated flow regime, the relative change of any property value or any parameter characterizing the model gas is insignificant on the length scale of the average distance between the gas particles or on the time scale of traveling time between consecutive collisions. Therefore, for a high frequency of collisions quantified by $P_v r$ value, the terms in the right hand of the equation above can be neglected, and the equation is reduced to the well-known continuity equation for compressible fluid:

$$\frac{\partial}{\partial t} n + \nabla \cdot (n \bar{u}) = 0, \tag{124}$$

which is identical to Equation (1). Since the vector differentiation of Equation (76) yields Equation (123), which eventually is reduced to Equation (1) and Equation (76) defines $\bar{u}$ at $t \geq 0$ as an implicit function of $\bar{r}$ on $\mathbb{R}^3$, and then Equation (76) is an implicit solution of Equation (1) in the region $\bar{r} \in \mathbb{R}^3$ occupied by the model gas.

### 4.6.2. Reducing the Integro-Differential Form of the Momentum Balance Equation into a Corresponding Vector Differential Form

Deriving a vector differential form of mass balance equation is shown:

Step 1. Normalizing Equation (80) by $P_v r$, applying the differential operator $\nabla$ to the left- and right-hands of the equation, and executing procedures of neglecting insignificant terms, which are similar to those described in Section 3. It leads to establishing the following equality:

$$\frac{1}{P_v r} \nabla \left[ \frac{1}{P_v r} \left( \frac{\partial}{\partial t} \left( \bar{u} n v_r \right) + 2 Z_v \bar{u} + \frac{1}{2} \nabla \left( n \bar{u} \bar{u} \right) - 2 n g + \frac{1}{m} \nabla p \right) \right] = -\iint_v Z_v \frac{1}{v_r} Q_{10} \bar{u}_{10} G dV' + \frac{1}{P_v r} \iint_v Z_v Q_{10} \nabla G dV' - 2 \iint_v Z_v \frac{1}{v_r} Q_{10} \bar{g} dV' \tag{125}$$

$$-2 \iint_v Z_v \frac{1}{v_r} Q_{10} \bar{g} dV' + 2 \frac{1}{P_v r} \iint_v Z_v \frac{1}{v_r} Q_{10} G dV'$$

$$-\iint_v Z_v \frac{1}{v_r} Q_{10} (G \cdot \bar{u}) \bar{u} dV' + \frac{1}{P_v r} \iint_v Z_v \frac{1}{v_r} Q_{10} (G \cdot \nabla) \bar{u} dV'$$
Step 2. Applying divergence operator \( \nabla \cdot \) to the left- and right-hands of the equation above and neglecting insignificant terms by the procedures, which is similar to the described in Section III:

\[
\nabla \cdot \left( \frac{1}{P v_{rel}} \int_{V_r} Z_r \frac{1}{V_r} \frac{1}{Q_{\infty}} \left[ \varphi_{\infty} \left( (\bar{u}g + \bar{g}u) + \bar{g}g \varphi_{\infty} \right) \right] \right) dV' \\
+ \frac{1}{P v_{rel}} \int_{V_r} Z_r \frac{1}{V_r} \frac{1}{Q_{\infty}} \left( \varphi_{\infty} \left( (\bar{u}g + \bar{g}u) + \bar{g}g \varphi_{\infty} \right) \right) \nabla G dV' \\
= P v_{rel} \int_{V_r} Z_r \frac{1}{V_r} \varphi_{\infty} \left[ \left( \bar{u} g + \bar{g} u \right) + \bar{g} g \varphi_{\infty} \right] dV' + \frac{1}{2} \left( \nabla \left( P v_{rel} \right) n \right) \\
+ \frac{1}{2} \nabla n + \frac{n}{V_r} \bar{g} - \frac{2 Z_r}{V_r} \bar{u} + \frac{1}{2} \nabla \left( P v_{rel} \right) \left( \frac{n}{V_r} \bar{u} \bar{u} \right) + \frac{1}{2} \nabla \left( \frac{n}{V_r} \bar{u} \bar{u} \right) .
\]

(126)

Step 3. Comparing the first four right-hand terms in the equation above and Equation (76), we may suggest a modification of Equation (80) by introducing a coefficient of proportionality to the mass, \( \frac{1}{V_r} \). This results in obtaining of the following equation:

\[
\frac{\partial}{\partial t} \left[ \frac{n}{V_r} \bar{u} \right] + \frac{Z_r}{V_r} \bar{u} + \frac{1}{2} \nabla \left( \frac{n}{V_r} \bar{v} \right) + \frac{1}{2} \nabla \cdot \left( \frac{n}{V_r} \bar{u} \bar{u} \right) - \frac{2 n}{V_r} \bar{g} + \frac{1}{m} \nabla \left( \frac{p}{V_r} \right) \\
= P v_{rel} \int_{V_r} Z_r \frac{1}{V_r} \varphi_{\infty} \left[ \left( \bar{u} g + \bar{g} u \right) + \bar{g} g \varphi_{\infty} \right] dV' + \frac{1}{2} \nabla \left( P v_{rel} \right) n \\
+ \frac{1}{2} \nabla n + \frac{n}{V_r} \bar{g} - \frac{2 Z_r}{V_r} \bar{u} + \frac{1}{2} \nabla \left( P v_{rel} \right) \left( \frac{n}{V_r} \bar{u} \bar{u} \right) + \frac{1}{2} \nabla \left( \frac{n}{V_r} \bar{u} \bar{u} \right) .
\]

(127)

Substituting the first five right-hand terms of Equation (126) by the left-hand terms of Equation (127) and rearranging terms, we finally obtain:

\[
\frac{\partial}{\partial t} \left[ \frac{n}{V_r} \bar{u} \right] + \nabla \left( \frac{n}{V_r} \bar{v} \right) + \nabla \cdot \left( \frac{n}{V_r} \bar{u} \bar{u} \right) - \frac{n}{V_r} \bar{g} + \frac{1}{m} \nabla \left( \frac{p}{V_r} \right) \\
\equiv \nabla \left[ \left( \frac{1}{P v_{rel}} \nabla \left[ \frac{1}{P v_{rel}} \varphi_{\infty} \left[ \left( \bar{u} g + \bar{g} u \right) + \bar{g} g \varphi_{\infty} \right] \right] + P v_{rel} n \bar{u} + \frac{1}{2} \nabla \left( n \bar{v} \right) + \frac{1}{2} \nabla \left[ n \bar{u} \bar{u} \right] \right) \\
- 2 n \bar{g} + \frac{1}{m} \nabla \left( \frac{p}{V_r} \right) \right] - \frac{1}{2} \nabla \left( n \bar{v} \right) - \frac{1}{2} \nabla \left( \frac{n}{V_r} \bar{u} \bar{u} \right) .
\]

(128)
Considering the conservation of mass by subtracting Equation (122) multiplied by \( \dot{\mathbf{u}} \) from Equation (128) and rearranging terms, the following reduced vector-differential form of the momentum balance equation is provided:

\[
\frac{\partial}{\partial t}[\dot{\mathbf{u}}] + \frac{\dot{\mathbf{V}}}{n} + (\mathbf{u} \cdot \nabla)\mathbf{u} + \frac{1}{mn} \nabla p - \mathbf{g} = \frac{\dot{\mathbf{V}}}{n} - \mathbf{u} \nabla \cdot \left\{ \frac{1}{P_v \nu_{rel}} \nabla \left[ \frac{\partial}{\partial t}[\dot{n} \mathbf{u}] + 2Z_r \mathbf{u} + \frac{1}{2} \nabla (nv_r^2) + \frac{1}{2} \nabla \cdot [n \mathbf{u} \mathbf{u}] - 2n \mathbf{g} \right] \right\} + \frac{1}{m} \nabla p
\]

Comparing the first, the second, the third, and the fourth left-hand terms of Equations (2) with the first, the third, the fourth, and the fifth left-hand terms of Equation (129), respectively, suggests that each pair of the compared terms is identical. The second left-hand term of the equation above is identified as a normalized self-diffusion force. Additional comparison of the right-hand terms of Equations (2) and (129) may suggest that stress inside the fluid has the following functional dependence:

\[
\frac{1}{\rho} \nabla \cdot \mathbf{T} = \frac{\dot{\mathbf{V}}}{n} - \mathbf{u} \nabla \cdot \left\{ \frac{1}{P_v \nu_{rel}} \nabla \left[ \frac{\partial}{\partial t}[\dot{n} \mathbf{u}] + 2Z_r \mathbf{u} + \frac{1}{2} \nabla (nv_r^2) + \frac{1}{2} \nabla \cdot [n \mathbf{u} \mathbf{u}] - 2n \mathbf{g} \right] \right\} + \frac{1}{m} \nabla p
\]

Since the vector differentiation of Equation (80) yields Equation (129), which resembles Equation (2) and Equation (80) defines \( \dot{\mathbf{u}} \) at \( t \geq 0 \) as an implicit function of \( \mathbf{r} \) on \( \mathbb{R}^3 \), then Equation (80) is an implicit solution of Equation (2) in the region \( \mathbf{r} \in \mathbb{R}^3 \) occupied by the model gas.

5. Conclusions

1) Modeling of fluid dynamics problems by the NAMDT is based on the recognition that each particle composing the model gas travels with a probability between any of two points in space occupied by the model gas while following a ballistic trajectory governed by a law of motion in free space. Each ballistic particle is treated as a property carrier transporting one or more of mass, momentum, and energy between the points of consecutive collisions and each point in space occupied by the model gas is both a sink accumulating property delivered by converging ballistic particles from the entire model gas system and a source dispersing property by diverging ballistic particles.

2) Based on the proposed model gas properties, we formulated the Ballistic Principle of the Property Balance in the Space occupied by the gas, application of
which may simplify and reduce computations in applications dealing with modeling of fluid dynamics problems.

3) Following the above principles, a general integro-differential form of the property balance equation is proposed. The general integro-differential form of the property balance is further modified to derive the integro-differential forms of mass balance, momentum balance, and energy balance.

4) The following two direct tests validate the BPPBS:
   a) demonstration of the mass-balance and momentum-balance conservation in a given non-moving point in three-dimensional unlimited incompressible gas space with a lack of gravitational force at the uniform temperature by analytical verification of the balance between the rate of mass and momentum influx and the rate of mass and momentum efflux, respectively, in the given point of the gas space;
   b) demonstration of the mass-balance and momentum-balance conservation in a given non-moving point in one-dimensional incompressible gas space confined between two parallel plates at the uniform temperature by analytical verification of the balance between the rate of mass and momentum influx and the rate of mass and momentum efflux, respectively, in the given point of the gas space.

5) The BPPBS is also validated by demonstrating that, in the collision dominated flow regime, the differential equations, which we converted from the derived integro-differential mass and momentum balance equations, are identical to the corresponding Navier-Stokes equations. This finding supports the assumption that, in the collision-dominated flow regime, the formulated integro-differential forms of the balance are exact implicit solutions for corresponding Navier-Stokes equations.

6) The analytical solution for determining the velocity profile induced in the model gas flow due to the pressure gradient along the channel is demonstrated. That the analytical solution is valid to explain velocity profiles in the wide range of gas pressure from the continuum through the slip and transition to free-molecule flow regimes gives additional confidence that the BM is valid.

7) From the practical viewpoint, the proposed method can be a fundamental base of a new generation of the CFD software in which the solver does not use the Navier-Stokes or Boltzmann equations. Using the solver operating on the exact implicit solutions of the balance equations will eliminate uncertainty, improve predictability, and shorten the computational time.

8) Although the BM is formulated to solve the CFD problems, it can be eventually used to model any dynamic system composed of presumably chaotically moving particles/elements, each carrying a specific amount of property/information.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.
References

https://www.grc.nasa.gov/www/k-12/airplane/nseqs.html


https://doi.org/10.1201/9781420050905.ch4

https://doi.org/10.1063/1.2393436


https://doi.org/10.1016/0370-1573(92)90090-M


https://doi.org/10.1088/0256-307X/21/9/028

https://en.wikipedia.org/wiki/Molecular_dynamics

https://doi.org/10.1016/j.compfluid.2015.03.023


https://doi.org/10.1201/9781439898212

https://doi.org/10.1177/1740349913486097


https://doi.org/10.1142/S0129183107010383


DOI: 10.4236/jamp.2020.86084

N. Kislov

Journal of Applied Mathematics and Physics


Collocation Method for Solving the Generalized KdV Equation

Turabi Geyikli

Department of Mathematics, Faculty of Education, Adiyaman University, Adiyaman, Turkey
Email: tgeyikli@adiyaman.edu.tr

Abstract

In this work, we have obtained numerical solutions of the generalized Korteweg-de Vries (GKdV) equation by using septic B-spline collocation finite element method. The suggested numerical algorithm is controlled by applying test problems including: single soliton wave. Our numerical algorithm, attributed to a Crank Nicolson approximation in time, is unconditionally stable. To control the performance of the newly applied method, the error norms, $L_2$ and $L_\infty$ and invariants $I_1$, $I_2$ and $I_3$ have been calculated. Our numerical results are compared with some of those available in the literature.

Keywords

Generalized Korteweg-de Vries Equation, Finite Element Method, Collocation, Septic B-Spline, Soliton

1. Introduction

Several physical processes for example dispersion of long waves in shallow water waves under gravity, bubble-liquid mixtures, ion acoustic plasma waves, fluid mechanics, nonlinear optics and wave phenomena in enharmonic crystals can be expressed by the KdV equation which was first introduced by Korteweg and de Vries [1]. The equation was solved analytically by Zabusky, Fornberg and Whitham, [2] [3]. Zabusky and Kruskal [4] were first obtained numerical solutions of the equation with finite difference method. Gardner et al. [5] demonstrated existence and uniqueness of solutions of the KdV equation. Several scientists have used various numerical methods including pseudospectral method [4], finite difference method [6] [7], finite element method [8]-[15] and heat balance integral method [15] to solve the equation. Numerical solutions of the KdV equation were obtained using differential quadrature method based on cosine ex-
pansion by B. Saka [16]. Like the KdV equation, in recent years, various numerical methods have been improved for the solution of the MKdV equation. Kaya [17] calculated the explicit solutions of the higher order modified Korteweg de-Vries equation by Adomian decomposition method. MKdV equation has been solved by using Galerkins’ method with quadratic B-spline finite elements by Biswas et al. [18]. Raslan and Baghdady [19] [20] indicated the accuracy and stability of the difference solution of the MKdV equation and they obtained the numerical aspects of the dynamics of shallow water waves along lakes’ shores and beaches modeled by the MKdV equation. A new variety of (3 + 1)-dimensional MKdV equations and multiple soliton solutions for each new equation were established by Wazwaz [21] [22]. Lumped Petrov-Galerkin and Galerkin methods were practiced to the MKdV equation by Ak et al. [23] [24].

GKdV equation has received much less attention, presumably because of its higher nonlinearity for \( p > 2 \). The symmetry group was calculated for the equation and several classes of solutions were obtained in [25]. Liu and Yi [26] developed and analyzed a Hamiltonian preserving DG method for solving the generalized KdV equation. The initial value problem of a kind of GKdV equations is considered by using Sobolev space theory and finite element method by Lai et al. [27]. Alvarado and Omel’yanov [28] create a finite differences scheme to simulate the solution of the Cauchy problem and present some numerical results for the problem of the solitary waves interaction. A class of fully discrete scheme for the generalized Korteweg-de Vries equation in a bounded domain \((0, L)\) has studied by Sepulveda and Villagrán [29]. Collocation finite element method based on quintic B-spline functions is applied to the generalized KdV equation by Ak et al. [30]. Solitary wave solution for the GKdV equation by using ADM has been obtained by Ismail et al. [31].

In this article, we will take in consideration for the following GKdV equation

\[
U_t + \varepsilon U^2 U_x + \mu U_{xxx} = 0, \tag{1}
\]

with the homogeneous boundary conditions

\[
U(a,t) = 0, \quad U(b,t) = 0,
\]

\[
U_x(a,t) = 0, \quad U_x(b,t) = 0, \quad t > 0
\]

and an initial condition

\[
U(x,0) = U_0(x), \quad a \leq x \leq b,
\]

where \( t \) is time, \( x \) is the space coordinate, \( \varepsilon \) and \( \mu \) are positive parameters.

One of the primary mathematical models for describing the theory of water waves in shallow channels is the following Korteweg de-Vries (KdV) equation:

\[
U_t + \varepsilon U U_x + \mu U_{xxx} = 0. \tag{4}
\]

The terms \( UU_x \) and \( U_{xxx} \) in the Equation (4) stand for the nonlinear convection and dispersion, respectively. In this paper, we have numerically solved the GKdV equation using collocation method with septic B-spline finite elements. We have investigated the motion of a single soliton wave to show the
performance and profiency of the proposed method. Also we have showed the suggested method is unconditionally stable applying the von-Neumann stability analysis.

2. Septic B-Spline Collocation Method

We think of a mesh \( a = x_0 < x_1 < \cdots < x_N = b \) as a uniform divide of the solution area \( a \leq x \leq b \) by the points \( x_m \) with \( h = \frac{b-a}{N} = x_{m+1} - x_m \). The septic B-splines \( \phi_m(x), (m = -3, -2, \cdots, N + 3) \) at the knots \( x_m \) are given by [32]

\[
\phi_m(x) = \frac{1}{h} \begin{cases} 
\left(x-x_{m-4}\right)^7 & [x_{m-4},x_{m-3}] \\
\left(x-x_{m-4}\right)^7 - 8\left(x-x_{m-3}\right)^7 & [x_{m-3},x_{m-2}] \\
\left(x-x_{m-4}\right)^7 - 8\left(x-x_{m-3}\right)^7 + 28\left(x-x_{m-2}\right)^7 & [x_{m-2},x_{m-1}] \\
\left(x-x_{m-4}\right)^7 - 8\left(x-x_{m-3}\right)^7 + 28\left(x-x_{m-2}\right)^7 - 56\left(x-x_{m-1}\right)^7 & [x_{m-1},x_m] \\
\left(x_{m+4}-x\right)^7 - 8\left(x_{m+3}-x\right)^7 + 28\left(x_{m+2}-x\right)^7 - 56\left(x_{m+1}-x\right)^7 & [x_{m+1},x_{m+2}] \\
\left(x_{m+4}-x\right)^7 - 8\left(x_{m+3}-x\right)^7 + 28\left(x_{m+2}-x\right)^7 & [x_{m+2},x_{m+3}] \\
\left(x_{m+4}-x\right)^7 - 8\left(x_{m+3}-x\right)^7 \\
0 & \text{otherwise}
\end{cases}
\]

The set of septic B-spline functions \( \{\phi_3(x), \phi_2(x), \cdots, \phi_{N+2}(x), \phi_{N+3}(x)\} \) forms a basis for the problem region of solution \( [a,b] \). The approximate solution \( U_N(x,t) \) to the exact solution \( U(x,t) \) in the form:

\[
U_N(x,t) = \sum_{m=-3}^{N+3} \phi_m(x) \delta_m(t)
\]

where \( \phi_m(x) \) are septic B-splines and \( \delta_m(t) \) are time dependent parameters to be identified from the boundary and collocation conditions. A characteristic finite interval \([x_m,x_{m+1}]\) is turn into the interval \([0,1]\) by a domestic coordinate conversion described by \( h\xi = x-x_m, \quad 0 \leq \xi \leq 1 \). So septic B-splines (5) in terms of \( \xi \) over \([0,1]\) can be written as

\[
\phi_{m-3} = 1-7\xi + 21\xi^2 - 35\xi^3 + 35\xi^4 - 21\xi^5 + 7\xi^6 - \xi^7,
\phi_{m-2} = 120 - 392\xi + 504\xi^2 - 280\xi^3 + 84\xi^4 - 42\xi^5 + 7\xi^6,
\phi_{m-1} = 1191 - 1715\xi + 315\xi^2 + 665\xi^3 - 315\xi^4 - 105\xi^5 + 105\xi^6 - 21\xi^7,
\phi_m = 2416 - 1680\xi + 560\xi^2 - 140\xi^3 + 35\xi^4,
\phi_{m+1} = 1191 + 1715\xi + 315\xi^2 - 665\xi^3 - 315\xi^4 + 105\xi^5 + 105\xi^6 - 35\xi^7,
\phi_{m+2} = 120 + 392\xi + 504\xi^2 + 280\xi^3 - 84\xi^4 - 42\xi^5 + 21\xi^6,
\phi_{m+3} = 1 + 7\xi + 21\xi^2 + 35\xi^3 + 35\xi^4 + 21\xi^5 + 7\xi^6 - \xi^7,
\phi_{m+4} = \xi^7.
\]

Using Equation (5) and Equation (6), the nodal values of \( U_m, U'_m, U''_m, U'''_m \) and \( U''''_m \) at the knots \( x_m \) are obtained as the following:
\[
U_{N}(x_{m},t) = U_{m} = \delta_{m-3} + 120\delta_{m-2} + 1191\delta_{m-1} + 2416\delta_{m} + 1191\delta_{m+1} + 120\delta_{m+2} + \delta_{m+3},
\]

\[
U_{m}^* = \frac{7}{h}(-\delta_{m-3} - 56\delta_{m-2} - 245\delta_{m-1} + 245\delta_{m+1} + 56\delta_{m+2} + \delta_{m+3}),
\]

\[
U_{m}^{**} = \frac{42}{h^2}(-\delta_{m-3} + 24\delta_{m-2} + 15\delta_{m-1} - 80\delta_{m} + 15\delta_{m+1} + 24\delta_{m+2} + \delta_{m+3}),
\]

\[
U_{m}^{***} = \frac{210}{h^3}(-\delta_{m-3} - 8\delta_{m-2} + 19\delta_{m-1} - 19\delta_{m+1} + 8\delta_{m+2} + \delta_{m+3}),
\]

\[
U_{m}^{iv*} = \frac{840}{h^4}(-\delta_{m-3} - 9\delta_{m-2} + 16\delta_{m} - 9\delta_{m+1} + \delta_{m+3})
\]

where the symbols \(',**,***\) symbolize differentiation according to \(x\), respectively.

Using (5) and (8) in the Equation (1) this guides to a set of ordinary differential equations of the form

\[
\frac{\delta_{m-3} + 120\delta_{m-2} + 1191\delta_{m-1} + 2416\delta_{m} + 1191\delta_{m+1} + 120\delta_{m+2} + \delta_{m+3}}{h^3 + \frac{7EZ_{m}}{h}(-\delta_{m-3} - 56\delta_{m-2} - 245\delta_{m-1} + 245\delta_{m+1} + 56\delta_{m+2} + \delta_{m+3}) + \frac{210\mu}{h^3}(-\delta_{m-3} - 8\delta_{m-2} + 19\delta_{m-1} - 19\delta_{m+1} + 8\delta_{m+2} + \delta_{m+3})} = 0,
\]

where

\[
Z_{m} = \left(\delta_{m-3} + 120\delta_{m-2} + 1191\delta_{m-1} + 2416\delta_{m} + 1191\delta_{m+1} + 120\delta_{m+2} + \delta_{m+3}\right)^2.
\]

If time parameters \(\delta_{i}\) and its time derivatives \(\dot{\delta}_{i}\) in Equation (9) are separated by the Crank-Nicolson form and finite difference approach, respectively:

\[
\delta_{i} = \frac{\delta_{i}^{n+1} + \delta_{i}^{n}}{2},
\]

and usual finite difference approximation

\[
\dot{\delta}_{i} = \frac{\delta_{i}^{n+1} - \delta_{i}^{n}}{\Delta t}
\]

we acquired a repetition relationship between two time levels \(n\) and \(n+1\) relating two unknown parameters \(\delta_{i}^{n+1}, \delta_{i}^{n}\) for \(i = m-3, m-2, \cdots, m+2, m+3\)

\[
\gamma_{1}\delta_{m-3}^{n+1} + \gamma_{2}\delta_{m-2}^{n+1} + \gamma_{3}\delta_{m-1}^{n+1} + \gamma_{4}\delta_{m}^{n+1} + \gamma_{5}\delta_{m+1}^{n+1} + \gamma_{6}\delta_{m+2}^{n+1} + \gamma_{7}\delta_{m+3}^{n+1} = \gamma_{1}\delta_{m-3}^{n} + \gamma_{2}\delta_{m-2}^{n} + \gamma_{3}\delta_{m-1}^{n} + \gamma_{4}\delta_{m}^{n} + \gamma_{5}\delta_{m+1}^{n} + \gamma_{6}\delta_{m+2}^{n} + \gamma_{7}\delta_{m+3}^{n},
\]

where

\[
\gamma_{1} = [1 - EZ_{m} - M],
\gamma_{2} = [120 - 56EZ_{m} - 8M],
\gamma_{3} = [1191 - 245EZ_{m} + 19M],
\gamma_{4} = [2416],
\gamma_{5} = [1191 + 245EZ_{m} - 19M],
\gamma_{6} = [120 + 56EZ_{m} + 8M],
\gamma_{7} = [1 + EZ_{m} + M],
\]

\[
m = 0, 1, \cdots, N, E = \frac{7E}{2h}, M = \frac{105\mu}{h^3}, \Delta t = M = \frac{105\mu}{h^3}.\]
The system (12) contains $(N+1)$ linear equations containing $(N+7)$ unknown coefficients $\left(\delta_3, \delta_2, \delta_{1}, \dots, \delta_{N+1}, \delta_{N+2}, \delta_{N+3}\right)^T$. To acquire a solution of this system, we require six additional restrictions. These are obtained from the boundary conditions (2) and can be used to remove $\delta_3, \delta_2, \delta_1$ and $\delta_{N+1}, \delta_{N+2}, \delta_{N+3}$ from the systems (12) which occurs a matrix equation for the $N+1$ unknowns $d^n = (\delta_0, \delta_1, \delta_{-1}, \delta_N)^T$ of the form

$$Ad^{n+1} = Bd^n.$$ (14)

The resulting system is effectively solved with a version of the Thomas algorithm and we implement an inner iteration $\delta'^n = \delta^n + \frac{1}{2}(\delta^n - \delta^{n-1})$ at each time step to overcome the non-linearity caused by $Z_m$. Before the beginning of the solution procedure, initial parameters $d^0$ are established by using the initial condition and following derivatives at the boundaries;

$$U_n(x,0) = U(x_n,0); \quad m = 0,1,2,\cdots,N$$ (15)

$$(U_n)'(a,0) = 0, \quad (U_n)'(b,0) = 0,$$ (16)

$$(U_n)_{xx}(a,0) = 0, \quad (U_n)_{xx}(b,0) = 0,$$ (17)

$$(U_n)_{xxx}(a,0) = 0, \quad (U_n)_{xxx}(b,0) = 0.$$ (18)

So, by taking account (18), we obtain the following matrix form for the initial vector $d^0$;

$$Wd^0 = b,$$

where

$$W = \begin{bmatrix}
1536 & 2712 & 768 & 24 \\
82731 & 210568.5 & 104796 & 10063.5 \\
81 & 81 & 81 & 81 \\
9600 & 96597 & 195768 & 96474 \\
81 & 81 & 81 & 81 \\
\vdots & & & & \\
\end{bmatrix}.$$ (19)

$$d^0 = (\delta_0, \delta_1, \delta_{-1}, \delta_{N-2}, \delta_{N-1}, \delta_N)^T$$

and

$$b = (U(x_0,0), U(x_1,0), \cdots, U(x_{N-1},0), U(x_N,0))^T.$$ (20)

3. Stability Analysis

To implement the von Neumann stability analysis, GKDv equation is linearized...
by thinking about the quantity \( U^p \) in the nonlinear term \( U^p U_x \) is locally invariable. Substituting the Fourier mode \( \delta_m^e = \xi e^{im\xi} \), \( i = \sqrt{-1} \) in which \( k \) is a mode number and \( h \) is the element size, into the Equation (12) gives the growth factor \( \xi \) of the form

\[
\xi = \frac{A-iB}{A+iB}
\]

where

\[
A = 2\cos(3kh) + 240\cos(2kh) + 2382\cos(kh) + 2416
\]

\[
B = 2(EZ_m + M)\sin(3kh) + 2(56Ez_m + 8M)\sin(2kh)
\]

\[
+2(245EZ_m - 19M)\sin(kh)
\]

The modulus of the (19) is found 1, hence the linearized algorithm is unconditionally stable.

4. Test Problems

In this section, we introduce some numerical examples including: motion of single soliton wave whose exact solution is known to test validity of our algorithm for solving GKDv equation. The initial boundary value problem (1)-(2) possesses following conservative quantities;

\[
I_1 = \int_{-\infty}^{\infty} U(x,t)dx,
\]

\[
I_2 = \int_{-\infty}^{\infty} U^2(x,t)dx,
\]

\[
I_3 = \int_{-\infty}^{\infty} \left[ U^{p+2}(x,t) - \frac{\mu(p+1)(p+2)}{2e} U^2_x(x,t) \right] dx
\]

which correspond to the mass, momentum and energy of the shallow water waves, respectively [33] [34]. To compare the numerical solution with the exact solution we use the following error norms:

\[
L_2 = \| U_{\text{exact}} - U_N \|_2 \simeq \sqrt{\frac{\sum_{j=0}^{N} |U_{\text{exact},j} - (U_N)_j|^2}{\sum_{j=0}^{N} |U_{\text{exact},j} - (U_N)_j|^2}}
\]

and

\[
L_\infty = \| U_{\text{exact}} - U_N \|_\infty \simeq \max_j |U_{\text{exact},j} - (U_N)_j|
\]

The Motion of Single Solitary Wave

For this test problem, Equation (1) is examined with the boundary conditions

\( U \rightarrow 0 \) as \( x \rightarrow \pm\infty \) and the initial condition \( U(x,0) = \text{Asech}^2 \left[ k(x-x_0) \right] \)

where \( A = \left[ \frac{c(p+1)(p+2)}{2e} \right]^{1/2} \) is amplitude and \( k = \frac{p}{2} \sqrt{\frac{c}{\mu}} \) is width of the single soliton. The exact soliton solution of the GKDv equation is

\[
U(x,t) = \text{Asech}^2 \left[ k(x-x_0 - ct) \right].
\]
where \( c \) and \( x_0 \) are arbitrary constants. In order to exemplify the validity of our numerical algorithm, we conceive the first case of a single soliton solution for the parameters \( p = 1, \ v = 1, \ \mu = 4.84 \times 10^{-4}, \ h = 0.01, \ \Delta t = 0.005, \ c = 0.3 \) and \( v = 3, \ \mu = 1, \ h = 0.1, \ \Delta t = 0.01 \) through the interval \([0, 80]\) to compare with that of previous papers [8] [10] [14] [16] [30]. All parameters are given in all refarans. For these parameters, the single solitary wave has the amplitude 0.9 and 0.3, respectively. The three invariants \( I_1, I_2, \) and \( I_3 \) together with the \( L_2, L_\infty \) error norms for the problem are documented and compared in Table 1 for times up to \( t = 1 \). As seen from the table that \( L_2 \) and \( L_\infty \) error norms are found small enough and the conservation of the invariants can be seen to be almost constant.

Solitary wave profiles are demonstrated at \( t = 0, 0.1, 0.2, \ldots, 1 \) in Figure 1 in which the soliton moves to the right at a nearly unchanged speed and amplitude as time increases, as expected.

### Table 1. Comparisons of results for invariants and error norms with \( p = 1, \ v = 1, \ \mu = 4.84 \times 10^{-4}, \ c = 0.3, \ h = 0.01, \ \Delta t = 0.005, \ x \in [0, 2] \) and \( v = 3, \ \mu = 1, \ c = 0.3, \ h = 0.1, \ \Delta t = 0.01 \).

<table>
<thead>
<tr>
<th>Method</th>
<th>Time</th>
<th>( I_1 )</th>
<th>( I_2 )</th>
<th>( I_3 )</th>
<th>( L_2 \times 10^3 )</th>
<th>( L_\infty \times 10^3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mu = 4.84 \times 10^{-4} ) Present Method</td>
<td>0.00</td>
<td>0.144598</td>
<td>0.086759</td>
<td>0.046850</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0.25</td>
<td>0.144598</td>
<td>0.086759</td>
<td>0.046850</td>
<td>0.02315</td>
<td>0.06802</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>0.144598</td>
<td>0.086759</td>
<td>0.046850</td>
<td>0.04525</td>
<td>0.12487</td>
</tr>
<tr>
<td></td>
<td>0.75</td>
<td>0.144598</td>
<td>0.086759</td>
<td>0.046850</td>
<td>0.06683</td>
<td>0.18353</td>
</tr>
<tr>
<td></td>
<td>1.00</td>
<td>0.144593</td>
<td>0.086759</td>
<td>0.046850</td>
<td>0.09082</td>
<td>0.23617</td>
</tr>
<tr>
<td>[16]</td>
<td>1.00</td>
<td>0.144598</td>
<td>0.086759</td>
<td>0.046850</td>
<td>0.13010</td>
<td>0.36895</td>
</tr>
<tr>
<td>[8]</td>
<td>1.00</td>
<td>0.144592</td>
<td>0.086759</td>
<td>0.016870</td>
<td>22.2</td>
<td></td>
</tr>
<tr>
<td>[14] MQ</td>
<td>1.00</td>
<td>0.144606</td>
<td>0.086759</td>
<td>0.046850</td>
<td>0.062</td>
<td>0.133</td>
</tr>
<tr>
<td>[14] IMQ</td>
<td>1.00</td>
<td>0.144623</td>
<td>0.086765</td>
<td>0.046847</td>
<td>2.751</td>
<td>5.018</td>
</tr>
<tr>
<td>[14] IQ</td>
<td>1.00</td>
<td>0.144598</td>
<td>0.086759</td>
<td>0.046849</td>
<td>1.013</td>
<td>2.090</td>
</tr>
<tr>
<td>[14] TPS</td>
<td>1.00</td>
<td>0.144261</td>
<td>0.086762</td>
<td>0.046842</td>
<td>2.606</td>
<td>6.345</td>
</tr>
<tr>
<td>[14] G</td>
<td>1.00</td>
<td>0.144601</td>
<td>0.086760</td>
<td>0.046850</td>
<td>0.046</td>
<td>0.136</td>
</tr>
<tr>
<td>[30]</td>
<td>1.00</td>
<td>0.144599</td>
<td>0.086759</td>
<td>0.046850</td>
<td>0.079</td>
<td>0.238</td>
</tr>
<tr>
<td>( \mu = 1 ) Present Method</td>
<td>0.00</td>
<td>2.190844</td>
<td>0.438176</td>
<td>0.078871</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0.25</td>
<td>2.190844</td>
<td>0.438176</td>
<td>0.078871</td>
<td>0.038</td>
<td>0.051</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>2.190858</td>
<td>0.438176</td>
<td>0.078871</td>
<td>0.040</td>
<td>0.048</td>
</tr>
<tr>
<td></td>
<td>0.75</td>
<td>2.190848</td>
<td>0.438176</td>
<td>0.078871</td>
<td>0.059</td>
<td>0.113</td>
</tr>
<tr>
<td></td>
<td>1.00</td>
<td>2.190873</td>
<td>0.438176</td>
<td>0.078863</td>
<td>0.083</td>
<td>0.155</td>
</tr>
</tbody>
</table>
Figure 1. Motion of single solitary wave for (a) \( p = 1, \varepsilon = 1, \mu = 4.84 \times 10^{-4}, c = 0.3, h = 0.01, \Delta t = 0.005 \) and (b) \( \varepsilon = 3, \mu = 1, c = 0.3, h = 0.1, \Delta t = 0.01 \).

For the second set, we choose the parameters \( p = 2, \varepsilon = 3, \mu = 1, h = 0.1, \Delta t = 0.01, c = 0.845 \) and 0.3 throughout the interval \([0, 80]\) just to be able to compare them with earlier papers [19] [23] [24]. These values yield the amplitude 1.3416 and 0.7746 and the computations are done until time \( t = 20 \) and \( t = 1 \). We calculate the values of the error norms \( L_2 \) and \( L_\infty \) and invariants \( I_1, I_2 \) and \( I_3 \) for different time levels and compare them with earlier papers in Table 2. This table indicates that the error norms obtained by our method are found much better than the others and the calculated values of invariants are in good conformity with the others. The motion of solitary wave using our scheme are plotted at times \( 0, 5, 10, 15, 20 \) and \( 0, 0.1, 0.2, \ldots, 1 \) in Figure 2 in which the soliton moves to the right at a nearly unchanged speed and amplitude as time increases, as expected.

Finally, we have taken the parameters \( p = 3, \varepsilon = 1, \mu = 4.84 \times 10^{-4}, h = 0.01, \Delta t = 0.005, c = 0.3 \) and \( \varepsilon = 6, \mu = 1, h = 0.1, \Delta t = 0.01, c = 0.6 \) over the region \([0, 80]\). Thereby, solitary wave has amplitude 1.44 and 1.0, respectively. Simulations are executed to time \( t = 1 \) to invent the error norms \( L_2 \) and \( L_\infty \) and the numerical invariants \( I_1, I_2 \) and \( I_3 \). The calculated values are presented in Table 3. As can be seen in Table 3, the error norms \( L_2 \) and \( L_\infty \) are sensibly small and the quantities of the invariants remain almost constant during the computer run. The behaviors of solutions for values of \( h = t = 0.125 \) at times \( t = 0.20 \) and 40 are depicted in Figure 3.

5. Conclusion

In this paper, a septic B-spline collocation method has been successfully applied to the GkDv equation to examine the motion of a single solitary wave whose analytical solution is known. To show how good and accurate the numerical solutions of the test problems, we have computed the error norms \( L_2 \) and \( L_\infty \) and conserved quantities \( I_1, I_2 \) and \( I_3 \). According to the tables in the paper, one can have easily seen that our error norms are enough small and the invariants are well conserved. Stability analysis has been done and the linearized numerical scheme has been obtained unconditionally stable. Thus, we can say
Figure 2. Motion of single solitary wave for (a) $p = 2$, $\varepsilon = 3$, $\mu = 1$, $h = 0.1$, $\Delta t = 0.01$, $c = 0.845$ and (b) $c = 0.3$, $h = 0.1$, $\Delta t = 0.01$.

Figure 3. Motion of single solitary wave for (a) $p = 3$, $\varepsilon = 3$, $\mu = 1$, $h = 0.1$, $\Delta t = 0.005$, $c = 0.845$ and (b) $c = 0.3$, $h = 0.1$, $\Delta t = 0.01$.

Table 2. Comparisons of results for invariants and error norms with $p = 2$, $\varepsilon = 3$, $\mu = 1$, $h = 0.1$, $\Delta t = 0.01$, $c = 0.845$ and $c = 0.3$, $h = 0.1$, $\Delta t = 0.01$.

<table>
<thead>
<tr>
<th>Method</th>
<th>Time</th>
<th>$I_1$</th>
<th>$I_2$</th>
<th>$I_3$</th>
<th>$L_\infty \times 10^1$</th>
<th>$L_\infty \times 10^1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c = 0.845$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Present Method</td>
<td>0</td>
<td>4.442865</td>
<td>3.676941</td>
<td>2.071335</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>4.442865</td>
<td>3.676941</td>
<td>2.073758</td>
<td>0.917706</td>
<td>0.562852</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>4.442865</td>
<td>3.676941</td>
<td>2.073900</td>
<td>1.265494</td>
<td>0.850150</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>4.442865</td>
<td>3.676941</td>
<td>2.073930</td>
<td>1.638275</td>
<td>1.096426</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>4.442865</td>
<td>3.676941</td>
<td>2.073948</td>
<td>1.983089</td>
<td>1.309575</td>
</tr>
<tr>
<td>[23]</td>
<td>20</td>
<td>4.442866</td>
<td>3.676941</td>
<td>2.073846</td>
<td>3.641638</td>
<td>2.285638</td>
</tr>
<tr>
<td>$c = 0.3$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Present Method</td>
<td>0.00</td>
<td>4.442815</td>
<td>2.190881</td>
<td>0.438173</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>0.25</td>
<td>4.442818</td>
<td>2.190881</td>
<td>0.438179</td>
<td>0.046</td>
<td>0.030</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>4.442818</td>
<td>2.190881</td>
<td>0.438191</td>
<td>0.073</td>
<td>0.042</td>
</tr>
<tr>
<td></td>
<td>0.75</td>
<td>4.442818</td>
<td>2.190881</td>
<td>0.438202</td>
<td>0.091</td>
<td>0.048</td>
</tr>
<tr>
<td></td>
<td>1.00</td>
<td>4.442818</td>
<td>2.190881</td>
<td>0.438213</td>
<td>0.105</td>
<td>0.051</td>
</tr>
<tr>
<td>[19]</td>
<td>1.00</td>
<td>4.44192</td>
<td>2.18994</td>
<td>0.437763</td>
<td>-</td>
<td>0.310</td>
</tr>
<tr>
<td>[19]</td>
<td>1.00</td>
<td>4.44198</td>
<td>2.18974</td>
<td>0.437642</td>
<td>-</td>
<td>0.325</td>
</tr>
</tbody>
</table>
Table 3. Values of the invariants and error norms for $p = 3$, $\varepsilon = 3$, $\mu = 1$, $h = 0.01$, $\Delta t = 0.005$, $c = 0.845$ and $c = 0.3$, $h = 0.1$, $\Delta t = 0.01$.

<table>
<thead>
<tr>
<th>Method</th>
<th>Time</th>
<th>$I_1$</th>
<th>$I_2$</th>
<th>$I_3$</th>
<th>$L_1 \times 10^3$</th>
<th>$L_2 \times 10^3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c = 0.845$</td>
<td>0</td>
<td>0.162456</td>
<td>0.144101</td>
<td>0.061758</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Present Method</td>
<td>0.25</td>
<td>0.162456</td>
<td>0.144100</td>
<td>0.061756</td>
<td>0.443</td>
<td>1.541</td>
</tr>
<tr>
<td></td>
<td>0.5</td>
<td>0.162458</td>
<td>0.144100</td>
<td>0.061755</td>
<td>0.926</td>
<td>3.062</td>
</tr>
<tr>
<td></td>
<td>0.75</td>
<td>0.162456</td>
<td>0.144099</td>
<td>0.061753</td>
<td>1.456</td>
<td>4.867</td>
</tr>
<tr>
<td></td>
<td>1.00</td>
<td>0.162450</td>
<td>0.144099</td>
<td>0.061752</td>
<td>2.008</td>
<td>6.645</td>
</tr>
<tr>
<td>$c = 0.3$</td>
<td>0.00</td>
<td>3.620369</td>
<td>2.226620</td>
<td>0.318081</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Present Method</td>
<td>0.25</td>
<td>3.620352</td>
<td>2.226620</td>
<td>0.318081</td>
<td>0.026</td>
<td>0.034</td>
</tr>
<tr>
<td></td>
<td>0.50</td>
<td>3.620352</td>
<td>2.226620</td>
<td>0.318080</td>
<td>0.029</td>
<td>0.031</td>
</tr>
<tr>
<td></td>
<td>0.75</td>
<td>3.620334</td>
<td>2.226620</td>
<td>0.318079</td>
<td>0.043</td>
<td>0.076</td>
</tr>
<tr>
<td></td>
<td>1.00</td>
<td>3.620342</td>
<td>2.226620</td>
<td>0.318075</td>
<td>0.062</td>
<td>0.111</td>
</tr>
</tbody>
</table>

that our numerical scheme is useful to obtain the numerical solutions of other important nonlinear problems in various fields.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References


DOI: 10.4236/jamp.2020.86085 1134 Journal of Applied Mathematics and Physics
Precursive Time, the Hidden Variable

Renè Burri

Free Academy of Modern Physics, Lugano, Switzerland
Email: r.burri@ietclab.org

Abstract

In this paper, a new complex variable defined as “precursive time” able to correlate general relativity (GR) and quantum field theory (QFT) in a single principle was characterized. The thesis was elaborated according to a hypothesis coherent with the “Einstein’s General Theory of Relativity”, making use of a new mathematical-topological variety called “time-space” developed on the properties of the hypersphere and explained mathematically through the quaternion of Hurwitz-Lipschitz algebra. In this publication we pay attention to the interaction between the weak nuclear force theory (EWT) and the nuclear mass of the Standard Model.

Keywords

Time Curvature, Precursive Time, Hidden Variable, Timespace Manifold, Chronotope, Quantum Compensation Spacetime

1. Introduction

In the previously published study (HDTSS 2016), we have exposed the thesis of a plausible “time curvature” induced by spacetime modifications. The hypothesis is coherent with GR which assumes that time curves jointly with space this is by definition.

The plausibility of the “time curvature” finds its theoretical hypothesis in the different classical representations of the “curvature” of spacetime, differently interpreted as a dynamic of a “twisting” process of spacetime.

From a conceptual point of view, the twisting dynamic implies a contraction effect of spacetime; the theoretical difference between “curvature” and “twist (torsion)” is decisive in the analysis of the topological variety of the system. The representation of space-time curvature is specifically described through the ten-

How to cite this paper: Burri, R. (2020) Precursive Time, the Hidden Variable. Journal of Applied Mathematics and Physics, 8, 1135-1154. https://doi.org/10.4236/jamp.2020.86086

Received: March 16, 2020
Accepted: June 19, 2020
Published: June 22, 2020

Copyright © 2020 by author(s) and Scientific Research Publishing Inc. This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access
sor equations as solved by the Einstein Field Equation (EFE), in other words, the contracting space-time places the time variable inside a concept equation:

\[ \text{<minor space} \Leftrightarrow \text{minor time} \Rightarrow \text{less space} \rightarrow \text{less time}> \]

2. Theoretical Synopsis

Theoretically, the time curvature can be described as: the projection of instantaneous time onto an imaginary abscissa shifted in advance towards the future.

However, more concretely we can see it as a “time discrepancy” (\(\Delta \tau\)) that originates between the event’s instantaneous time and its imaginary projection in spacetime. These results in a coherent representation of curved spacetime referred to a generic mass but specified by Planck constants. The topological characteristic of the representation of a time discrepancy or a spacetime-induced time warp was developed by placing as a reference theoretical concept: Minkowski spacetime\(^4\) \([1]-[7]\). The Minkowski manifold highlights the position in Euclidean space of an event recorded in an instant “chronotope” time, however it does not make possible a connection to an imaginary temporal base. In order to operate on the imaginary time variable, the study has developed a specific topological manifold applied to the properties of the Hypersphere 3S Sphere manifold, which made it possible to represent a complex imaginary time related to the frame of reference of instantaneous time. The studied manifold variety, defined as “time-space”, characterizes a pseudo-connection of relative order between the time referred to the Minkowski space, and the projection on an imaginary timeframe of reference.

Time-space manifold resolves the correlation between \(<\text{Minkowski4 space}>\) and the projection of imaginary timeframe of reference elaborated based on the higher dimensions’ geometry of the 3-s Sphere hypersphere, running an entanglement-type correlation. The time-space variety allows us to specify the imaginary part of the time axis leaving the space coordinates invariant, as shown in Figure 1 For the sake of simplicity, the graphical representation of the hypersphere is dashed at a semi-circle.

- \(P_e\): chronotope event
- \(t_o\): time event
- \(C\): pseudo connection
- \(\varepsilon_o\): Imaginary time \((\varepsilon_0 \equiv t_0)\) The reference point \((t_0)\) of chronotope time component expressed by M4 space is correlated and in phase with its imaginary time frame of reference component \((\varepsilon_0)\) when in a theoretical flat space-time condition.

\(^3\)Minkowski4 spacetime provides a logical representation of our Universe considering that there is no event without a reference to a specific instant in time. Any event of ponderable nature is necessarily characterized by a precise and unique set of “space + time" coordinates, called for this reason “chronotope”. The representation of the chronotope on a hyperplane or four-dimensional space establishes the uniqueness of that precise spacetime event. The chronotropic coordinates of an event establish the inviolability defined by the “Pauli exclusion principle”. In fact, in the same place and at the same time there can be just one fermion. The chronotope representation on 4-dimensional space was the synthesis of the study developed by Hermann Minkowski to provide a coherent geometry to relativistic theories.
Figure 1. Correlation time event \( \rightarrow \) projection Imaginary time in flat space \( (e_0) \) and curved spacetime \( (e_0^-) \). In the case of flat space, \( (e_0 \equiv t_0) \) the two components (real and imaginary) coincide.

- B: hypersphere 3-s Sphere \( (e_0) \) is centre of gravity
- D: discrepancy time – time-space curvature \( (\Delta qt) \)
- \( e_0^- \): precursive time \( (t_0 \Leftrightarrow e_0^-) \) Chronotope event “\( P_k \)” with time component \( (e_0^-) \) establishes a new correlation \( (e_0^-) \) becomes the new point of hypersphere centre of gravity.

3. TimeSpace Manifold

The use of the term “timespace”, at variance from spacetime, makes sense because in the thesis we take the parameter time as the independent function and space as the dependent function. In fact, the term “chronotope” \[13\] \[14\] is composed, in the order, by: chronos (time) + topos (space) \( \equiv \) time space.

The “time-space manifold” can be understood as an entanglement-type connection with symbol \( \langle \langle \langle \theta \rangle \rangle \rangle \) between [Minkowski space] and \langle 3-s Sphere hyperspace\rangle, mathematically it is expressed according to the quaternion algebra.

\[ a(x); b(y); c(z), \] and time component: \([\{t_0 \rightarrow e_0^-\}]; \] pure quaternion part:

\[
\begin{pmatrix}
\begin{bmatrix}
-x_0 + t_0 \\
x_0 - x_j \\
y_0 - y_j \\
z_0 - z_k \\
\end{bmatrix}
\end{bmatrix} = \sigma, \ hyperspace coordinate o (\sigma \ sigma-point)
\]

The expression can be described according by Hurwitz-Lipschitz \[15\] \[16\], since the components are unitary quantities:

\[
\begin{bmatrix}
\{-e_0 + t_0,0,0,0\}, \{0,x_0 - x_j,0,0\}, \{0,0,y_0 - y_j,0\}, \{0,0,0,z_0 - z_k\}
\end{bmatrix}
\]
The “precursive time” \( (e_0) \) is the variable that quantifies the “time curvature” effect that is generated between the instant time of the event and the projection on spacetime.

\[
\begin{bmatrix}
-e_0 - i_0, x_0 - x, y_0 - y, z_0 - z
\end{bmatrix}
\]

The concept of time curvature is expressed by the “time-space” complex manifold (Figure 2).

The timespace manifold layout is represented by two hemispheres which respectively indicate: (red) the mass relative to the event and (blue) the hyper-dimensional component temporally shifted with respect to the inertial frame reference of the chronotope.

- \( P_k \): chronotope (event)
- \( t_0 \): real time event
- \( e_0 \): Imaginary projection event time (Hypersphere centre of gravity)
- \( e_0^- \): precursive time event
- \( M_4 \): Minkowski space
- \( 3-\text{Sph} \): hypersphere 3-sphere

The spacetime variation induced by a generic mass (red hemisphere) is quantifiable by its time curvature that is generated between the instantaneous real time \( (e) \) and the correlated complex imaginary time: \( (e_0 \rightarrow e_0^-) \).

The manifold “timespace” allows us to depict the concept of precursive time \( (\Delta r, e_0^-) \). The connection points are respectively: the centre of gravity \( (P_k) \) and the hypersphere centre of gravity (3-Sph).

### 3.1. Time Curvature Conceptual Meaning

The proportionality correlation that is determined between the inertial mass referred to the event and the time curvature that it generates on spacetime allows us to describe the spacetime variations by operating via quantum field theory (QFT). The invariance to the reference system as defined by the Lorentz transformations is preserved by the event’s frame of reference (Figure 3).

Calculations of time curvature \( (\Delta r) \) and precursive time \( (e_0^-) \) as referred to a generic inertial mass.

The theses and concepts from which the main equations were obtained were described in the previous HDTSS publication, [17]; in this study we make some improvements in the calculation and further analyze and evaluate the procedure. We will examine a fundamental correlation that occurs between: <nuclear mass - weak nuclear force-orbital levels>.

### 3.2. Time Curvature Equation

The time-curvature \( (\Delta qf) \) is calculated in Planck time.

The Equation (1) allows us to determine time-space curvatures \( (\Delta r) \) of a generic inertial mass:
Figure 2. Timespace manifold depiction: The connection points are respectively: the centre of gravity the mass event and the centre of gravity the hypersphere 3-Sph (red hemisphere). The "precursive" time discrepancy ($\Delta \tau$) specifies exactly he inertial mass value (blue hemisphere). Notice that the instantaneous time imaginary projection undergoes a shift as a function of the inertial mass.

Figure 3. In this figure, we distinguish the event on the arrow of time as in the classic scalar representation and on the imaginary abscissa the related "conjugate time" (-sec.).

$$\begin{align*}
M (P_t) &= \frac{h}{t_p c^2} \rightarrow t (P_t) = \frac{h}{M (P_t) c^2} \rightarrow \frac{1}{h} (s)^{-1} \Delta \tau (\text{joules/sec}) \quad (1)
\end{align*}$$

- $t_p = \sqrt{\frac{hG}{c^3}}$ Planck time
- $M (P_k)$ = Planck mass event
- $c^2 = \text{Speed of Light Squared}$

From the Equation (1) we get the value of the precursive time ($\epsilon_{0\tau}$) (-seconds) referred to the instant time inertial frame of the reference event.

3.3. Precursive Time Meaning

$$\tau_{\epsilon} = \frac{1}{h} \text{ joules/sec}$$

As Figure 4 shows, Equation (1) determines the amount of deflected time in-
duced by the time curvature of the inertial mass of the event; we can distinguish:

- $(\Delta \tau)$; time curvature $(\Delta \tau)$ indicates the deflection effect of the real time coordinate induced by the time curvature.
- $(\tau_0)$; precursive time: the precursive time is specifically the amount of time deflected by the time curvature. We remember that the time deflection effect is dictated by the equation:

$$<\text{minor space} \Leftrightarrow \text{minor time}> \Rightarrow <\text{less space} \rightarrow \text{less time}>.$$ From a relativistic point of view, the dynamics (torsion $\rightarrow$ contraction) of spacetime simultaneously determines a deflation effect of space (less space) together with a deflection of time (less time):

$$|\text{time deflection} \rightarrow \text{space deflation}| \Rightarrow \text{spacetime variation}$$

Consequently, the time curvature becomes the primary cause of the spacetime curvature according general theory of relativity (Figure 5).

It is important to note that $(\tau_0)$ expressed in Planck time (joule/sec) coincides with the amount of energy of the inertial mass:

$$\text{energy (} \tau_0 \text{) = inertial mass} \Rightarrow (\Delta \tau) = M_i$$

The equation states that the inertial mass is induced by the time curvature; more specifically we can state that the inertial mass is exactly the value the time curvature.

Figure 4. Equation (1) specifies the time curvature generated by the event in spacetime. The dynamic is centripetal, because it precursive, whereby involves from the future $(e_{\nu})$ toward the time event $(t_0)$.

Figure 5. Depiction as “spacetime curvature” and “spacetime deflation”. The depiction of spacetime in twist dynamics allows us to act on hyper-dimensional coordinate working with quantum correlations.
Two other fundamental parameters are acquired from Equation (1):

1) \((-\sigma)\) time deflection in second (Planck time in precursive second).

2) \((\sigma)\) space deflation. It coincides with the spacetime curvature of the GR (Figure 5).

Equation (1) shows that spacetime variations can be interpreted as classically described by GR through quantum physics.

Equation (1) also confirms that:

\(<\text{For each event with [mass > 0] there is an imaginary time component proportional to the value of the inertial mass}.\>\)

Equation (2) extends the concept of complex variable mass in which the inertial frame of reference is a function of the time-space curvature \((\Delta \tau)\), the equality \((mi = \Delta \tau)\) indicates that:

The inertial mass in the absence of interacting gravitational forces is a physical quantity determined by the time curvature.

Note:

Conceptually, the hypothesis of time curvature deals with two important inexplicable issues that cannot be demonstrated through the classic depiction of space-curved GR:

1) It unifies general relativity and quantum field theory.
2) It unifies inertial and gravitational mass phenomena.

3.4. Time Space Quadrant as a Hyper-Dimensional Manifold

The representation in TimeSpace quadrant allows us to get a photograph of the event in instant time, as a “snapshot” of the imaginary timespace curvature \((\sigma_{\tau})\). The coordinates \((x, y, z)\) can be resolved as “hyperdimensional” since the complementary reference base \((e_0 \rightarrow e_0)\) is placed on an abscissa dimensionally shifted with respect to the event (instantaneous time) (Figure 6).

\(\text{Figure 6. Representation in quadrant depiction; this allows us to have a comprehensive picture of the coordinates as referred to the event.}\)
• Coordinates \( (x_i, y_j, z_k) \) define the “curvature” of the deflated space.
• Sigma point \( (\sigma_\omega) \) which indicates the point of deflated edge space (is the concrete physical space even if calculated with imaginary coordinates).
• \( (e_{\omega_0}) \) is a deflection Imaginary time coordinate assuming the local observables reference frame, induced by the inertial mass \( (P_k) \) on time space (time curvature) \( (e_{\omega_0}) \) is positioned in advance (precursive) with respect to \( (t_0) \), because <less space → less time>.

3.5. Time Curvature Coordinates

From Timespace manifold:

\[
\forall [P_i \in \mathbb{R} (x_0, y_0, z_0) (t_0)] \subseteq \mathbb{C} (x_i, y_j, z_k, e_{\omega_0})
\]

\[
M4 \mathbb{R} (x_0, y_0, z_0, t_0) ; 3 \& \mathbb{C} (x_i, y_j, z_k, e_{\omega_0})
\]

\[
P_k (t_0) (x_0, y_0, z_0) ; P'_k (e_{\omega_0}) (x_i, y_j, z_k)
\]

Summing up, each event is characterized by “seven” coordinates:

\[
P_k (t_0) (x_0, y_0, z_0) ; P'_k (e_{\omega_0}) (\sigma_\omega)
\]

real coordinates referring to the chronotropic position of the M4 event:

\[
P_k : \mathbb{R} (x_0, y_0, z_0) (t_0)
\]

and four 3S hyperdimensional coordinates:

\[
\sigma_\omega : \mathbb{C} (x_i, y_j, z_k) (e_{\omega_0})
\]

where, \( (t_0) \) is the event time of inertial frame of reference.

3.6. Calculation of Inertial Mass as a Function of Spacetime

From the analyses carried out in this study, we can assume that matter and spacetime are in fact a single entity consisting of two joint parts, real part and spacetime the complex hyperdimensional joint part, the interaction <inertial mass ⇔ spacetime> assumes the concept of a bivalent set. Ref. timespace manifold (Figure 2).

From Equation (1), we can infer correlation (3) that introduces a concept of mass expressed in Planck parameters, directly related to the spacetime interaction properties.

It is a new expression of mass calculation as a function of the precursive time, that is, time-space curvatures, a new concept of mass that we distinguish with the name “mass-matter” \( (\psi) \).

Equation (3): calculation of inertial mass according to the precursive time equation:

\[
\tau_\psi = \frac{1}{h} \rightarrow \psi = \frac{m_p \tau_\psi}{\Gamma_p K} \pi^F
\]

\( \psi \) : mass-matter (inertial mass as function of time curvature)
- $m_{\psi}$: chronotope mass ($P_{i}$) in Planck mass
- $\tau_{p}$: precursive time
- $\Gamma_{p}$: quantum specific Heat capacity
- $K$: Boltzmann Constant
- $\pi_{F}$: quantum compensation spacetime, (F) Fibonacci coefficient [18] [19] [20] [21].

The equation of the calculation of the inertial mass as a function of the time-space curvature implies a direct correlation to Quantum Heat capacity ($\Gamma_{p}$):

**Q. Heat capacity:**

$$\Gamma_{p} = \frac{\tau_{p}}{K} C^2$$  \hspace{1cm} (4)

The equation expressed with ($\tau_{p}$) is calculated as Planck temperature referred to ($\tau_{p}$).

Equation (3) highlights that: $\psi f(\Gamma_{p} \rightarrow \tau_{p})$

this leads to asserting that ($\Gamma_{p}$) is a state property of matter.

From the calculation of Q.Heat capacity it is pointed out that is a characteristic of the mass as a function of the time curvature; in fact the correlation provides us with a reading key to the interpretation of the rebus of absolute zero: $\psi \rightarrow \Gamma_{p} > 0 \hspace{1cm} K$

From the correlations expressed, we can say that: a generic mass $|m_{i} > 0|$ has a heat state of Q. Heat capacity determined by its time curvature. This assumption implies that it cannot degrade below the Absolute Zero (0 K).

3.7. Quantum Compensation Spacetime ($\pi_{F}$)

(F) Fibonacci coefficient

Let’s highlight the remarkable coefficient ($\pi_{F}$) obtained from theoretical processing of spacetime twist dynamics: quantum compensation spacetime ($q.c.s.$). This coefficient allows us to obtain a precise measure of the value of the inertial mass as a function of the precursive time. The representation of spacetime calculated as a function of the Fibonacci exponential faithfully describes the variation of spacetime in twist (torsion) dynamics with an accuracy of the time curvatures of ($10^{-15}$), as shown below in Figure 8 and Figure 9 with regard to nuclear masses. The constant of “quantum compensation spacetime”, describes the development layout of the twist-contraction dynamics of spacetime (Figure 7).

Given the coordinates: ($y_{i}(c_{i})$; $x_{i}(c_{i})$) the graphic elaboration model draws a centripetal cycloid where the space-time (eigenvectors) contracts towards the centre of gravity of the chronotope. The four segments ($F$) represent equipotential space-time areas with a different frame of reference.

It is important to underline that the different areas or equipotential regions ($F_{1}, F_{2}, F_{3}, F_{4}$) are also iso-temporal, this being in accordance with Equations (1) and (2). The summation of Fibonacci’s hyperdimensional regions: complex time + complex space, is equivalent to the value of the inertial mass as referred to the chronotope: ($F_{n}$): $\tau_{x_{1}} + \tau_{x_{2}} + \tau_{x_{3}} + \tau_{x_{4}} = \psi$
Figure 7. Left image: computer processing of space-time torsion contraction, Coordinates \((Y_j \text{ and } X_i)\) draw a centripetal cycloid with a Fibonacci progression \((\pi')\) (Spiral sinks complex eigenvalues graphics processing). Right image: comparison with a spiral generated by an electron in a bubble chamber - Photo: Harvard project, Elementary Particles [22].

The hypothesis under study seems to draw an important meaning as a physical consequence, that is: mass and spacetime are not two different interacting entities as we would interpret classically, but we can state that:

<inertial mass and spacetime are constituents of a single entity in which precursive time is the mediator>.

3.8. Energy Equation

The energy equation \(\psi\) function of time curvature, becomes a concept of mass peculiarly related to spacetime interaction. As a result of this, the equation: \((E = mc^2)\), can be reexamined in:

\[
E = \psi c^3
\]

Equation (5) which confirms the thesis, in fact it can take over from Einstein’s equation.

\[
E = mc^2 \equiv \psi c^3
\]

From the exposed correlations [matter \(\rightarrow\) precursive time \(\rightarrow\) energy], we can support the principle stating that: [matter \(\Leftrightarrow\) spacetime] uniquely phenomena.

4. Unified General Relativity \(\Leftrightarrow\) Quantum Field Theory

In this unpublished second part of the publication we submit the working hypothesis to the test of coherence applied to the nuclear masses. The theses that are here developed involve an atypical research approach, as general relativity (GR) and quantum field theory (QFT):

4.1. Correlation: Timespace Curvature \(\Leftrightarrow\) Nuclearenergy

Figure 8 shows an example of application of Equations (1)-(3)-(5) to the nuclear masses of the following atomic elements.

Figure 8 confirms that the correlations are consistent, in fact the value of the energy calculated in GeV is correct.
The value of the energy GeV confirms: \[ E = mc^2 \equiv \psi c^3 \]
- Column (\(\Delta \tau\) J/sec) indicates the precursive time (time curvature) calculated from Equation (1).
- Column (\(\Psi\)) indicates “mass-matter” calculated from Equation (3).
- The energy calculation (GeV) is obtained from Equation (5).

**Figure 8** confirms the thesis that:
1. We can specify spacetime in quantum physics.
2. Spacetime can also be calculated for atomic nuclei.

In **Figure 9**, it is possible to quantify the spacetime variation even for nuclear masses. We apply Equations (1)-(5) to the nuclear masses of the following atomic elements:

### 4.2. Correlation: Timespace Curvature ⇔ Electron Energy Levels

The column (picometers) is calculated:
- \((\Delta t): \tau, \Rightarrow Ptc^2 = \text{orbital radius}\)
- nuclear mass \(\psi c^3 = \text{GeV}\)

**Figure 9** correlates time curvatures (\(\Delta \tau\)) and the distance (\(\sigma\)) in picometers.

We can note that time curvatures generated by the nucleus’ mass determines the distance of the electronic orbitals’ levels (\(\sigma\)).

<table>
<thead>
<tr>
<th>elements</th>
<th>un.atomic mass (u)</th>
<th>Mass (Kg)</th>
<th>plank mass</th>
<th>(\Delta t) (joules/sec)</th>
<th>(\psi) mass-matter</th>
<th>(\psi C3)</th>
<th>GeV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hydrogen</td>
<td>1,0379400E+00</td>
<td>1,6726200E-27</td>
<td>7,6900000E-20</td>
<td>3,228974E+15</td>
<td>5,57933E-36</td>
<td>1,50322E-10</td>
<td>6,39822E+20</td>
</tr>
<tr>
<td>Li</td>
<td>6,9400000E+00</td>
<td>1,1500000E-26</td>
<td>5,2950000E-19</td>
<td>8,472945E+15</td>
<td>3,8414E-35</td>
<td>1,03503E-09</td>
<td>6,46012E+10</td>
</tr>
<tr>
<td>Na</td>
<td>2,298979E+01</td>
<td>3,1875409E-26</td>
<td>1,7540000E-18</td>
<td>1,542113E+16</td>
<td>1,27249E-34</td>
<td>3,42859E-09</td>
<td>21,3999E+02</td>
</tr>
<tr>
<td>K</td>
<td>3,999830E+01</td>
<td>6,4924300E-26</td>
<td>2,9830000E-18</td>
<td>2,011074E+16</td>
<td>2,1641E-34</td>
<td>5,83094E-09</td>
<td>36,3938E+02</td>
</tr>
<tr>
<td>Rb</td>
<td>8,546780E+01</td>
<td>1,4192000E-25</td>
<td>6,5210000E-18</td>
<td>2,973435E+16</td>
<td>4,73083E-34</td>
<td>1,27468E+08</td>
<td>79,5590E+01</td>
</tr>
<tr>
<td>Cs</td>
<td>1,3290545E+02</td>
<td>2,2609470E-25</td>
<td>1,0140000E-17</td>
<td>3,707350E+16</td>
<td>7,35633E-34</td>
<td>1,98209E-08</td>
<td>123,712E+02</td>
</tr>
<tr>
<td>Fr</td>
<td>2,2300000E+02</td>
<td>3,7030000E-25</td>
<td>1,7010000E-17</td>
<td>4,802347E+16</td>
<td>1,23404E-33</td>
<td>3,24998E-08</td>
<td>207,529E+03</td>
</tr>
</tbody>
</table>

**Figure 8**. This figure shows how the time curvatures equation can determine the energy value of the relevant respective nuclear. The calculated values are correct and confirm our thesis.

<table>
<thead>
<tr>
<th>elements</th>
<th>(\Delta t) (joules/sec)</th>
<th>(PlanckTime) Pt m^2/sec</th>
<th>Pt m^2/sec</th>
<th>(dm)Picosimeters</th>
<th>GeV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hydrogen</td>
<td>3,22897E+15</td>
<td>1,741E-28</td>
<td>1,56473E-11</td>
<td>15,64733</td>
<td>0,938282E+24</td>
</tr>
<tr>
<td>Li</td>
<td>8,47295E+15</td>
<td>4,568E-28</td>
<td>4,10551E-11</td>
<td>41,05513E-06</td>
<td>6,460129</td>
</tr>
<tr>
<td>Na</td>
<td>1,54211E+16</td>
<td>8,314E-28</td>
<td>7,47225E-11</td>
<td>34,72350E-06</td>
<td>21,3999E+02</td>
</tr>
<tr>
<td>K</td>
<td>2,01107E+16</td>
<td>1,084E-27</td>
<td>9,74251E-11</td>
<td>97,42506E-13</td>
<td>36,3938E+02</td>
</tr>
<tr>
<td>Rb</td>
<td>2,97344E+16</td>
<td>1,603E-27</td>
<td>1,4407E-10</td>
<td>144,0704E-02</td>
<td>79,5590E+01</td>
</tr>
<tr>
<td>Cs</td>
<td>3,70735E+16</td>
<td>1,999E-27</td>
<td>1,79661E-10</td>
<td>179,6611E-02</td>
<td>123,712E+02</td>
</tr>
<tr>
<td>Fr</td>
<td>4,80234E+16</td>
<td>2,589E-27</td>
<td>2,32688E-10</td>
<td>232,6877E-02</td>
<td>207,529E+03</td>
</tr>
</tbody>
</table>

**Figure 9**. This figure calculates the time curvature (\(\Delta \tau\)) generated by the mass of the nucleus. The space gap in picometers is exactly the electron distance from the nucleus according to its energy level.
5. The Electromagnetic, Weak, and Strong Interactions

5.1. Correlation: Nuclear Mass ⇔ Nuclear Electronic Orbital Levels

In the following diagram we have highlighted the correlation ratio determined between the orbitals i.e. Na -sodium (2, 8, 1) and the respective distance from the nucleus. The energy contained in orbitals (2, 8.1) is directly related to the distance between the nucleus and the orbital radius.

– Values in red refer to the energies ((\(\Delta \tau\) J/sec)) of the different orbital levels.
– Values in blue refer to the positions (Pm) in picometers attributed to orbitals.

The orbital takes on an energy value as a function of the distance determined from the nucleus.

From the Equation (1) \((\tau_v)\) it is possible to calculate the levels of the orbitals and the number of electrons between orbitals according to the distance from the nucleus.

**Figure 10** shows the position of the three sodium levels and the overall energy equal to the number of electrons on that level:

5.2. Correlation: \(\Delta \tau \rightarrow \Gamma_p\) Exothermic ↔ Endothermic Q Value

The movement of electrons between orbital levels in emission or absorption form is indeed a variation of the orbital radius, the correlation is:

\[
[r_v] \rightarrow (\sigma_{\mu}) \rightarrow \psi c^3
\]

Each level is characterized by a space position coordinate \((\sigma_{\mu})\) (**Figure 11**) and a corresponding amount of energy as a function of precursive time \((\tau_v)\).

The complexity of the system is due to the double correlation that specifies the coordinate set \((\mathbb{R} + \mathbb{C})\) for each Orbital level \((K, L, M, N)\):

- \(\Sigma K : C(x_{ki}, y_{ki}, z_{ki}) (\tau_{v_{ki}}) \rightarrow \psi_{ki} c^3\)
- \(\Sigma L : C(x_{li}, y_{li}, z_{li}) (\tau_{v_{li}}) \rightarrow \psi_{li} c^3\)
- \(\Sigma M : C(x_{mi}, y_{mi}, z_{mi}) (\tau_{v_{mi}}) \rightarrow \psi_{mi} c^3\)
- \(\Sigma N : C(x_{ni}, y_{ni}, z_{ni}) (\tau_{v_{ni}}) \rightarrow \psi_{ni} c^3\)

Nucleus:

- chronotope nucleus position (nucleus coordinate): \(P_i : \mathbb{R}(x_0, y_0, z_0)(t_0)\)
- Nucleus energy: \((\Delta q \tau)c^3\)
- Atom cross sections: \(\sigma_{\mu} : C(x_{r}, y_{r}, z_{r}))(e_{\mu})\)
- orbital (N) stretch dynamic: \(\frac{1}{n_{\psi}} = \tau_{N1} - \tau_{N2}\)

\[
\frac{me^4}{8\varepsilon_0^2 h^4} \tau_{N2} \rightarrow \Delta \tau (-\Gamma_p) \rightarrow \text{orbital contraction (Figure 12).}
\]

\[
\frac{me^4}{8\varepsilon_0^2 h^4} \tau_{N1} \rightarrow \Delta \tau (+\Gamma_p) \rightarrow \text{orbital stretch (Figure 12).}
\]

In summary, each orbital consists of a space position \((\sigma_{\mu})\) (hyperspace posi-
tioning) and a time quantum level ($\tau_v$) which determines its energy (Figure 11). The correlation is extremely significant, as it places the nucleus mass as the cause of the energies distributed over the orbital levels (Figure 12 general diagram).

**Figure 10.** Orbital sodium levels obtained from a calculation model from Equation (1).

![Orbital sodium levels](image)

**Figure 11.** The correlation ratio between nucleus mass and orbital radius defines the electroweak force.

![Correlation ratio](image)

**Figure 12.** General diagram: The dynamic of the <exothermic ⇔ endothermic> reaction can be very similarly imagined to a (stretching dynamic). *Slide show 2019 LANR/CF Colloquium at MIT, March 23-24, 2019. Cambridge MA 02139 (USA).*

--

**Table 1:**

<table>
<thead>
<tr>
<th>τp (proton)</th>
<th>τv (electron)</th>
<th>Lev H</th>
<th>(1st gap) Gap1</th>
<th>Lev 2</th>
<th>Gap 2</th>
<th>Lev 3</th>
<th>σ(ω)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.227924E+15</td>
<td>7.53426E+13</td>
<td>15.6383</td>
<td>0.3653</td>
<td>2</td>
<td>8</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>21.3996 GeV</td>
<td>0.5104 MeV</td>
<td>0.0020</td>
<td>3.2025</td>
<td>4.0833</td>
<td>5.1041</td>
<td>0.5104</td>
<td>18.2915</td>
</tr>
</tbody>
</table>

**Table 2:**

<table>
<thead>
<tr>
<th>μ</th>
<th>15,6383</th>
<th>Pm</th>
<th>0.3653</th>
<th>0.7301</th>
<th>2.1899</th>
<th>2.9205</th>
<th>3.6507</th>
<th>0.3650</th>
<th>10,3213</th>
</tr>
</thead>
<tbody>
<tr>
<td>σ(ω)</td>
<td>7.53426E+13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>21,3996 GeV</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

DOI: 10.4236/jamp.2020.86086 1147 Journal of Applied Mathematics and Physics
6. Theoretical Fallout

6.1. Extending the Standard Model in Hyper-Dimensional Mechanics

As shown on the chart in Figure 11, the electrons are emitted from quantum time precursive orbital, with respect to the nucleus chronotypical position, (the frame of reference of the nucleus is delayed compared electronic orbitals), it follows that the electron does not have a specific position “x” and momentum “p”, but is in a weak interactions’ regime in the latency energy form of.

The deductive conclusions suggest an understanding of the structure of a four-dimensional atom much more complex than the one presented by the current “standard model”.

The concepts presented so far are coherent with the standard model, Bohr model representation models hitherto analyzed, however the electroweak interaction dynamics defined by the new variable suggest a mathematical structure which goes under the name the Hopf fibration \[23\] \[24\] \[25\].

Hopf’s fibrational dynamics configuration allows the standard model to be extended to hyperspace correlations between nucleus and electronic orbitals. Fibrational mechanics allow a consistent representation of the theory and allow to calculate the energies of the electronic orbitals (Figure 13 referred to Figure 10)

The fibrational dynamics explain the reason why there is no conflict between the negative charge of the electron and the positive charge of the nucleus, the system to continue in balance because it is the nucleus itself the one generating the spacetime variations that give rise to the electron charge of the electro-weak interaction.

- fibration positioning \((s^1 \rightarrow s^2 \rightarrow s^3)\).
- energy levels \((ψ1 \rightarrow ψ3 \rightarrow ψ2)\).

The proposed atom fibration depiction dynamics does not modify the energy levels ascertained by the standard model, it varies the rationality of the origin causality of the Electroweak interactions.

As regards Quantum Chromodynamics (QCD) theory of the strong interaction between quarks and gluons, a review the interpretations in this new key of reading necessarily arises.

The hypothesis studying a “hyperdimensional atom” explains the reason for the incomprehensible conflict “The Wave-particle duality relation” and confirms Louis de Broglie’s Research on Quantum Theory hypothesized that all matter can be represented as a de Broglie wave in the manner of light.

6.2. Heisenberg Uncertainty Principle, Why?

The uncertainty principle is basically evincible through the electron levels in fibrational dynamic, since it is generally not possible to predict the value of a quantity with absolute certainty. In fact, any measurement test induces the parameters \((X \text{ position and } P \text{ momentum})\) change in \((σ_P)\) of the fibers, just as demonstrated by Heisenberg’s uncertainty principle.
6.3. How Is Gravitational Force Generated and Why Is It Always Attractive?

We have explained the concept of time curvarture as an effect of “temporal discrepancy” originating between the chronotropic mass referred to the event and the spacetime indicated in Planck time defined as precursive time.

The Equations (1) and (2) indicate that the mass and time curvature are proportional. This leads us to infer that time curvature is primarily responsible for gravitational mass.

In fact, Equation (6) confirms the thesis:

\[ g = G \frac{m_1}{r^2} \Rightarrow g = G \frac{\psi}{r \Delta \tau} \]  \hspace{1cm} (6)

- \( g \): gravity.
- \( r \): radius.
- \( G \): gravitational constant.
- \( \Delta \tau \): time-space curvature.

The reason why the inertial mass and gravitational mass are related is because both forces originate from the same causal event.

The direct correlation exists between inertial and gravitational mass, so the gravitational mass value acquires an objective value defined Gravitation as a <specific attraction degree \( A_0 \)>.

\[ A_0 = \Delta_{\tau r} = g \]

Time curvature defines the value of the specific attraction capacity degree of a body \( A_0 \).

The two masses attract each other \( g = (FA_01 - FA_02) \) so the result is the difference between the two sections of time curvature (Interferential of the Fibonacci areas).

Brief note about it.

In gravitational interactions, areas \( F \) are consistently added into quaternion matrix as described according to Hurwitz-Lipschitz, since the components are unitary quantities.
However, in this study we limit the exposure to the gravitational mass equation or gravitational attraction force, understood as a specific attraction capacity of a body as a function of the time curvature (space deflated effect). About the interactions between bodies, that is, between gravitational masses, the theory extends the concept to the principle of interference, resulting from the mutual action of the precursive areas \((e^0)\) (Figure 14).

7. Discussion

Compared to the first publication of the theory, we collected reviews from many colleagues who highlighted the difficulty in understanding exactly the sense of time curvature and precursive time. The difficulty lies in acquiring a new knowledge of “binary” matter including of a hyper-dimensional part. As explained, the study basically demonstrates that matter and spacetime are two constituents of a single entity. The same principle is applicable on a nuclear scale by establishing a correlation at the base of the electroweak force, between the nucleus and the electronic orbitals, so even the Higgs boson, can be considered as the minimum time discrepancy amount that can give rise to a ponderable unit of mass through the LHC.

Some colleagues asked us if the theory went towards a GUT hypothesis or if ToE, given that in the same work it is exposed to gravitational force and electroweak force. However, the fact that the publication exposes aspects of general relativistic physics, and nuclear physics in the same study, confirms the hypothesis of a Theory of ToE Unification. This made it difficult even to characterize the physics area for the publication of this study.

As a final point, we can ask ourselves whether there are any scientific clues leading to acceptance of this hypothesis.

Regarding this difficult question we are sure that the theoretical research was born from the need to identify a scientific hypothesis consistent with the theories and phenomenologies gained, to provide an interpretation of the inexplicability/conflictuality of the current physics. To this aim, we can say that until now neither evidence nor clues have emerged, that exclude it from a valid research hypothesis.

Figure 14. \((e^0)\): precursive time events.
8. Conclusions

In this study we have made a subsequent deepening of the theoretical hypothesis of temporal curvature as the basic phenomenon of the interaction between matter and spacetime. The study summarized the electroweak interaction between the nucleus and orbital levels, applying the same equations describing spacetime and gravitation.

The study pointed out the following demonstrations:

1) CORRELATION: TIMESPACE CURVATURE ⇔ NUCLEAR ENERGY
2) CORRELATION: TIMESPACE CURVATURE ⇔ ELECTRON ENERGY LEVELS
3) CORRELATION: NUCLEAR MASS ⇔ NUCLEAR ELECTRON ORBITAL ENERGY LEVELS
4) CORRELATION: ∆τ → Γp Exothermic ↔ endothermic Q Value

Mathematical analyses have been demonstrated with Equations: (1) (3)-(6).

From Equation (1) we calculate:
1) Time curvature (Planck time)
2) Precursive time (joule/second) inertial mass
3) Timespace deflation (σ) (second * c^2) gravitational mass
4) Nuclear Electronic orbital (Planck time * c^2) - Electroweak interaction.

From Equation (3) we calculate:
Mass-matter or “hypermatter”: calculation of the mass as a function of the induced space-time curvature.

Equation (3) modifies the concept of mass intended as a scalar quantity, replaces the concept of quantitative aggregate as a derivative of a function composite. The symbol (ψ) indicates the complex mass.

The equation introduces two new relationships:
1) Quantum specific heat capacity (QSHC): Planck temperature as a function of time curvature.
2) QCS (π^2) coefficient

The processing of the graphical representation models has characterized the QCS (π^2) coefficient, allowing to reproduce the centripetal dynamics moment of the twisting effect of spacetime.

From Equation (5) we calculate:
energy equation:  \( E = mc^2 \equiv ψc^3 \)

The ability to calculate the energy (\( E = mc^2 \)) through the hypermatter calculation equation (ψ), confirms consistency.

The study has been carried out according to a pragmatic theoretical development:

<Idea → hypothesis → thesis → theory>

The following stages were developed: conceptual analysis; topological and mathematical conceptual connections; analysis of phenomenological coherence with classical theories.

Key concepts elaborated:
• Timespace correlative manifold variety
• Quaternion algebra; hyperbolic geometry - hypersphere - Minkowski spacetime
• Quadrant Complex depiction (complex time + complex space)
• Quantum compensation spacetime \( \pi F \); \( F \) Fibonacci coefficient
• Hopf fibration mathematical structure

The study was subject to “Theoretical Stress Test”:

• Inconsistencies
• Theoretical inconsistencies
• Phenomenological inconsistencies
• Calculation inconsistencies

**Theoretical Definitions**

In the study unpublished definitions were suggested to make exposure understandable. However, the study proposes them as investigational indicators, suggested to the scientific community for all changes that the research will consider necessary:

*Time curvature.* It is understood as the discrepancy time that occurs between inertial frame of reference and spacetime. The theoretical research that has been carried out has shown that the spacetime curvature is a phenomenon resulting from the complex interaction between the three exclusive ingredients that make up our universe: matter-space-time. The notion makes it possible to combine and unify quantum physics with classical and relativistic theories.

*Precursive time.* defines the value of the time curvature in quantities of Planck time or \((-\text{sec})\) of. It is the variable that quantifies the temporal curvature that the mass induces on spacetime. It can be considered as the physical mediator between matter and spacetime.

*Space deflation - Time deflection.* In the twist representation, the contraction of space (deflation) is related to a reduction of time (deflection). In summary, it takes less time to travel less space.

*Chronotope.* Indicates the spatial position of an event at a specific instant time. The chronotope establishes a condition of exclusivity, as there can be no overlap of events having the same space and time coordinates

*Mass-matter.* The equation with symbol \( \Psi \), enshrines a concept of complex mass.

*Timespace manifold.* The topological variety of relationships between Minkowski space and the projection of the complex time variable. The variety was developed to represent in a comprehensible form the points of conjunction between the real parameters of the event and its imaginary projection. The connection is hypothesized according to the principle of entanglement correlation.

*Timespace Quadrant.* Representation chart of all real and complex components of the event at the time instant.

*Q. Heat capacity.* Planck temperature related to time curvature. It is a parameter for the calculation of mass-matter. It expresses a thermal characteristic of
matter \((m_i > 0)\) as a function of spacetime.

*Q.c.s. Quantum compensation spacetime \((\pi^i)\):* coefficient for the calculation of the mass-matter. It determines the angular momentum of spacetime.

\((A_0)\) specific attraction degree: The study shows that inertial and gravitational mass derive from the same cause described as time curvature. It follows that a body’s gravitational attraction force can be simply resolved as a specific attraction degree.

**Acknowledgements**

I wish to thank all my trusted and intangible collaborators for all the support that they have poured into me during these years of study, starting with: Hermann Minkowski who despite his extraordinary synthesis of spacetime in the science of a thousand inexplicables is still somewhat unrecognized. To Albert who is always willing to whisper advice to me, to David Bohm whom I consider my putative father. I thank all the other collaborators for what they have given me their knowledge: Max Planck, W. R. Hamilton, Ludwig Boltzmann, J. H. Poincaré, J. Krishnamurti, Max Born. Last but not least to Giordano Bruno who feed the flame of my knowledge.

**License**

The publication is licensed under Academic Free License (“AFL”) v. 3.0 a Creative Commons Attribution 4.0 International License.

**Conflicts of Interest**

The authors declare no conflicts of interest regarding the publication of this paper.

**References**


Steady State Gas Flow in Pipeline Networks: Existence and Uniqueness of Solution

Agegnehu Atena, Wondimu Tekalign, Tilahun Muche

Department of Mathematics, Savannah State University, Savannah, GA, USA
Email: aatena@savannahstate.edu, tekalignw@savannahstate.edu, muche@savannahstate.edu


Received: May 9, 2020
Accepted: June 19, 2020
Published: June 22, 2020

Abstract
In this paper we discuss the uniqueness and existence of solution to a real gas flow network by employing graph theory. A directed graph is an efficient way to represent a gas network. We consider steady state real gas flow network that includes pipelines, compressors, and the connectors. The pipelines and compressors are represented as edges of the graph and the interconnecting points are represented as nodes of the graph representing the network. We show that a unique solution of such a system exists. We use monotonicity property of a mapping to proof uniqueness, and the contraction mapping theorem is used to prove existence.

Keywords
Gas Flow Network, Uniqueness, Existence

1. Introduction
In this paper we investigate the existence and uniqueness of solution to a real gas flow pipeline network by representing the network by graph. A directed graph is an efficient way to represent a gas network. The pipelines and compressors are represented as edges of the graph and the interconnecting points are represented as nodes of the graph representing the network. In a directed graph representation of a gas network, each edge has been assigned a flow direction. Existence and Uniqueness of Solutions to the Generalized Riemann Problem for Isentropic Flow is discussed in [1]. In that paper a gas flow in networks of pipelines is considered. Their models are based on the generalized Riemann problem formulation, where the flow in each connected pipe section is described by the hyperbolic conservation law supplemented by initial conditions within each section of the flow network. In our case, we consider a steady state and isothermal flow in the pipelines, which enables us to simplify the three conservation laws (mass,
momentum, and energy) into a single equation for each pipeline, see (3.1). The flow equation of the gas in a compressor is given in (3.2). As we are employing graph theory to study the gas flow network, reviews of some preliminary concepts from graph theory are presented in the first section of this paper. In the second section we present the mathematical model for steady state real gas flow in a pipeline network. Section three contains the discussion about existence and uniqueness of solution for a real gas flow network.

2. Preliminary

Graph Theory

A graph \( G = (V, E) \) is a structure consisting of a finite set \( V \) of elements called vertices or nodes and a set \( E \) whose elements are called edges [2]. A directed graph or digraph is a graph in which each edge has a direction from one node to another. A walk of a graph \( G \) is defined as a finite alternating sequence of vertices and edges, beginning and ending with vertices, such that each edge is incident with the two vertices immediately preceding and following it. A walk in which no vertex appears more than once is called a path. A path beginning and ending with the same vertex is called a cycle.

A graph \( G \) is said to connected if there is at least one path between every pair of vertices in \( G \). A tree is a connected graph with no cycles. A spanning tree \( T \) of \( G \) is a tree consisting of all vertices in \( G \). For a given spanning tree \( T \) of \( G \), any edge in \( G \) which is not in the tree \( T \) is called a chord. A basic result from graph theory states that adding a chord to a spanning tree \( T \) will create exactly one cycle. Such a cycle formed by adding a chord to a spanning tree is called a fundamental cycle.

**Theorem 1:** A tree with \( n \) vertices has \( n - 1 \) edges.

**Proof:** The theorem will be proved by induction on the number of vertices. It is easy to see that the theorem is true for \( n = 1, 2, 3 \). Assume that the theorem holds true for all trees with fewer than \( n \) vertices. Let us now consider a tree \( T \) with \( n \) vertices. In \( T \) let \( e_k \) be an edge with end vertices \( v_i \) and \( v_j \). Since there is one and only one path between every pair of vertices in a tree, there is no other path between \( v_i \) and \( v_j \) other than \( e_k \). Therefore, deletion of \( e_k \) from \( T \) will disconnect the graph. Furthermore, \( T - e_k \) consists of exactly two components, and there are no cycles in \( T \), each of these components is a tree. Both of these trees have fewer than \( n \) vertices each, and therefore, by the induction assumption, each contains one less edge than the vertices in it. Thus \( T - e_k \) has \( n - 2 \) edges. Hence \( T \) has exactly \( n - 1 \) vertices.

**Theorem 2:** Let \( n \) and \( e \) be the number of vertices and edges, respectively, in a connected graph \( G \). Let \( T \) be a spanning tree of \( G \). Then

1) The number of edges in \( T \) is \( n - 1 \) and the number of chords corresponding to the spanning tree \( T \) is \( e - n + 1 \) chords.

2) The number of the fundamental cycles corresponding to the spanning tree \( T \) is exactly the number of chords, \( e - n + 1 \).
**Proof:**

1) Since a spanning tree is a tree, theorem 1 can be used to show the first part of 1). Since a chord is an edge in $G$ which is not in the given spanning tree $T$, the number of chords corresponding to $T$ is the total number of edges $e$ in the graph $G$ minus those $n-1$ edges which belong to the spanning tree $T$, i.e. $e-n+1$.

2) From part 1), we know that we have only $e-n+1$ chords. Each chord creates a fundamental cycle, and hence the number of the fundamental cycles with respect to the given spanning tree is $e-n+1$.

**Incidence matrix of a digraph**

Let $G$ be a directed graph with $n$ vertices and $e$ edges, with no circuits (cycles). The node-edge incident matrix $A$ of $G$ is an $n \times e$ matrix, defined as

$$a_{ij} = \begin{cases} 
1, & \text{if edge } j \text{ is coming in to node } i \\
-1, & \text{if edge } j \text{ is going in to node } i \\
0, & \text{otherwise.} 
\end{cases}$$

**Theorem 3:** If $A$ is the incident matrix of a connected graph with $n$ nodes, the rank of $A$ is $n-1$.

**Proof:** Since there are exactly one $-1$ and one $1$ in every column of the incidence matrix $A$, the sum of all these row vectors is $0$. Thus the $n$ vectors are not linearly independent. Therefore, the rank of $A$ is less than $n$. Since the graph $G$ is connected, it cannot be partitioned like

$$A = \begin{pmatrix} A_1 & 0 \\ 0 & A_2 \end{pmatrix}$$

such that $A_1$ is with $m$ rows and $A_2$ with $n-m$ rows. In other words, no $m$ by $m$ sub matrix of $A$ can be found, for $m \leq n-1$, such that the sum of those $m$ rows is equal to zero. Since there are only three constants $-1$, $0$, and $1$ in this field, the additions of all vectors taken $m$ at a time for $m=1,2,\ldots,n-1$ exhausts all possible linear combinations of $n-1$ row vectors. Thus we have just shown that no linear combination of $m$ row vectors of $A$, for $m \leq n-1$ can be equal to zero. Therefore, the rank of $A$ must be at least $n-1$. Since the rank of $A$ is no more than $n-1$ and no less than $n-1$, it must be exactly equal to $n-1$.

A matrix $A_f$, which is $n-1$ by $e$ obtained from $A$ by deleting one of its rows is called **reduced incidence matrix**.

**Theorem 4:** The reduced incidence matrix of a tree is invertible.

**Proof:** $A$ is the incidence matrix of a tree means, by theorem 2, it has $n$ rows and $n-1$ columns. Hence, $A_f$ is an $n-1$ by $n-1$ matrix. By theorem 3, $A$ has rank $n-1$ implies that $A_f$ is of rank $n-1$. Since $A_f$ is a square matrix of order $n-1$ and has rank $n-1$, $A_f$ is invertible.

**Cycle (circuit) matrix of a digraph**

Let $G'$ be the un-directed version of a digraph $G$, i.e., $G'$ is the graph $G$ with out considering the directions. Each cycle in $G'$, after being assigned an arbitrary orientation, can be represented by a vector whose components are are $-1$, $1$, or $0$ according to whether and how the edge is included in the cycle. A
cycle matrix $B$ is a matrix where each row corresponds to a cycle vector, and is defined by

$$b_j = \begin{cases} 
1, & \text{if cycle } i \text{ contains edge } j \text{ and their orientation coincides} \\
-1, & \text{if cycle } i \text{ contains edge } j \text{ and their orientation are opposite} \\
0, & \text{if the } i\text{th cycle does not include edge } j
\end{cases}$$

A cycle is also called circuit.

A set of fundamental cycles with respect to any spanning tree in a connected digraph are the only independent cycles in the digraph, since each contains an edge not in any of the others. The rest of cycles can obtained as linear combinations of these fundamental cycles. A sub matrix $fB$ of the matrix $B$ in which all rows correspond to the fundamental cycles is called Fundamental Cycle matrix.

If $n$ is the number of vertices and $e$ the number of edges in a connected digraph, then $B_f$ is an $e-n+1$ by $e$ matrix, because the number of fundamental cycles is $e-n+1$, each fundamental cycle being produced by one chord. Since the fundamental cycles are linearly independent the rank of $B_f$ is $e-n+1$.

**Theorem 5:** Let $B$ and $A$ be, respectively, the circuit matrix and incidence matrix of a digraph such that the columns in $B$ and $A$ are arranged using the same order of edges. Then

$$AB^T = B^T A = 0$$

**Proof:** Consider the $m$th row in $B$ and the $k$th row in $A$. If the cycle $m$ does not include any edge incident on vertex $K$, the product of the two rows is clearly zero. If, on the other hand, vertex $k$ in cycle $m$, there are exactly two edges (say $x$ and $y$) incident on $K$ that are also in cycle $m$. This situation can occur in only four different ways. The possible entries in row $k$ of $A$ and row $m$ of $B$ in column positions $x$ and $y$ are tabulated in Table 1 for each of these four cases. In each case, the dot product is zero.

### 3. Mathematical Model of a Steady State Gas Flow in Pipeline Networks

A typical gas network consists of one or more gas sources, one or more gas deliveries, pipelines, compressors, and other devices, such as valves and regulators [3] [4]. The compressors are installed in the network to increase the gas pressure so that the gas can flow through the pipeline to the locations where it is consumed. Valves and regulators provide control of the gas flow rate, prevent excessive growth of pressure in the network.

<table>
<thead>
<tr>
<th>Case</th>
<th>Row $k$</th>
<th>Row $m$</th>
<th>Dot Product</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>column $x$</td>
<td>column $y$</td>
<td>column $x$</td>
</tr>
<tr>
<td>1)</td>
<td>$-1$</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2)</td>
<td>1</td>
<td>$-1$</td>
<td>$-1$</td>
</tr>
<tr>
<td>3)</td>
<td>$-1$</td>
<td>$-1$</td>
<td>1</td>
</tr>
<tr>
<td>4)</td>
<td>1</td>
<td>1</td>
<td>$-1$</td>
</tr>
</tbody>
</table>
We consider the three basic components of a gas network to model the gas flow in pipeline networks:
1) pipelines,
2) compressors, and
3) nodes (interconnection points).

A directed graph is an efficient way to represent a gas network. The pipelines and compressors are represented as edges of the graph and the interconnecting points are represented as nodes of the graph representing the network. In a directed graph representation of a gas network, each edge has been assigned a flow direction. If the actual direction of the gas flow coincides with the direction of the edge, the flow rate is positive, otherwise the flow rate is negative.

3.1. Gas Flow in Pipes

Gas flow in a pipe is described by the three conservation laws (mass, momentum, and Energy) closed by an equation of state. If we consider the flow to be steady state and isothermal, the gas flow can be described by the following equations

\[
\frac{\partial}{\partial x} \left( \rho u^2 + p \right) = -\frac{f u |u|}{2D}
\]

where \( \rho \) is the gas density, \( u \) is the average flow velocity, \( P \) is the pressure, \( T \) is the temperature, \( f \) is friction factor, \( x \) is position, and \( D \) is the diameter of the pipe. If we use the volumetric flow rate \( Q \) and pressure \( P \) as flow variables, and use the substitution \( c = \frac{P}{\rho} \) the above equations can be written as:

\[
\frac{\partial}{\partial x} \left( \frac{Q^2 c + p}{P} \right) = -\frac{f Q |Q|}{2D}
\]

If we consider that \( c \) is the average of its values at the two end points of the pipe, we can integrate the above equation to get

\[
\frac{P_u^2}{2} - cQ^2 \log P_u - \frac{P_d^2}{2} + cQ^2 \log P_d + \frac{f Q |Q| L}{2D} = 0
\]

The logarithmic terms are because of the inertia term included in eqn which is neglected by many authors.

3.2. Compressor Equations

Compressors are installed in the gas network to transport gas and compensate for the loss of energy due to frictional resistance which results in a loss of pressure at the downstream of the pipe. The flow rate through the compressor and pressures at the upstream and downstream of the compressor are related by the following equation.

\[
\eta W = z w \frac{RTQ_c}{\gamma - 1} \left[ \left( \frac{P_u}{P_d} \right)^{\gamma - 1} - 1 \right] \quad \text{for } c = 1 : t
\]
where:

- $P_{in}$ is inlet pressure.
- $P_{out}$ is outlet pressure.
- $Q_c$ is flow rate through the compressor.
- $W$ is power of the compressor.
- $\eta$ is compressor efficiency.
- $\gamma$ is ratio of specific heats.
- $z_{in}$ is compressibility factor at entry.

3.3. Node Equations

We can categorize the nodes in the network as source nodes (where gas is supplied to the system), junction nodes (where two or more edges are connected), and sink nodes (where gas is delivered for consumption). At a source node either the pressure or the flow rate is specified. While at a junction or a sink node a flow rate balance is made.

3.3.1. Pressure Node

If the pressure value $S_p$ is specified at a node of the network, then we get the equation

$$P = S_p$$

at this node.

3.3.2. Flow Node

The node at which we balance the flow rates coming into and going out of the is called flow node. Suppose a flow node connects $N$ edges (pipes or compressors) and has a nodal flow rate $S_q$. $S_q$ is positive if gas is added into the network, negative if gas is tapped out of the network, and zero if there is no gas entering or leaving the network at this node.

Then balancing the flow rate at this node gives us

$$\sum_{j \in In_i} Q_j - \sum_{j \in Out_i} Q_j = S_q \text{ for } i = 1 : N_q$$

$In_i$ is the set of incoming flows into the $i^{th}$ Flow Node.

$Out_i$ is the set of outgoing flows from the $i^{th}$ Flow Node.

Therefore, the steady state, isothermal gas flow in pipeline networks is described by a non-linear simultaneous equation containing (3), (4), (5), and (6).

This system of equations can be written as

$$F(x) = y, \quad x = (P, Q, Q_c)$$

4. Existence and Uniqueness of the Solution to a Network Which Can Be Represented by a Tree

In the present section, proofs for existence and uniqueness of solutions to a gas flow network will be given. First, we present a proof of uniqueness of solutions which is based on a monotonicity property of a mapping. Existence is then
proved with the help of the contraction mapping theorem (4.2). Assume the gas flow network of n nodes can be represented by a tree. Suppose the pressure value $S_p$, of one node, say node 1 is given. By theorem (1) this tree has $n-1$ edges. Let $A_t$ and $A$ be, respectively, the pipe-node incidence matrix and the edge-node incidence matrix of the tree. Let $A_f$ be the reduced incidence matrix. Now the network equations can be written as:

$$\left[ A_f^T \left( \frac{P^2}{2} - cQ_p^2 \otimes \log P \right) \right] = \phi(Q),$$

where $Q_p$ and $Q$ are, respectively, pipe flows and compressor flows. Since $A_f$ is obtained from $A$ by deleting the row corresponding to the reference node, it is $n-1$ by $n-1$ matrix and is invertible. Hence, there is a unique solution for the last Equation in (8). If we have t compressors involves 2 pressures, there fore the t pressures can be determined if the values of the other t pressures are known. Assuming the t pressures from the compressor equations and one pressure from the reference node are known, we have only $n-1-t=m$ unknown pressures. By substituting the known flow rates and the known pressures in to the pipe equations we get $n-1-t$ equations for $n-1-c$ unknown pressures. Since the rows of $A_f^T$ are rows of the reduced incidence matrix $A_f$, they are linearly independent. This implies $A_f^T$ is invertible and we get the unique solution $\frac{P^2}{2} - cQ_p^2 \otimes \log P$ for the first Equation of (8). This gives us unique $P$ as $\frac{P^2}{2} - cQ_p^2 \otimes \log P$ is one-to-one.

4.1. Uniqueness of the Solution to a Gas Flow Network

In this section, let us assume we have a network of pipes, with no compressor, which can be represented by directed graph. Suppose the gas flow network has n nodes. Assume the pressure $S_p$ is given at a reference node, say at node 1. Then our network system of Equations (7) takes the form:

$$\left[ A_f^T \left( \frac{P^2}{2} - cQ_p^2 \otimes \log P \right) \right] = \phi(Q),$$

where $Q_p$ and $Q$ are, respectively, pipe flows and compressor flows. Since $A_f$ is obtained from $A$ by deleting the row corresponding to the reference node, it is $n-1$ by $n-1$ matrix and is invertible. Hence, there is a unique solution for the last Equation in (8). If we have t compressors involves 2 pressures, there fore the t pressures can be determined if the values of the other t pressures are known. Assuming the t pressures from the compressor equations and one pressure from the reference node are known, we have only $n-1-t=m$ unknown pressures. By substituting the known flow rates and the known pressures in to the pipe equations we get $n-1-t$ equations for $n-1-c$ unknown pressures. Since the rows of $A_f^T$ are rows of the reduced incidence matrix $A_f$, they are linearly independent. This implies $A_f^T$ is invertible and we get the unique solution $\frac{P^2}{2} - cQ_p^2 \otimes \log P$ for the first Equation of (8). This gives us unique $P$ as $\frac{P^2}{2} - cQ_p^2 \otimes \log P$ is one-to-one.

4.1. Uniqueness of the Solution to a Gas Flow Network

In this section, let us assume we have a network of pipes, with no compressor, which can be represented by directed graph. Suppose the gas flow network has n nodes. Assume the pressure $S_p$ is given at a reference node, say at node 1. Then our network system of Equations (7) takes the form:

$$\left[ A_f^T \left( \frac{P^2}{2} - cQ_p^2 \otimes \log P \right) \right] = \phi(Q),$$

where $Q_p$ and $Q$ are, respectively, pipe flows and compressor flows. Since $A_f$ is obtained from $A$ by deleting the row corresponding to the reference node, it is $n-1$ by $n-1$ matrix and is invertible. Hence, there is a unique solution for the last Equation in (8). If we have t compressors involves 2 pressures, there fore the t pressures can be determined if the values of the other t pressures are known. Assuming the t pressures from the compressor equations and one pressure from the reference node are known, we have only $n-1-t=m$ unknown pressures. By substituting the known flow rates and the known pressures in to the pipe equations we get $n-1-t$ equations for $n-1-c$ unknown pressures. Since the rows of $A_f^T$ are rows of the reduced incidence matrix $A_f$, they are linearly independent. This implies $A_f^T$ is invertible and we get the unique solution $\frac{P^2}{2} - cQ_p^2 \otimes \log P$ for the first Equation of (8). This gives us unique $P$ as $\frac{P^2}{2} - cQ_p^2 \otimes \log P$ is one-to-one.

4.1. Uniqueness of the Solution to a Gas Flow Network

In this section, let us assume we have a network of pipes, with no compressor, which can be represented by directed graph. Suppose the gas flow network has n nodes. Assume the pressure $S_p$ is given at a reference node, say at node 1. Then our network system of Equations (7) takes the form:

$$\left[ A_f^T \left( \frac{P^2}{2} - cQ_p^2 \otimes \log P \right) \right] = \phi(Q),$$

where $Q_p$ and $Q$ are, respectively, pipe flows and compressor flows. Since $A_f$ is obtained from $A$ by deleting the row corresponding to the reference node, it is $n-1$ by $n-1$ matrix and is invertible. Hence, there is a unique solution for the last Equation in (8). If we have t compressors involves 2 pressures, there fore the t pressures can be determined if the values of the other t pressures are known. Assuming the t pressures from the compressor equations and one pressure from the reference node are known, we have only $n-1-t=m$ unknown pressures. By substituting the known flow rates and the known pressures in to the pipe equations we get $n-1-t$ equations for $n-1-c$ unknown pressures. Since the rows of $A_f^T$ are rows of the reduced incidence matrix $A_f$, they are linearly independent. This implies $A_f^T$ is invertible and we get the unique solution $\frac{P^2}{2} - cQ_p^2 \otimes \log P$ for the first Equation of (8). This gives us unique $P$ as $\frac{P^2}{2} - cQ_p^2 \otimes \log P$ is one-to-one.
Now let us consider the system

\begin{align}
B_f \phi(Q) &= 0 \\
A_f Q &= S_q
\end{align}

The system (11) contains only the flow variable \( Q \). The first equation contains \( e-n+1 \) equations and the second contains \( n-1 \) equations. Hence, system (11) consists of \( e \) equations for \( n \) unknown flow rate variables. First, we show that the solution of the system (11) is unique, and then it can be shown that the first two equations of system (10)

\begin{align}
A^T \left( \frac{P^2}{2} - cQ^2 \otimes \log P \right) &= \phi(Q) \\
P_f &= S_p \\
B_f \phi(Q) &= 0 \\
A_f Q &= S_q
\end{align}

give unique \( P \). System (12) consists of \( e+1 \) equations for \( n \) unknowns. But, we can eliminate \( e-n+1 \) equations as they are linear combinations of the others, since \( e-n+1 \) \( \phi(Q) \) are expressed in terms of the others in the equation \( B_f \phi(Q) = 0 \).

Now let us show that the solution of system (11) is unique.

**Definition:** A mapping \( \Phi : H \to H \) is said to be strictly monotonic if for every \( x, y \in H \) we have

\[
(\Phi(x) - \Phi(y), x - y) \geq 0,
\]

and equality holds if and only if \( x = y \).

**Theorem 6:** Let \( \Phi : \mathbb{R}^d \to \mathbb{R}^d \), for some positive integer \( d \), is given by

\[
\Phi(x) = (\Phi_1(x_1), \Phi_2(x_2), \Phi_3(x_3), \ldots, \Phi_d(x_d))
\]

where

\[
\Phi_j(x_j) = c_j x_j, \quad \text{for } 1 \leq j \leq d
\]

with \( c_j > 0 \). Then \( \Phi \) is strictly monotonic.

**Proof:** For every \( x = [x_1, x_2, x_3, \ldots, x_d]^T, y = [y_1, y_2, y_3, \ldots, y_d]^T \in \mathbb{R}^d \),

\[
(\Phi(x) - \Phi(y), x - y) = \sum_{j=1}^d c_j (x_j - y_j)(x_j - y_j)
\]

The function \( h(s) = s \|s\| \) is a strictly increasing function for all \( s \). Hence, each term on the right-hand side is non-negative. Thus,

\[
(\Phi(x) - \Phi(y), x - y) \geq 0.
\]

Equality holds if and only if every term at the right hand side is zero, so
\[ x_j = y_j, \text{ for every } j. \] Therefore, \( \Phi \) is strictly monotonic.

**Definition:** Let \( r > 0, t > 0 \), be two integers, and \( d = r + t \). We say an \( r \times d \) matrix \( A \) and an \( t \times d \) matrix \( B \) are perpendicular to each other if they satisfy

1) \( \text{rank}(A) = r \), \( \text{rank}(B) = t \);
2) \( AB^T = BA^T \)

Let \( M = x \in \mathbb{R}^d : Ax = 0 \) and \( N = y \in \mathbb{R}^d : By = 0 \). Then we have \( M \) perpendicular to \( N \) and

\[ \mathbb{R}^d = M \oplus N. \]

**Theorem 7:** Let matrices \( A \) and \( B \) be perpendicular to each other. Suppose \( \Phi : \mathbb{R}^d \rightarrow \mathbb{R}^d \) is strictly monotonic, then for every \( s \in \mathbb{R}^d \), the solution to the system of equations

\[
\begin{align*}
AQ &= s \\
B\Phi(Q) &= 0
\end{align*}
\]

is unique.

**Proof:** Suppose both \( u \) and \( v \) are solutions of system (13), then

\[
\begin{align*}
A(u - v) &= 0 \\
B(\Phi(u) - \Phi(v)) &= 0
\end{align*}
\]

Hence, \( u - v \in M \), and \( \Phi(u) - \Phi(v) \in N \). Thus \( \Phi(u) - \Phi(v) = 0 \). Since \( \Phi \) is strictly monotonic, the above equation implies that \( u = v \). Hence, the solution is unique.

By applying the above theorem, and taking \( A = A_j \), \( B = B_j \), and \( s = s_j \), it is shown the solution of the system (11) is unique.

### 4.2. Existence of the Solution for a Gas Flow Network

In this section, we discuss the existence of solution to a general gas flow network that contains pipes and compressors. Let us consider a network of \( m \) pipes, \( t \) compressors, and \( n \) nodes. Assume the pressure is given, say at node 1, i.e. \( P_1 = S_p \). We use the **The Contraction Mapping Theorem** to show existence of a solution for such gas flow network.

**Definition:** Let \( (X,d) \) be a metric space. A function \( T : X \rightarrow X \) is called a contraction mapping if there exists \( k \in \mathbb{R} \) such that \( 0 < k < 1 \) and

\[
d(T(x), T(y)) \leq kd(x, y)
\]

for any \( x, y \in X \).

**Theorem 8:** ([The Contraction Mapping Theorem]) If \( (X,d) \) is a complete metric space and \( T : X \rightarrow X \) is a contraction mapping, then \( T \) has one and only one fixed point, (i.e., there exists exactly one \( x \in X \) such that \( T(x) = x \)).

Sketch of the proof of the Contraction Mapping Theorem. The complete proof is found in [5]. The proof proceeds in several steps:

1) starting from an arbitrary \( x_0 \in X \), construct the sequence \( x_n \subset X \) by taking \( x_n = T(x_{n-1}) \);
2) prove that \( d(x_n, x_{n+1}) \leq k^nd(x_0, x_1) \), and therefore
\[ d(x_n, x_m) = \frac{k^n - k^m}{1 - k} d(x_0, x_1) \] for any natural numbers \( n, m \) with \( m > n \);

3) conclude that \( \{x_n\} \) is a Cauchy sequence, and thus, since \( X \) is complete, converges;

4) prove that \( x = \lim_{n \to \infty} x_n \) is a fixed point of \( T \);

5) prove that \( x \) is the only fixed point of \( T \).

Now, we want to show that the equation

\[ T(x) = y \]

has locally unique solution. The Contraction Mapping Theorem can be applied as follows: If we can find \( x_0 \) such that the jacobian determinant, \( J \), of \( T \) at \( x_0 \) is different from zero, then we can introduce

\[ G(x) = x - J^{-1}(x_0)(T(x) - y). \]

The equation \( x = G(x) \) is equivalent to \( T(x) = y \). If \( G \) is a contraction mapping, then \( G \) has a unique fixed point \( \bar{x} \) and thus \( T(\bar{x}) = y \).

The jacobian \( J(x) \) is expressed in the following form

\[
J(x) = \begin{bmatrix}
A(x) & D(x) \\
B(x) & C(x)
\end{bmatrix}
\]

where

\[
a_{ij}(x) = \left\{ \begin{array}{ll}
a_{ij}(x) & \text{if Node is the upstream node of Pipe } i \\
\frac{\partial F_i(x)}{\partial P_{ia}} = -\left( P_{ia} - CQ_i \right), & \text{if Node is the upstream node of Compressor } c \\
\frac{\partial F_{m+c}}{\partial P_{ca}} = -\left( \eta W \frac{1}{\gamma} + bQ_c P_{cw}^{-1} \right), & \text{if Node is the downstream node of Pipe } i \\
\frac{\partial F_{m+c}}{\partial P_{cd}} = bQ_c P_{cd}^{-1}, & \text{if Node is the downstream node of Compressor } c \\
0, & \text{else}
\end{array} \right.
\]

\[
D(x) = \begin{bmatrix}
d_{ij}(x) & \text{if } i \leq m \\
\frac{\partial F_i(x)}{\partial P_{ia}} = \left( C \log \frac{P_{ia}}{P_{dd}} + \frac{CFLQ_i}{D} \right) & \text{if } i \leq m \\
\frac{\partial F_{m+c}}{\partial P_{ca}} = b \frac{1}{\gamma - 1} \left( \frac{P_{ca}}{P_{cm}} \right)^{\frac{1}{\gamma - 1}} & \text{if } m < i \leq m + t \\
d_{ij}(x) = 0 & \text{if } i \neq j
\end{bmatrix}
\]

\[
B(x) = \begin{bmatrix}
b_{ij}(x) & \text{if Node } j \text{ is the } i^{\text{th}} \text{ Pressure Node} \\
0, & \text{else}
\end{bmatrix}
\]
\[ C(x) = \left( C_y(x) \right)_{x \in \mathbb{R}^{m+i}} \]

\[
C_y(x) = \begin{cases} 
1, & \text{if flow } j \text{ flows into the } i^{th} \text{ FlowNode} \\
-1, & \text{if flow } j \text{ flows out of the } i^{th} \text{ Flow Node} \\
0, & \text{else}
\end{cases}
\]

Choose \( x_0 \) such that \( d_y(x_0) \neq 0 \), \( a_{iu}(x_0) < 0 \) and \( a_{id}(x_0) > 0 \)

Then \( \det(J(X_0)) \neq 0 \)

Example: Let us consider the following network with 3 pipes, 1 compressor, and 5 junctions.

The Jacobian of the matrix of the system of equations of the gas flow network depicted in Figure 1 is given by

\[
J = \begin{pmatrix}
\begin{array}{cccc|cccc}
 a_{iu} & a_{id} & 0 & 0 & 0 & d_{i1} & 0 & 0 \\
 0 & a_{iu} & a_{kd} & 0 & 0 & d_{22} & 0 & 0 \\
 0 & 0 & a_{iu} & a_{kd} & 0 & 0 & d_{33} & 0 \\
 0 & a_{kd} & a_{kd} & 0 & 0 & 0 & 0 & 0 \\
 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 \\
 0 & 0 & 0 & 0 & 0 & -1 & 0 & 1 \\
 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
\end{array}
\end{pmatrix}
\]

By performing elementary row and column operations on \( J \)

\[
J \sim \begin{pmatrix} 0 & I_{(m+i)} \\ B^* & 0 \end{pmatrix}
\]

where \( B^* = \left( b^*_y \right)_{x \in \mathbb{R}} \)

\[
b^*_y = \begin{cases} 
 b_y, & \text{if node } i \text{ is a pressure node} \\
 -\frac{a_{iu}}{d_{ik}}, & \text{if flow } k \text{ flows from node } j \text{ to node } i \\
 \frac{a_{id}}{d_{ik}}, & \text{if flow } k \text{ flows from node } i \text{ to node } j \\
 \sum_{k \in \text{out}} \frac{a_{iu}}{d_{ik}} - \sum_{k \in \text{in}} \frac{a_{id}}{d_{ik}}, & \text{if } i = j \text{ and node } i \text{ is a flow node} \\
 0, & \text{else}
\end{cases}
\]

\[ \det(J) \neq 0 \iff \det(B^*) \neq 0 \]

For our example given above

\[ \text{Figure 1. Example of a gas flow network.} \]
Let $B^*_d$ be the matrix obtained by deleting the rows and columns containing the Pressure Nodes, then

$$\text{det}(B^*) \neq 0 \iff \text{det}(B^*_d) \neq 0$$

Properties of $B^*_d$ are

1) Diagonally dominant (column), strictly dominant at column $j$, where node $j$ is connected to a Pressure Node (there is at least one Pressure Node).
2) Irreducible.

Hence, $B^*_d$ is invertible.

$\Rightarrow B^*$ is invertible.

$\Rightarrow J$ is invertible.

Define $G(x) = x - J^{-1}(x_0)\left(F(x) - y\right)$

$\Rightarrow G'(x) = I - J^{-1}(x_0)J(x)$

Since $G'(x)$ is continuous at $x_0$

$\forall \epsilon > 0 \exists \delta$ such that $\|x - x_0\| < \delta \Rightarrow \|G'(x)\| < \epsilon$

Choose $\epsilon$ such that $0 < \epsilon < 1$

$\forall x_1, x_2 \text{ in } B(x_0, \delta)$

$$\|G(x_1) - G(x_2)\| \leq \|G'(x)\|\|x_1 - x_2\| < \epsilon\|x_1 - x_2\|$$

If $y \in B\left(\frac{1 - \epsilon}{J^{-1}(x_0)}\delta\right)$, then $G(B(x_0, \delta)) \subseteq B(x_0, \delta)$

$\Rightarrow G$ is a contraction in $B(x_0, \delta)$

$\Rightarrow G$ has a unique fixed point.

$\Rightarrow F(x) = y$ has a unique solution in $B(x_0, \delta)$

5. Conclusion

In this paper, we discussed the existence and uniqueness of solution to the real gas flow network. A review of some basic concepts from graph theory is given as we discussed the flow network as a graph. The flow network is represented by a graph in which pipes and compressors are represented as edges and the junction points as nodes of the graph. The flow is assumed to be steady state and isothermal. The Contraction Mapping Theorem is applied to show the existence of the solution to the considered gas flow network, and its uniqueness is proved.
using monotonicity property of a mapping.

**Conflicts of Interest**

The authors declare no conflicts of interest regarding the publication of this paper.

**References**


**Nomenclature**

- $\rho$ Gas density
- $u$ Flow velocity
- $P$ Pressure
- $P_u$ Upstream pipe inlet pressure
- $P_d$ Downstream pipe outlet pressure
- $T$ Temperature
- $Q$ Volumetric flow rate in a pipe
- $x$ Position
- $D$ Pipe diameter
- $L$ Pipe length
- $\eta$ Compressor efficiency
- $W$ Compressor power
- $Z_{cu}$ Comprehensibility factor of a gas at compressor entry
- $R$ Gas constant
- $Q_c$ Flow rate through the compressor
- $\gamma$ Ratio of specific heats
- $P_{cu}$ Inlet compressor pressure
- $P_{cd}$ Outlet compressor pressure
On the Uphill Domination Polynomial of Graphs

Thekra Alsalomy, Anwar Saleh, Najat Muthana, Wafa Al Shammakh

Department of Mathematics, Faculty of Science, University of Jeddah, Jeddah, Saudi Arabia
Email: talsalomy.stu@uj.edu.sa, math.msfs@gmail.com, nmuthana@uj.edu.sa, wsalshamahk@uj.edu.sa

Abstract
A path \( \pi = [v_1, v_2, \ldots, v_k] \) in a graph \( G = (V, E) \) is an uphill path if \( \text{deg}(v_i) \leq \text{deg}(v_{i+1}) \) for every \( 1 \leq i \leq k \). A subset \( S \subseteq V(G) \) is an uphill dominating set if every vertex \( v_j \in V(G) \) lies on an uphill path originating from some vertex in \( S \). The uphill domination number of \( G \) is denoted by \( \gamma_{up}(G) \) and is the minimum cardinality of the uphill dominating set of \( G \). In this paper, we introduce the uphill domination polynomial of a graph \( G \). The uphill domination polynomial of a graph \( G \) of \( n \) vertices is the polynomial \( UP(G, x) = \sum_{i=\gamma_{up}(G)}^{n} up(G, i) x^i \), where \( up(G, i) \) is the number of uphill dominating sets of size \( i \) in \( G \), and \( \gamma_{up}(G) \) is the uphill domination number of \( G \).

Keywords
Domination, Uphill Domination, Uphill Domination Polynomial

1. Introduction
In this paper, we are concerned with simple graphs which are finite, undirected with no loops nor multiple edges. Throughout this paper, we let \( |V(G)| = n \) and \( |E(G)| = m \). In a graph \( G = (V, E) \), the degree of \( v \in V(G) \) denoted by \( \text{deg}(v) \) is the number of edges that incident with \( v \). A path in \( G \) is an alternating sequence of distinct vertices. A path is an uphill path if for every \( 1 \leq i \leq k \) we have \( \text{deg}(v_i) \leq \text{deg}(v_{i+1}) \) [1].

The bistar graph \( S_{k_1,k_2} \) with \( n = 2k_1 + 2 \) vertices is obtained by joining the non-pendant vertices of two copies of star graph \( S_{k_1} \) by new edge. The corona of two graphs \( G_1 \) and \( G_2 \) with \( n_1 \) and \( n_2 \) vertices, respectively, denoted by...
$G = G_1 \circ G_2$ is obtained by taking one copy of $G_1$ and $n_i$ copies of $G_2$ and joining the $i$th vertex of $G_1$ with an edge to every vertex in the $i$th copy of $G_2$. The corona $G \circ K_i$ (in particular) is the graph constructed by a copy of $G$, where for each vertex $v \in V(G)$ a new vertex $v'$ and a pendant edge $vv'$ are added. The tadpole graph $T_{s,k}$ is a graph consisting of a cycle graph $C_s$ on at least three vertices and a path graph $P_k$ on $k$ vertices connected with bridge. The wheel graph $W_n$ is a graph formed by connecting a single vertex to all vertices of a cycle graph $C_n$. The book graph is a Cartesian product $B_{m,s} = S_m \times P_s$, where $S_m$ is the star graph with $m+1$ vertices and $P_s$ is the path graph on two vertices. Also, the windmill graph $W_d(s,k)$ is a graph constructed for $s \geq 2$ and $k \geq 2$ by joining $k$ copies of the complete graph $K_s$ at a shared universal vertex. The friendship graph $F_k$ is a graph that constructed by joining $k$ copies of the cycle graph $C_3$ and observes that $F_k$ is a special case of $D(s,k)$. Finlay, the firefly graph $F_{s,t,k}$ with $s,t,k \geq 0$ and $n = 2s + 2t + k + 1$ vertices is defined by consisting of $s$ triangles, $t$ pendant paths of length 2 and $k$ pendant edges, sharing a common vertex. Any terminology not mentioned here we refer the reader to [2].

A set $S \subseteq V$ of vertices in a graph $G$ is called a dominating set if every vertex $v \in V$ is either $v \in S$ or $v$ is adjacent to an element of $S$. The uphill dominating set "UDS" is a set $S \subseteq V$ having the property that every vertex $v \in V$ lies on an uphill path originating from some vertex in $S$. The uphill domination number of a graph $G$ is denoted by $\gamma_{up}(G)$ and is defined to be the minimum cardinality of the UDS of $G$. Moreover, it’s customary to denote the UDS having the minimum cardinality by $\gamma_{up}(G)$-set, for more details in domination see [3] and [4].

Representing a graph by using a polynomial is one of the algebraic representations of a graph to study some of algebraic properties and graph’s structure. In general graph polynomials are a well-developed area which is very useful for analyzing properties of the graphs.

The domination polynomial [5] and the uphill domination of a graph [6], motivated us to introduce and study the uphill domination polynomial and the uphill domination roots of a graph.

2. Uphill Domination Polynomial

**Definition 2.1.** For any graph $G$ of $n$ vertices, the uphill domination polynomial of $G$ is defined by

$$UP(G,x) = \sum_{i=\gamma_{up}(G)}^n \gamma_{up}(G,i)x^i,$$

where $\gamma_{up}(G,i)$ is the number of uphill dominating sets of size $i$ in $G$. The set of roots of $UP(G,x)$ is called uphill domination roots of graph $G$ and denoted by $Z_{up}(G)$.

**Example 2.2.** The uphill domination polynomial of House graph $H$ (as shown
in Figure 1) with 6 vertices and \( \gamma_{up}(H) = 2 \) is given by

\[ UP(H,x) = 2x^2 + 7x^3 + 9x^4 + 5x^5 + x^6. \]

Furthermore, \( Z_{up}(H) = \{0, -1, -2\} \).

The following theorem gives the sufficient condition for the uphill domination polynomial of \( r \)-regular graph.

**Theorem 2.3.** Let \( G \) be connected graph with \( n \geq 2 \) vertices. Then,\[ UP(G,x) = (1+x)^n - 1 \]if and only if \( G \) is \( r \)-regular graph.

**Proof.** Let \( G \) be a connected graph of \( n \geq 2 \) vertices. Suppose that the uphill domination polynomial of \( G \) is given by

\[ UP(G,x) = nx + \binom{n}{2}x^2 + \cdots + x^n. \]

Since the first coefficient of the polynomial is \( n \), then it is easily verified that for every \( v \in V(G) \), the singleton vertex set \( \{v\} \) is an UDS in \( G \). Assume that \( G \) is not \( r \)-regular graph. Hence there exists a vertex \( u \in V(G) \) such that \( \deg(u) = s \neq r \). Now, we have two cases:

Case 1: If \( s > r \), then the set \( \{u\} \) is not UDS which contradict that every singleton vertex set is an UDS in \( G \).

Case 2: If \( s < r \), then for all \( u \neq v \) with \( \deg(v) = r \), we get the set \( \{v\} \) is not UDS which is also contradict that every singleton vertex set is an UDS in \( G \).

Thus, \( G \) must be \( r \)-regular graph.

On the other hand, suppose that \( G \) is \( r \)-regular graph with \( n \geq 2 \) vertices. We have \( \gamma_{up}(G) = 1 \), then there exist \( n \) UDS of size one, while for \( i = 2 \) there are \( \binom{n}{2} \) UDS and so on. Thus, we can write the uphill domination polynomial as

\[ UP(G,x) = nx + \binom{n}{2}x^2 + \cdots + x^n = (1+x)^n - 1. \]

**Corollary 2.4.** Let \( G \) ba a graph with \( s \) vertices. If \( G \) is a cycle \( C_s \), or complete graph \( K_s \), then \( UP(G,x) = (1+x)^s - 1 \).

**Corollary 2.5.** The uphill domination polynomial for the regular graph \( G = C_s \times C_t \) with \( sk \) vertices is given by \( UP(G,x) = (1+x)^k - 1 \).

**Corollary 2.6.** [6] Let \( G \) be a graph with \( m \) components. Then,

\[ \gamma_{up}(G) = \sum_{i=1}^{m} \gamma_{up}(G_i). \]

**Proposition 2.7.** If a graph \( G \) with \( n \) vertices consists of \( m \) components \( G_1, G_2, \ldots, G_m \), then

![Image](image.png)

**Figure 1.** The House graph.
\[
UP(G, x) = \prod_{i=1}^{m} UP(G_i, x).
\]

**Proof.** By using mathematical induction we found that for \( m = 1 \) the statement is true and the proof is trivial. Suppose that the statement is true when \( m = k \) such that

\[
UP(G, x) = \prod_{i=1}^{k} UP(G_i, x).
\]

Now, we prove that the statement is true when \( m = k + 1 \). Let \( G \) consists of \( k + 1 \) components that mean \( G = G_1 \cup G_2 \cup \cdots \cup G_{k+1} \). If the set \( \{r_1, r_2, \ldots, r_{k+1}\} \) represent the uphill domination number for the components of \( G \) respectively, such that \( \gamma_{up}(G_i) = r_i \ \forall 1 \leq i \leq k + 1 \). Then, by Corollary (2.6) it easily to see that

\[
\gamma_{up}(G) = \gamma_{up}\left( \bigcup_{1 \leq i \leq k+1} G_i \right) = \sum_{1 \leq i \leq k+1} \gamma_{up}(G_i) = r_1 + \cdots + r_{k+1} = r.
\]

Thus, \( up(G, r) \) is exactly equal the number of way for choosing an UDS of size \( r \) in \( G_i \) and an UDS of size \( r_i \) in \( G_2 \) and so on. Hence, \( up(G, r) \) is the coefficient of \( x^r \) in \( UP(G_1, x)UP(G_2, x)\cdots UP(G_{k+1}, x) \) and in \( UP(G, x) \).

In the same argument we can proof for all \( up(G, j) \), where \( r \leq j \leq n \) that

\[
up(G, j) = up(G_1, j)\cdots up(G_{k+1}, j) = \prod_{i=1}^{k+1} up(G_i, j).
\]

Thus, for \( m = k + 1 \) the statement is true and the proof is done.

**Theorem 2.8.** For any path \( P_n \) with \( n \geq 3 \) vertices, \( UP(G, x) = x^2 (1 + x)^{n-2} \). Furthermore, \( Z_{up}(P_n) = \{0, -1\} \).

**Proof.** Let \( G \) be a path graph \( P_n \) with \( n \geq 3 \). We know that \( \gamma_{up}(P_n) = 2 \), then there is only one UDS of size two. For \( i = 3 \) there are \( n-2 \) UDS of size three and so on. Thus, we get

\[
UP(G, x) = x^2 + \left( \frac{n-2}{1} \right) x^3 + \left( \frac{n-2}{2} \right) x^4 + \cdots + \left( \frac{n-2}{n-2} \right) x^n
\]

\[
= x^2 \left[ 1 + \sum_{i=1}^{n-2} \left( \frac{n-2}{i} \right) x^i \right]
\]

\[
= x^2 \left[ \sum_{i=0}^{n-2} \left( \frac{n-2}{i} \right) x^i \right]
\]

\[
= x^2 (1 + x)^{n-2}.
\]

**Theorem 2.9.** For any graph \( G \). \( UP(G, x) = x^n \) if and only if \( G \cong \overline{K}_n \).

**Proof.** Let \( G \) be a graph with \( UP(G, x) = x^n \). Since, \( UP(\overline{K}_1, x) = x \), then we can write that

\[
UP(G, x) = x^n
\]

\[
= x \cdot x \cdots x \quad \text{a times}
\]

\[
= UP(\overline{K}_1, x) \cdot UP(\overline{K}_1, x) \cdots UP(\overline{K}_1, x)
\]

\[
= UP(\overline{K}_n, x).
\]
Thus, $G \cong \overline{K}_n$. On the other hand if $G \cong \overline{K}_n$, then by Proposition (2.7) we get $UP(G, x) = x^n$.

**Corollary 2.10.** A graph $G$ has one uphill domination root if and only if $G \cong \overline{K}_n$.

**Theorem 2.11.** Let $G$ be a bistar graph $S_{k_1, k_1}$ with $n = 2k_1 + 2$ vertices. Then, $UP(G, x) = x^{2k_1} (1 + x)^3$. Furthermore, $Z_{up}(G) = \{0, -1\}$.

**Proof.** Let $G$ be a bistar graph $S_{k_1, k_1}$ with $n = 2k_1 + 2$ vertices, we have $\gamma_{up}(G) = 2k_1$. Then, there is only one UDS of size $2k_1$ and for $i = 2k_1 + 1$ there are two UDS. Finally, for $i = 2k_1 + 2 = n$ there is only one UDS. Thus, the result will be as following

$$UP(G, x) = x^{2k_1} + 2x^{2k_1+1} + x^{2k_1+2}$$

$$= x^{2k_1} \left[1 + 2x + x^2\right]$$

$$= x^{2k_1} (1 + x)^2.$$

**Theorem 2.12.** For any graph $G \cong K_{r, s}$ with $r < s$ and $r + s \geq 3$ vertices, $UP(G, x) = x^s (1 + x)^r$. Furthermore, $Z_{up}(K_{r, s}) = \{0, -1\}$.

**Proof.** Let $G$ is a complete bipartite graph $K_{r, s}$ with $r < s$, then we have $\gamma_{up}(K_{r, s}) = s$. There is only one UDS of size $s$. Now, for $i = s + 1$ there exist $r$ UDS. For $i = s + 2$ there exist $\binom{r}{2}$ UDS and so on. Thus, we get

$$UP(G, x) = x^s + \binom{r}{1} x^{s+1} + \binom{r}{2} x^{s+2} + \cdots + \binom{r}{r} x^{s+r}$$

$$= x^s + \sum_{i=1}^{r} \binom{r}{i} x^{s+i}$$

$$= x^s \left[\sum_{i=0}^{r} \binom{r}{i} x^i\right]$$

$$= x^s (1 + x)^r.$$

**Corollary 2.13.** For any graph $G \cong S_r$ with $r + 1$ vertices, $UP(G, x) = x^r (1 + x)^2$. Furthermore, $Z_{up}(G) = \{0, -1\}$.

The generalization of Theorem 0.12 is the following result.

**Theorem 2.14.** For any graph $G \cong K_{r_1, \ldots, r_k}$ where $r_1 < r_2 < \cdots < r_k$ with $n = \sum_{i=1}^{k} r_i$ vertices, $UP(G, x) = x^n (1 + x)^{n-r_k}$. Furthermore, $Z_{up}(K_{r_1, \ldots, r_k}) = \{0, -1\}$.

**Proof.** Let $G$ be a complete $k$-partite graph $K_{r_1, \ldots, r_k}$ with $r_1 < r_2 < \cdots < r_k$, we have $\gamma_{up}(K_{r_1, \ldots, r_k}) = r_k$. There is only one UDS of size $r_k$ for $i = r_k + 1$ there are $n-r_k$ UDS of size $r_k + 1$. Also, for $i = r_k + 2$ there are $\binom{n-r_k}{2}$ and so on. Thus,
T. Alsalomy et al.

PROPOSITION 2.15. For any graph \( G \equiv K_{r_1, r_2, \ldots, r_k} \) with \( n = \sum_{i=1}^{k} r_i \) vertices we have the following:

1) If \( r_1 \leq r_2 \leq \cdots \leq r_{k-1} < r_k \), such that at least two partite sets of the same size, then \( UP(G, x) = x^{r_1} (1 + x)^{r_k} \).

2) If \( r_1 = r_2 = \cdots = r_k \), then the graph is regular and \( UP(G, x) = (1 + x)^n - 1 \).

THEOREM 2.16. For any graph \( G \equiv K_{r_1, r_2, \ldots, r_k} \) with \( n = \sum_{i=1}^{k} r_i \) vertices, where \( r_1 \leq r_2 \leq \cdots < r_{k-1} = r_k \). Then,

\[
UP(G, x) = \sum_{k=1}^{n} \sum_{r_1 \leq \cdots \leq r_k} \left( \prod_{i=1}^{k} \binom{2r_i}{r_i} \right) x^{n-2r_k}
\]

Proof. Let \( G \) be a complete \( k \)-partite graph \( K_{r_1, r_2, \ldots, r_k} \) with \( r_1 \leq r_2 \leq \cdots < r_{k-1} = r_k \), then we have \( \gamma_{up}(K_{r_1, r_2, \ldots, r_k}) = 1 \). Let divide the vertices of a graph into two sets \( R_1 \) and \( R_2 \) where \( R_1 \) contains the vertices of \( r_k \) and \( r_{k-1} \), which means \( R_1 \) is of cardinality \( 2r_k \) while \( R_2 = V(G) \setminus R_1 \) this implies that \( R_2 \) is of cardinality \( n - 2r_k \). Thus, we get

\[ up(G, 1) = \binom{2r_k}{1} \binom{n - 2r_k}{0} = 2r_k. \]

We have for \( up(G, 2) \),

\[ up(G, 2) = \binom{2r_k}{2} \binom{n - 2r_k}{0} + \binom{2r_k}{1} \binom{n - 2r_k}{1}. \]

Also, for \( up(G, 3) \) we get

\[ up(G, 3) = \binom{2r_k}{3} \binom{n - 2r_k}{0} + \binom{2r_k}{2} \binom{n - 2r_k}{1} + \binom{2r_k}{1} \binom{n - 2r_k}{2}. \]

And so on we get for all \( up(G, h) \), where \( 1 \leq h \leq n \)

\[ up(G, h) = \sum_{r_1 \leq \cdots \leq r_k} \binom{2r_k}{r_k} \binom{n - 2r_k}{r_2}. \]

Thus, the proof is done.

THEOREM 2.17. For any graph \( G \equiv W_s \) with \( s + 1 \) vertices and \( s > 3 \), then \( UP(G, x) = (1 + x) [ (1 + x)^s - 1 ] \).

Proof. Let \( G \) be a wheel graph \( W_s \) (\( s > 3 \)), then we have \( \gamma_{up}(W_s) = 1 \). There
are $s$ UDS of size one. For $i = 2$, there are $\binom{s+1}{2}$ UDS of size two and so on. Thus,

$$UP(G, x) = sx + \left(\frac{s+1}{2}\right)x^2 + \left(\frac{s+1}{3}\right)x^3 + \cdots + \left(\frac{s+1}{s}\right)x^{s+1}$$

$$= \sum_{i=0}^{s} \left(\frac{s+1}{i}\right)x^i - (x+1)$$

$$= (x+1)^{s+1} - (x+1)$$

$$= (x+1)\left[(x+1)^i - 1\right].$$

**Corollary 2.18.** For any wheel graph $W_s$ and $s > 3$ we have

$$Z_{up}(W_s) = \begin{cases} \{0,-1,-2\}, & \text{if } s \text{ is even.} \\ \{0,-1\}, & \text{if } s \text{ is odd.} \end{cases}$$

### 3. Uphill Domination Polynomials of Graphs under Some Binary Operations

**Theorem 3.1.** Let $G \cong P_r \times P_s$ be a grid graph with $rs$ vertices and $r, s \geq 4$. Then, $UP(G, x) = x^r (1+x)^{r-4}$.

**Proof.** Let $G$ be a grid graph with $rs$ vertices and $r, s \geq 4$, then we have $\gamma_{up}(G) = 4$. Note that, there is only one UDS of size four. For $i = 5$, there are $rs-4$ UDS of size five and so on. Thus, we get

$$UP(G, x) = x^4 + \left(\frac{rs-4}{4}\right)x^5 + \cdots + \left(\frac{rs-4}{rs-4}\right)x^{rs}$$

$$= x^4\left[\sum_{i=0}^{rs-4} \left(\frac{rs-4}{i}\right)x^i\right]$$

$$= x^4(1+x)^{r-4}.$$

**Theorem 3.2.** Let $G \cong C_r \circ K_s$ be a corona graph with $rs + r$ vertices. Then, $UP(G, x) = x^r (1+x)^r$.

**Proof.** Let $G$ be a corona graph with $rs + r$ vertices, we have $\gamma_{up}(C_r \circ K_s) = rs$. For $rs$ vertices, there is only one UDS of size $rs$. For $rs + 1$ vertices, there are $r$ UDS and so on. Thus, we get

$$UP(G, x) = x^{rs} + \left(\frac{r}{1}\right)x^{rs+1} + \cdots + \left(\frac{r}{r}\right)x^{rs+r}$$

$$= \sum_{i=0}^{r} \left(\frac{r}{i}\right)x^{rs+i}$$

$$= x^{rs}\left[\sum_{i=0}^{r} \left(\frac{r}{i}\right)x^i\right]$$

$$= x^{rs}(1+x)^r.$$

**Corollary 3.3.** Let $G \cong C_r \circ K_i$ be a corona graph with $2r$ vertices. Then, $UP(G, x) = x^r (1+x)^r$.
Theorem 3.2 can generalize in the following result.

**Theorem 3.4.** For any nontrivial connected graph $H$ with $r$ vertices, if $G \cong H \circ \overline{K}_s$, then, $UP(G, x) = x^s (1 + x)^r$.

**Proof.** The proof similarly to the proof of Theorem 3.2.

**Theorem 3.5.** Let $G$ be a book graph $B_m = P_2 \times S_m$ with $2m + 2$ vertices. Then,

$$UP(G, x) = 2^m x^m + \left[ m \left( 2^{m-1} \right) + 2^{m+1} \right] x^{m+1} + \sum_{i=2}^{2m+1} \left( \begin{array}{c} m \\ i-1 \end{array} \right) 2^{m-i+1} x^{m+1} + \sum_{i=2}^{2m+1} \left( \begin{array}{c} m \\ i-2 \end{array} \right) 2^{m-i+2} x^{m+1} + \left[ 1 + m2^2 + \left( \begin{array}{c} m \\ 2 \end{array} \right) \right] x^{2m} + (2m + 1)x^{2m+1} + x^{2m+2}.$$

**Proof.** Suppose we have the book graph $B_m = P_2 \times S_m$ with $2m + 2$ vertices, then we have $\gamma_{up}(B_m) = m$. Let divide the vertices of $B_m$ into $m+1$ sets “as shown in Figure 2” let the set $R_i = \{u_i, v_i\}$ i.e., $1 \leq i \leq m$ while $R_{m+1} = \{u, v\}$. Since $\gamma_{up}(B_m) = m$, then for $up(G, m)$ we have to take one vertex from each $R_i$ $(i \neq m+1)$ so, there exist $2^m$ UDS of size $m$. For $up(G, m+1)$ we have,

$$up(G, m+1) = \sum_{\eta: |\eta|=m+1} \left( \begin{array}{c} 2 \\ 1 \end{array} \right) \ldots \left( \begin{array}{c} 2 \\ 1 \end{array} \right) \ldots \left( \begin{array}{c} 2 \\ 1 \end{array} \right) \left( \begin{array}{c} 2 \\ 0 \end{array} \right) \left( \begin{array}{c} 2 \\ 0 \end{array} \right) \ldots \left( \begin{array}{c} 2 \\ 0 \end{array} \right)$$

$$= 2^{m+1} + m2^{m-1}.$$

Also, for $up(G, m+2)$ we get

$$up(G, m+2) = \sum_{\eta: |\eta|=m+2} \left( \begin{array}{c} 2 \\ 1 \end{array} \right) \ldots \left( \begin{array}{c} 2 \\ 1 \end{array} \right) \ldots \left( \begin{array}{c} 2 \\ 1 \end{array} \right) \left( \begin{array}{c} 2 \\ 1 \end{array} \right) \ldots \left( \begin{array}{c} 2 \\ 1 \end{array} \right) \left( \begin{array}{c} 2 \\ 1 \end{array} \right) \left( \begin{array}{c} 2 \\ 1 \end{array} \right) \ldots \left( \begin{array}{c} 2 \\ 1 \end{array} \right)$$

$$+ \sum_{\eta: |\eta|=m+2} \left( \begin{array}{c} 2 \\ 1 \end{array} \right) \ldots \left( \begin{array}{c} 2 \\ 1 \end{array} \right) \ldots \left( \begin{array}{c} 2 \\ 1 \end{array} \right)$$

$$= \left( \begin{array}{c} m \\ 2 \end{array} \right) 2^{m-2} + \left( \begin{array}{c} m \\ 1 \end{array} \right) 2^m + \left( \begin{array}{c} m \\ 0 \end{array} \right) 2^m$$

$$= \left( \begin{array}{c} m \\ 2 \end{array} \right) 2^{m-2} + 2^m + 2^m.$$

**Figure 2.** A Book Graph $B_m.$
Therefore, for \( \text{up}(G, m+3) \) we have
\[
\text{up}(G, m+3) = \sum_{\eta_1, \eta_2, \eta_3 \geq 3} \left( \begin{array}{c} 2 \\ \eta_1 \\ \eta_2 \\ \eta_3 \\ \end{array} \right) \left( \begin{array}{c} 1 \\ 0 \\ 0 \\ \end{array} \right) + \sum_{\eta_1, \eta_2, \eta_3 \geq 2} \left( \begin{array}{c} 2 \\ \eta_1 \\ \eta_2 \\ \eta_3 \\ \end{array} \right) \left( \begin{array}{c} 2 \\ 2 \\ 2 \\ \end{array} \right) \\
+ \sum_{\eta_1, \eta_2, \eta_3 \geq 1} \left( \begin{array}{c} 2 \\ \eta_1 \\ \eta_2 \\ \eta_3 \\ \end{array} \right) \left( \begin{array}{c} 2 \\ 2 \\ \end{array} \right)
\]
\[
= \left( \begin{array}{c} m \\ \end{array} \right) 2^{m-3} + \left( \begin{array}{c} m \\ \end{array} \right) 2^m + \left( \begin{array}{c} 2 \\ \end{array} \right) 2^m
\]
\[
= \left( \begin{array}{c} m \\ \end{array} \right) 2^{m-3} + \left( \begin{array}{c} m \\ \end{array} \right) 2^m + 2^m.
\]

And so on, we use the same argument until \( \text{up}(G, 2m-1) \). After that, for \( \text{up}(G, 2m) \) we have
\[
\text{up}(G, 2m) = \left( \begin{array}{c} 2m + 2 \\ 2m + 1 \\ \end{array} \right) = 2m + 2 \quad \& \quad \text{up}(G, 2m + 2) = 1.
\]

Finally,
\[
\text{up}(G, 2m + 1) = \left( \begin{array}{c} 2m + 2 \\ 2m + 1 \\ \end{array} \right) = 2m + 2
\]

Thus, the proof is completed.

**Theorem 3.6.** Let \( G \) be a graph. If \( G \cong P_s \times C_t \) with \( sk \) vertices, then
\[
\text{UP}(G, x) = \sum_{s, t \geq 2} \left( \begin{array}{c} s \\ \eta_1, \eta_2, \eta_3 \geq 2 \end{array} \right) \left( \begin{array}{c} \eta_1 \\ \eta_2 \\ \eta_3 \\ \end{array} \right) \left( \begin{array}{c} sk - 2s \\ \eta_1 \\ \eta_2 \\ \eta_3 \\ \end{array} \right) x^t
\]

**Proof.** Let \( G \cong P_s \times C_t \) with \( sk \) vertices, then we have \( \gamma_{up}(P_s \times C_t) \geq 2 \). We first divide the vertices of \( G \) into three sets called them \( R_1, R_2 \) and \( R_3 \), where \( R_1 \) (resp. \( R_2 \)) is contains the vertices of the outer cycle (resp. inner cycle) which every vertex is of degree three. The third set \( R_3 \) contains the vertices of the middle cycles, where every vertex is of degree four. Note that, any UDS should contain at least one vertex form \( R_1 \) and one vertex from \( R_2 \). Thus, for \( \text{up}(G, 2) \)
\[
\text{up}(G, 2) = \left( \begin{array}{c} s \\ 1 \\ 0 \\ \end{array} \right) \left( \begin{array}{c} sk - 2s \\ \eta_1 \\ \eta_2 \\ \eta_3 \\ \end{array} \right) = s^2.
\]

For \( \text{up}(G, 3) \) we have
\[
\text{up}(G, 3) = \sum_{\eta_1, \eta_2, \eta_3 \geq 2} \left( \begin{array}{c} s \\ \eta_1 \\ \eta_2 \\ \eta_3 \\ \end{array} \right) \left( \begin{array}{c} sk - 2s \\ \eta_1 \\ \eta_2 \\ \eta_3 \\ \end{array} \right)
\]

And so on, we use the same argument for all \( \text{up}(G, t) \) i.e., \( 3 \leq t \leq sk \) and
the proof is done.

**Theorem 3.7.** Let \( G \) be a tadpole graph \( T_{s,k} \) with \( s+k \) vertices. Then,

\[
UP(G,x) = (s-1)x^2 + \sum_{i=1}^{s+k} \left[ \sum_{r_2+r_3 \geq 1} \binom{k}{r_2} \binom{s-1}{r_3} \right] x'.
\]

**Proof.** Let \( G \) be a tadpole graph \( T_{s,k} \) with \( s+k \) vertices, we have \( \gamma_{up}(T_{s,k}) = 2 \). We first divide the vertices of \( T_{s,k} \) into three sets called them \( R_1, R_2 \) and \( R_3 \) such that \( R_1 \) is a singleton set that contains the pendant vertex, \( R_2 \) has \( k \) vertices each of them is of degree two except one vertex is of degree three while the last set \( R_3 \) has \( s-1 \) vertices each of them of degree two which are the vertices that lies in a cycle part of a graph. Notice that, any UDS of \( T_{s,k} \) should contains the pendant vertex and at least one vertex from \( R_3 \). Now, for \( up(G,2) \) we have to take the pendant vertex with one vertex from \( R_3 \), so there exist \( s-1 \) UDS of size two. For \( up(G,3) \) we get

\[
up(G,3) = \sum_{r_2+r_3 \geq 2} \binom{k}{r_2} \binom{s-1}{r_3}.
\]

And so on, we use the same argument for all \( up(G,t) \) i.e., \( 3 \leq t \leq s+k \) and the proof is completed.

**Theorem 3.8.** Let \( G \) be a windmill graph \( Wd(s,k) \) with \( k(s-1)+1 \) vertices. Then,

\[
UP(G,x) = (s-1)^k x^k + \sum_{i=1}^{k(s-1)+1} \left[ \sum_{h_1+\cdots+h_{i+1} = t} \binom{s-1}{h_1} \cdots \binom{s-1}{h_{i+1}} \binom{1}{r_{k+1}} \right] x'.
\]

**Proof.** Let \( G \) be a windmill graph with center vertex \( w \), we have \( \gamma_{up}(G) = k \). Any minimum uphill domination set must contains one vertex from each copy of \( K_s \) without the center vertex \( w \), that means, we have \( (s-1)^k \) uphill dominating set of size \( k \). Suppose \( R_i \) be the set of vertices of the \( i \)-th copy of \( K_s \) without the center vertex \( w \) and \( R_w \) be the singleton, with the center vertex \( w \). To get the number of uphill dominating sets of size \( t = k+j \), where \( j = 1,2,\cdots,(k(s-2)+1) \), we need to select \( r_j \) vertices from each \( R_i \), and \( r_{k+1} \) from \( R_w \) where \( i = 1,2,\cdots,k \), \( \sum_{i=1}^{k+1} r_i = t \) and \( r_i \geq 1 \) for all \( i = 1,2,\cdots,k \). Hence,

\[
up(G,t) = \sum_{i=1}^{k+1} \binom{s-1}{r_1} \cdots \binom{s-1}{r_{k+1}} \binom{1}{r_{k+1}}.
\]

Thus,

\[
UP(G,x) = (s-1)^k x^k + \sum_{i=k+1}^{k(s-1)+1} \left[ \sum_{h_1+\cdots+h_{i+1} = t} \binom{s-1}{h_1} \cdots \binom{s-1}{h_{i+1}} \binom{1}{r_{k+1}} \right] x'.
\]

**Proposition 3.9.** Let \( G \) be a dutch windmill graph \( D(s,k) \) with \( s>3 \) and \( k(s-1)+1 \) vertices. Then,
Theorem 3.10. Let $G$ be a firefly graph $F_{s,t,k}$ with $s,t,k \geq 0$, $n = 2s + 2t + k + 1$ vertices and $\gamma_{up}(G) = s + t + k = b$. Then,

$$Up(G,x) = 2^s x^b + \left[2^t (t+1) + 2^{s+1} (s)\right]x^{s+1}$$

$$+ \sum_{k=b+2}^{n} \left[\sum_{\eta_1, \ldots, \eta_{s+1}=b}^{s+1} \left(\eta_1 r_1 \ldots \eta_s r_s \eta_{s+1} r_{s+1}\right)\right]x^b.$$ 

Proof. Let $G$ be a firefly graph $F_{s,t,k}$ with $n$ vertices and $\gamma_{up}(G) = s + t + k = b$. First, let us divide the vertices of $G$ into $s+2$ sets and let $u$ be the shared vertex in $G$. Suppose that $R_1 \subset V(G)$ contains the vertices of the first triangle without $u$, this implies $R_1$ has two vertices each of them are of degree two, also we mean by $R_2 \subset V(G)$ the set that contains the vertices of the second triangle without $u$ and so on for all $R_i$, where $1 \leq i \leq s$. Now, the subset $R_{s+1} \subset V(G)$ contains $u$ in addition the $t$ vertices of the pendant paths that adjacent to $u$ which means $R_{s+1}$ is of cardinality $t+1$. Finally, $R_{s+2} \subset V(G)$ contains all the leaves vertices of $G$ which are exactly of cardinality $t+k$. Notice that, any UDS of $G$ should contain all the vertices of $R_{s+2}$ with at least one vertex from each $R_i$. Thus, for $up(G,b)$ we have

$$up(G,b) = \sum_{\sum_{i=1}^{s+1} \eta_i = b}^{s+1} \left[\begin{array}{c} 2 \\ \eta_1 \\ \vdots \\ \eta_s \\ r_s \\ r_{s+1} \end{array}\right]$$

$$= \sum_{\sum_{i=1}^{s+1} \eta_i = b}^{s+1} \left[\begin{array}{c} 2 \\ \eta_1 \\ \vdots \\ \eta_s \\ r_s \\ r_{s+1} \end{array}\right] \left[\begin{array}{c} (t+1) \\ r_s \end{array}\right] \left[\begin{array}{c} t+k \\ r_{s+1} \end{array}\right]$$

$$= 2 \times 2 \times \cdots \times 2 = 2^s.$$

For $up(G,b+1)$ we get

$$up(G,b+1) = \sum_{\sum_{i=1}^{s+1} \eta_i = b+1}^{s+1} \left[\begin{array}{c} 2 \\ \eta_1 \\ \vdots \\ \eta_s \\ r_s \\ r_{s+1} \end{array}\right]$$

$$= \sum_{\sum_{i=1}^{s+1} \eta_i = b+1}^{s+1} \left[\begin{array}{c} 2 \\ \eta_1 \\ \vdots \\ \eta_s \\ r_s \\ r_{s+1} \end{array}\right] \left[\begin{array}{c} (t+1) \\ r_s \end{array}\right] \left[\begin{array}{c} t+k \\ r_{s+1} \end{array}\right]$$

$$= 2^{s+1} (s) + 2^t (t+1).$$

And for $up(G,b+2)$ we have

$$up(G,b+2) = \sum_{\sum_{i=1}^{s+1} \eta_i = b+2}^{s+1} \left[\begin{array}{c} 2 \\ \eta_1 \\ \vdots \\ \eta_s \\ r_s \\ r_{s+1} \end{array}\right]$$

$$= \sum_{\sum_{i=1}^{s+1} \eta_i = b+2}^{s+1} \left[\begin{array}{c} 2 \\ \eta_1 \\ \vdots \\ \eta_s \\ r_s \\ r_{s+1} \end{array}\right] \left[\begin{array}{c} (t+1) \\ r_s \end{array}\right] \left[\begin{array}{c} t+k \\ r_{s+1} \end{array}\right].$$

In the same argument we can find all $up(G,h)$, where $b+2 \leq h \leq n$ and the proof is completed.

Corollary 3.11. Let $G$ be a friendship graph $F_k$ with $2k+1$ vertices. Then,
$UP(G, x) = 2^k x^k + \left[ 2^k + k 2^{k-1} \right] x^{k-1} + \sum_{i=2}^{k+1} \left[ \sum_{\eta_i = \cdots = \eta_{k+1} = 2} \left( 2 \right) \left( \frac{2}{r_i} \right) \left( \frac{1}{r_{k+1}} \right) \right] x^i.$

4. Open Problems

Finally, for feature work we state the following definition.

**Definition 4.1.** Two graphs $G$ and $H$ are said to be uphill-equivalent if $UP(G, x) = UP(H, x)$. The uphill-equivalence classes of $G$ noted by $G_{up} = \{ H : H$ is uphill-equivalent to $G \}$. 

**Example 4.2.**
1. $K_n_{up} = \{ H : H$ is regular graph of $n$ vertices $\}$.
2. The windmill graph $Wd(s, k)$ and Dutch windmill graph $D(s, k)$ are uphill-equivalent.

We state the following open problems for feature work:
1) which graphs have two distinct uphill domination roots?
2) which families of graphs have only real uphill domination roots?
3) which graphs satisfy $G_{up} = \{ G \}$?
4) determine the uphill-equivalence classes for some new families of graphs.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References


Application of Stability Theory in Study of Local Dynamics of Nonlinear Systems

Tahmineh Azizi, Gabriel Kerr

Department of Mathematics, Kansas State University, Manhattan, KS, USA
Email: tazizi@ksu.edu, gdkerr@math.ksu.edu

Abstract

Investigating local dynamics of equilibrium points of nonlinear systems plays an important role in studying the behavior of dynamical systems. There are many different definitions for stable and unstable solutions in the literature. The main goal to develop stability definitions is exploring the responses or output of a system to perturbation as time approaches infinity. Due to the wide range of application of local dynamical system theory in physics, biology, economics and social science, it still attracts many researchers to play with its definitions to find out the answers for their questions. In this paper, we start with a brief review over continuous time dynamical systems modeling and then we bring useful examples to the playground. We study the local dynamics of some interesting systems and we show the local stable behavior of the system around its critical points. Moreover, we look at local dynamical behavior of famous dynamical systems, Hénon-Heiles system, Duffing oscillator and Van der Pol equation and analyze them. Finally, we discuss about the chaotic behavior of Hamiltonian systems using two different and new examples.

Keywords

Local Dynamics, Hénon-Heiles System, Duffing Oscillator, The Van Der Pol Equation, Hamiltonian Function

1. Introduction

A dynamical system describes the evolution of a system over time using a set of mathematical laws. Also, it can be used to predict the interactions between different components of a system [1] [2]. There are two main methods to model the dynamical behaviors of a system, continuous time modeling, discrete-time mod-
eling [1] [2] [3]. When the time between two measurements is negligible, the continuous time modeling governs the evolution of the system, however, when there is a gap between two measurements, discrete-time system modeling comes to play. Ordinary differential equations are the tool to model a continuous system and iterated maps represent the discrete generations [4] [5].

In this paper we will be concerned with continuous dynamical systems which are defined by differential equations. Indeed, some famous examples of dynamical systems can be written in terms of differential equations: the harmonic oscillator, the pendulum and double pendulum, or the N-body problem [4]-[9]. A dynamical system is a triple \((M, \Phi, K)\) where \(M\) is called the phase space and is usually a smooth manifold or a subset of \(\mathbb{R}^n\), \(\Phi : M \times K \rightarrow M\), called the evolution, is a smooth action of \(K\) in \(M\) and \(K\) is either a subset of \(\mathbb{R}\) in the case of a continuous time dynamical system or a subset of \(\mathbb{Z}\) in the case of a discrete time dynamical system. The smooth action \(\Phi(x)\) describes the evolution with time \(t \in K\) of a point \(x\) in the phase space \(M\) [4] [5] [6] [7].

Stability of a system is one of the most important parts of the studying the dynamical behavior of a system. Generally speaking, an unstable and also a chaotic system are not useful and we like to work with a system with stable and or periodic behavior (although chaos is a known behavior for many systems and sometimes people look for different strategies for chaotification of a system for different purposes [10] [11] [12]).

There are different definitions for stability, however, all they have this common fact that a system is stable if perturbation, external input and or intentionally applied signals cannot make the system get away from the equilibrium point [13] [14]. There are three possibilities for dynamical behavior of a system after applying a perturbation [5]:

1) The system state would return to the equilibrium state.
2) The system state would not return to the equilibrium state but stays near to that state.
3) The system state diverges from the equilibrium state.

Mathematically speaking, the equilibrium state \(x^*\) is stable if for each initial conditions \(x(0)\) close enough to \(x^*\), the corresponding trajectory \(x(t)\) remains near \(x^*\) for all \(t \geq 0\).

\[
\forall \epsilon > 0 \; \exists \delta > 0 : \|x(0) - x^*\| < \delta \; \Rightarrow \|x(t) - x^*\| < \epsilon, \; \forall t \geq 0
\]

In this paper, we present some results regarding the study of local dynamics of non-linear continuous time dynamical systems. We provide different examples to display stable and unstable limit cycles and we demonstrate the numerical results for each case. Also, we study the local dynamics of three well known physical systems, Henon-Heiles system, Duffing oscillator and Van der pol equations and we display the evolution of solutions of the system in time. Finally, we discuss about the chaos in Hamiltonian systems and we provide two examples to show chaos in Hamiltonian systems.
2. Dynamical Systems Playground

When we start to analyze the local dynamics of non-linear systems, the first step is finding the critical points and then exploring how the trajectories of the system evolving in the neighborhood of critical point. This analysis helps us to find out how other solutions or trajectories of the system behave when they get close to the critical points. Another step to analyze a dynamical system is studying the trajectories which trace out a limit cycle or a closed curve. In this case, the solution \( x(t) \) of the system will go around and create a closed curve \( C \) with a certain period \( T \). Therefore, the solutions \( x(t) = (x(t), y(t)) \) of the system when it becomes periodic change to be \( x(t+T) = x(t), y(t+T) = y(t) \) for all \( t \). Any trajectories which are close to the limit cycle \( C \) follow the same behavior as the limit cycle \( C \). For instance, they can get spiral in toward \( C \), or they can spiral away from \( C \), which demonstrates if the closed curve \( C \) is stable or unstable. See Figure 1.

The root point for \( F_1(X,Y) = (-X^2 - Y^2)\exp(-X^2 - Y^2) \) is \( (X,Y) = (0,0) \). The Taylor expansion for \( F_1(X,Y) \) at \( (X,Y) = (0,0) \) has the following form:

\[
T_a(X,Y) = -Y^2 e^{-Y^2} + X^2 e^{-X^2} (X^2 - 1) - \frac{1}{2} X^4 e^{-X^2} (Y^2 - 2) + O(X^5)
\] (1)

As it is clear from Figure 1, the point \( (0,0) \) is a stable fixed point and also the maximum of \( F_1(X,Y) \) occurs at \( (X,Y) = (0,0) \) and it is:

\[
\max \left\{ (-X^2 - Y^2)\exp(-X^2 - Y^2) \right\} = 0
\] (2)

In and Figure 2, we see \( F_2(X,Y) = (X^2 + Y^2)\exp(-X^2 - Y^2) \). The root for \( F_2(X,Y) \) is \( (X,Y) = (0,0) \). The Taylor expansion for \( F_2(X,Y) \) at \( (X,Y) = (0,0) \) has the following form:

\[
T_a(X,Y) = Y^2 e^{-Y^2} - X^2 e^{-X^2} (Y^2 - 1) + \frac{1}{2} Y^4 e^{-Y^2} (Y^2 - 2) + O(X^5)
\] (3)

The minimum of \( F_2(X,Y) \) happens for \( (X,Y) = (0,0) \) which is equal:

\[
\min \left\{ (X^2 + Y^2)\exp(-X^2 - Y^2) \right\} = 0
\] (4)

For \( F_3(X,Y) \) the point \( (X,Y) = (0,0) \) is unstable.

Another example, \( F_3(X,Y) = X \exp(-X^2 - Y^2) \) which has been displayed in Figure 3. The maximum of \( F_3(X,Y) \) occurs at \( (X,Y) = (-1,0) \) and \( (X,Y) = (1,0) \) and it equals to:

\[
\max \left\{ X^2 \exp(-X^2 - Y^2) \right\} = \frac{1}{e}
\] (5)

Also, the Taylor expansion for \( F_3(X,Y) \) at \( X = 0 \) has the following form:

\[
T_a(X,Y) = X^2 e^{-Y^2} - X^2 e^{-X^2} (Y^2 - 1) - \frac{1}{2} X^4 e^{-X^2} (Y^2 - 2) + O(X^5)
\] (6)

As we can see in Figure 3, \( (X,Y) = (-1,0) \) and \( (X,Y) = (1,0) \) are stable.

In Figure 4, we can see for \( F_4(X,Y) = Y^2 \exp(-X^2 - Y^2) \), \( (X,Y) = (0,-1) \) and \( (X,Y) = (0,1) \) are stable.
Figure 1. $F_1(X,Y) = (-X^2 - Y^2) \exp(-X^2 - Y^2)$, local dynamics.

Figure 2. $F_2(X,Y) = (X^2 + Y^2) \exp(-X^2 - Y^2)$, local dynamics.

Figure 3. $F_3(X,Y) = X \exp(-X^2 - Y^2)$, local dynamics.

Figure 4. $Z = Y^2 \exp(-X^2 - Y^2)$, local dynamics.
The maximum of \( F_4(X, Y) \) occurs at \( (X, Y) = (0,-1) \) and \( (X, Y) = (0,1) \) and equals to

\[
\max \left\{ Y^2 \exp \left( -X^2 - Y^2 \right) \right\} = \frac{1}{e} \tag{7}
\]

Also, the Taylor expansion for \( F_4(X, Y) \) at \( X = 0 \) has the form

\[
T_n(X, Y) = Y^2 e^{-Y^2} - X^2 e^{-X^2} Y^2 + \frac{1}{2} X^4 e^{-X^2} Y^2 + O(X^5) \tag{8}
\]

For \( F_5(X, Y) = (X^2 Y^2) \exp( -X^2 - Y^2 ) \) (Figure 5), \( (X, Y) = (0, 0) \) is a root and Taylor expansion for \( F_5(X, Y) \) at \( X = 0 \) has the form

\[
T_n(X, Y) = X^2 e^{-X^2} Y^2 - X^4 e^{-X^2} Y^2 + \frac{1}{2} X^6 e^{-X^2} Y^2 - \frac{1}{6} X^8 e^{-X^2} Y^2 + O(X^9) \tag{9}
\]

here, \( (X, Y) = (-1, -1) \) and \( (X, Y) = (-1, 1) \) give the maximum of \( F_5(X, Y) \) which is

\[
\max \left\{ X^2 Y^2 \exp \left( -X^2 - Y^2 \right) \right\} = \frac{1}{e^2} \tag{10}
\]

Finally, in Figure 6, \( F_6(X, Y) = -(X^2 Y^2) \exp( -X^2 - Y^2 ) \) has a root at \( (X, Y) = (0, 0) \) and Taylor expansion for \( F_6(X, Y) \) at \( X = 0 \) has the form

\[
T_n(X, Y) = -X^2 e^{-X^2} Y^2 + X^4 e^{-X^2} Y^2 - \frac{1}{2} X^6 e^{-X^2} Y^2 + \frac{1}{6} X^8 e^{-X^2} Y^2 + O(X^9) \tag{11}
\]

here, \( (X, Y) = (-1, -1) \) and \( (X, Y) = (-1, 1) \) give the minimum of \( F_6(X, Y) \) which is

\[
\text{Figure 5. } F_5(X, Y) = (X^2 Y^2) \exp( -X^2 - Y^2 ), \text{ local dynamics.}
\]

\[
\text{Figure 6. } F_6(X, Y) = -(X^2 Y^2) \exp( -X^2 - Y^2 ), \text{ local dynamics.}
\]
\[
\min \left\{ -X^2y^2 \exp\left( -X^2 - Y^2 \right) \right\} = -\frac{1}{e^2}
\]  

(12)

3. Application of Continuous Dynamical Systems Modeling

3.1. Hénon-Heiles System

The Hénon-Heiles potential is one of the simplest examples of classical mechanics and Hamiltonian systems [15] [16] [17] [18]. The Hénon-Heiles Hamiltonian demonstrates the emotion of stars around a galactic center. In 1964, Michael Hénon and Carl Heiles simplified the problem of the emotion of stars around a galactic center by using a Hamiltonian to describe the motion of stars near the equilibrium [16]. The Hénon-Heiles system has a wide application in studying chaotic dynamics in a system. If the energy of the motion becomes close to the bounding energy of the potential sink which is surrounding the center of the potential, this system displays chaotic dynamics feit 1984 wave.

Consider the following nonlinear system of ordinary differential equations

\[
\begin{align*}
\frac{dx}{dt} &= y \\
\frac{dy}{dt} &= x - x^2
\end{align*}
\]  

(13)

The Hamiltonian function for this system has the form

\[
H(x, y) = \frac{x^2}{2} - \frac{x^3}{2} + \frac{x^3}{2}
\]  

(14)

For any \(x, y\) satisfying (13), we have \(\frac{dH}{dt} = 0\). For any solution \((x(t), y(t))\) of system (13), the Hamiltonian \(H(x(t), y(t))\) is constant, it means \(\frac{d}{dt} H(x(t), y(t)) = 0\). This is a very nice property of Hamiltonian function which is a conserved quantity for a system of ordinary differential equations and it is constant along all solution curves of the system.

The solution curves are given by \(H(x, y) = C\). Here, there are two non-degenerate critical points \((0, 0)\) and \((-1, 0)\). The critical point \((0, 0)\) is a saddle point and the eigenvectors corresponding to this critical points are \((1, -1)^T\) and \((1, 1)^T\). The critical point \(-(1, 0)\) is a center. Figure 7 displays the level curves or contours of four different Hamiltonian functions.

Consider the following Hamiltonian functions

\[
\begin{align*}
H_1(x, y) &= -\frac{y^2}{2} + \frac{x^2}{2} + \frac{x^3}{2} \\
H_2(x, y) &= \frac{y^2}{2} - \frac{x^2}{2} + \frac{x^3}{2} \\
H_3(x, y) &= \frac{y^2}{2} + \frac{x^2}{2} + \frac{x^3}{2} \\
H_4(x, y) &= \frac{y^2}{2} + \frac{x^2}{2} - \frac{x^3}{2}
\end{align*}
\]  

(15)  

(16)  

(17)  

(18)
These Hamiltonian functions (15)-(18), are corresponding to different system of ordinary differential equations. The Hamiltonian function $H_1(x, y)$ has a critical point at $(1, 0)$ and the Hamiltonian function $H_2(x, y)$ has a critical point at $(-1, 0)$. As we can see in Figure 7, the stable and unstable manifolds from the origin for $H_1(x, y)$ and $H_2(x, y)$ form a homoclinic orbit which we can not see this property in Hamiltonian functions $H_3(x, y)$ and $H_4(x, y)$. This homoclinic loop connects the critical point $(0, 0)$ to itself and it takes infinite amount of time to make connection. For Hamiltonian functions $H_1(x, y)$ and $H_2(x, y)$, the critical point $(0, 0)$ is called a saddle-node equilibrium and the Jacobian matrix of the system has a zero eigenvalue at this equilibrium point. However, the critical point $(0, 0)$ for the Hamiltonian functions $H_3(x, y)$ and $H_4(x, y)$ demonstrates another kind of dynamics and it is called the Bogdanov-Takens equilibrium point and the Jacobian matrix in this case has two zero eigenvalues. As it can be seen in Figure 7, the critical point $(0, 0)$ is unstable which is the property of Bogdanov-Takens equilibrium point.

3.2. Duffing Oscillator

The Duffing oscillator is a single ordinary differential equation which represents a nonlinear damped driven oscillator. This simple nonlinear system displays different kinds of dynamical behaviors from periodic and regular behaviors to chaos. When we add a driving force and friction, we can see this simple equation exhibit chaotic behavior [19] [20] [21]. The Duffing oscillator equation has the following form

$$y'' + \alpha y' + \beta y + \gamma x^3 = 0, \quad y(0) = A, \quad y'(0) = B$$

(19)

The law of energy conservation mentions that this is impossible to see chaotic motion in a single degree of freedom. Therefore, with adding a driving force and damping, the energy conservation would be eliminated. Then, the equations of motion has the form

$$\begin{align*}
\frac{d^2y}{dt^2} &= y(2) \\
\frac{dy}{dt} &= -by(2) - \alpha y(1) - \beta y^3(1) + amp \sin(\omega t)
\end{align*}$$

(20)
We have demonstrated different dynamical behaviors of (20) in Figures 8-10. The Duffing oscillator can be used to model different physical phenomena such as stiffening springs, beam buckling, nonlinear electronic circuits, superconducting Josephson parametric amplifiers, and ionization waves in plasmas [22].

**Figure 8.** Chaotic solutions of Duffing oscillator (20) for $amp = 0.42$, $b = 0.5$, $\alpha = -1.0$, $\beta = 1.0$, $w = 1.0$. Periodic solutions of Duffing oscillator (20) for $amp = 0.35$, $b = 0.75$, $\alpha = -1.0$, $\beta = 1.0$, $w = 1.0$.

**Figure 9.** Periodic solutions of Duffing oscillator (20) for $amp = 0.45$, $b = 0.45$, $\alpha = -1.0$, $\beta = 1.0$, $w = 0.75$. Chaotic solutions of Duffing oscillator (20) for $amp = 0.4$, $b = 0.49$, $\alpha = -1.0$, $\beta = 1.0$, $w = 1.1$.

**Figure 10.** Periodic solutions of Duffing oscillator (20) for $amp = 0.43$, $b = 0.51$, $\alpha = -1.0$, $\beta = 1.0$, $w = 1.05$. Chaotic solutions of Duffing oscillator (20) for $amp = 0.39$, $b = 0.47$, $\alpha = -1.0$, $\beta = 1.0$, $w = 0.9$. 
3.3. The Van Der Pol Equation

Van der Pol equation which is a well known second order ordinary differential equation with cubic nonlinearity has attracted many researchers in different field of science. This self oscillatory system, Van der Pol oscillator, has been considered as very useful mathematical model for many complicated systems [23] [24] [25]. Mathematical representation of the Van der Pol system has the form

\[ x'' + \mu(x^2 - 1)x' + x = 0 \]  

(21)

where constant \( \mu \) is a positive parameter depending on the tube constants. This equation represents describes the current \( x(t) \) in a certain type of vacuum tube. We can write (21) as a first order system of differential equations:

\[
\begin{align*}
\frac{dx}{dt} &= y \\
\frac{dy}{dt} &= -\mu(x^2 - 1)x' - x
\end{align*}
\]

(22)

The numerical integration of Equation (22) has been represented in Figures 11-14.

As we can see, depending on different values for \( \mu \), solutions look like periodic motion. When \( \mu \) small values, this motion is nearly sinusoidal, however for larger values of \( \mu \), the solutions seem to be relaxation oscillations which means solutions are similar to a series of step functions and jump twice per cycle between the positive and the negative values.

4. Chaos in Continuous Dynamical Systems

In this section, we assume that there is a Hamiltonian function with two degrees of freedom and it is given by \( H = H_0 + \epsilon H_1 \). Here, we consider \( \epsilon \) to be a very small parameter, \( H_0 \) an integrable Hamiltonian system and \( H_1 \) makes \( H \) to be non-integrable. For \( \epsilon = 0 \) and also for \( 0 < \epsilon \ll 1 \), there exist quasi periodic cycles which are known as KAM tori. However, under perturbation, these quasi periodic cycles will be deformed and KAM tori will be dissolved gradually as we increase \( \epsilon \). This phenomenon can be observed in Figure 15 and Figure 16.

![Figure 11. Solutions of Van der Pol Equation (22) with \( \mu = 0.75 \) and \( \mu = 5 \).](image-url)
Figure 12. Solutions of Van der Pol Equation (22) with $\mu = 0.5$ and $\mu = 3$.

Figure 13. Solutions of Van der Pol Equation (22) with $\mu = 0.25$ and $\mu = 10$.

Figure 14. Solutions of Van der Pol Equation (22) with $\mu = 0.1$ and $\mu = 15$.

Figure 15. Dissolving the KAM tori caused by perturbation, $H_I(x,y) = -(1/2)x^2 + \varepsilon \sin((\pi/2)y) + \sin((\pi/2)(y - \beta))$.
According to, KAM theory when \( x \) is irrational, then the torus is preserved for small perturbation \( \varepsilon \). But, proportional tori and adjacent irrational tori would be destroyed. Also, the stable manifold and unstable manifold of the saddle point which are intersecting transversely, appear to be Smale horseshoe and chaotic motion. As \( \varepsilon \) increases gradually, these chaotic layers grow and they envelope larger area in phase space [26] [27].

5. Conclusion

Dynamic systems modeling have been frequently used to describe different physical systems and have a very important role in predicting the interactions between multiple components of a system over time. In the present study, we explored different dynamical behaviors of some continuous dynamical systems, from stable and regular motions to periodic and limit cycles, and then chaotic and irregular oscillations. We started with studying the local dynamics of some vector fields and we demonstrated the local stable behavior of the system around its critical points. We continued this paper with studying the well known problems which have been used a lot for different physical purposes. Hénon-Heiles system, Duffing oscillator and Van der Pol equation are three important dynamical examples which have been widely studied numerically. We demonstrated the stable and unstable manifolds from the origin form a homoclinic orbit in Hénon-Heiles system and we discussed about the local dynamical behaviors of its critical points. We showed that the critical point \((0,0)\) is a saddle point and critical point \((1,0)\) is a center. For Duffing oscillator, which can be used to model different physical phenomenon, we showed the periodic and chaotic motions of the system using time series. Also, for Van der Pol equation, we presented the limit cycle solutions and periodic behavior of the system. We concluded that depending on different values for \( \mu \), solutions look like periodic motion. When \( \mu \) small values, this motion is nearly sinusoidal, however for larger values of \( \mu \), the solutions seem to be relaxation oscillations, which means solutions are similar to a series of step functions and jump twice per cycle between the positive and the negative values. Finally, we discussed about the
chaos in Hamiltonian systems and we provided two interesting and different examples which exhibit chaotic behaviors. We assume a Hamiltonian function with two degrees of freedom and it can be obtained by adding an integrable Hamiltonian system and a non-integrable Hamiltonian system. We showed that for $\varepsilon = 0$ and also for $0 < \varepsilon \ll 1$, there exist quasi periodic cycles which are known as KAM tori. However, under perturbation, these quasi periodic cycles will be deformed and KAM tori will be dissolved gradually as we increase $\varepsilon$.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References


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

https://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: https://www.scirp.org
Subscription: sub@scirp.org
Advertisement: service@scirp.org