Encyclopaedia

Author and Editor

Elzein Salah Elzein

IEEE Member

- Papers in Applied Math

IEEE Publisher, UK



This Content is confidential

AppliedMath

MDPI

Article

Modeling the Impact of Misinformation on the Transmission Dynamics of COVID-19

Ziyi Su † and Ephraim Agyingi *,†

School of Mathematics and Statistics, Rochester Institute of Technology, Rochester, NY 14623, USA; zs4768@rit.edu
* Correspondence: eoasma@rit.edu
† These authors contributed equally to this work.

Abstract: The threat posed by the COVID-19 pandemic has been accompanied by an epidemic of misinformation, causing confusion and mistrust among the public. Misinformation about COVID-19 whether intentional or unintentional takes many forms, including conspiracy theories, false treatments, and inaccurate information about the origins and spread of the virus. Though the pandemic has brought to light the significant impact of misinformation on public health, mathematical modeling emerged as a valuable tool for understanding the spread of COVID-19 and the impact of public health interventions. However, there has been limited research on the mathematical modeling of the spread of misinformation related to COVID-19. In this paper, we present a mathematical model of the spread of misinformation related to COVID-19. The model highlights the challenges posed by misinformation, in that rather than focusing only on the reproduction number that drives new infections, there is an additional threshold parameter that drives the spread of misinformation. The equilibria of the model are analyzed for both local and global stability, and numerical simulations are presented. We also discuss the model's potential to develop effective strategies for combating misinformation related to COVID-19.

Keywords: COVID-19; misinformation; mathematical model

check for updates

Citation: Su, Z.; Agyingi, E. Modeling the Impact of misinformation on the Transmission Dynamics of COVID-19. *AppliedMath* **2024**, *4*, 544–560. https://doi.org/10.3390/appliedmath4020029

Academic Editor: Thomas Woolley

Received: 1 March 2024
Revised: 24 April 2024
Accepted: 26 April 2024
Published: 30 April 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

COVID-19, also known as the coronavirus disease, caused by the novel coronavirus SARS-CoV-2, has had a profound impact on global health since its emergence in late 2019. The first case was reported in Wuhan, China, and the virus rapidly spread worldwide, with multiple waves of infections observed in different regions. After more than 118,000 cases in 114 countries and 4291 deaths, the World Health Organization (WHO) declared COVID-19 a pandemic on 11 March 2020 [1,2]. COVID-19 is primarily transmitted through respiratory droplets when an infected individual coughs, sneezes, talks, or breathes heavily in close proximity to others [3]. It exhibits a wide spectrum of clinical manifestations, ranging from mild flu-like symptoms to severe respiratory distress and multi-organ failure [4,5]. Several factors including social behavior have influenced the disease's transmission dynamics. COVID-19 continues to evolve, with variations in the prevalence and severity of the disease due to emerging variants, vaccination campaigns, and public health measures implemented by governments and health organizations. Globally, as of 27 September 2023, there have been 770,875,433 confirmed cases of COVID-19, including 6,959,316 deaths, reported to the WHO [6].

During the pandemic, many countries put in place a series of measures to contain the spread of COVID-19, such as social distancing, masking, self-quarantine and eventually vaccinations when they became available, which essentially helped to regulate the transmission of the coronavirus disease. However, the lack of clear and concise messaging from public health officials also contributed to the spread of misinformation related to COVID-19. The constantly evolving nature of the pandemic and the lack of understanding of the virus

early on led to conflicting information and messages from public health officials, leading to confusion and mistrust among the public. This led some individuals to turn to alternative sources of information, including misinformation, for guidance.

Misinformation related to COVID-19 has taken various forms [7–9]. Conspiracy theories related to COVID-19 have been widely circulated, often claiming that the virus was intentionally created and spread as a bioweapon or part of a global plot [10]. For example, the theory that 5G technology is responsible for the spread of COVID-19 has been shared widely on social media platforms, despite the lack of scientific evidence to support this claim [11]. False treatments and cures for COVID-19 have also been promoted, leading to dangerous consequences for individuals who follow them. For instance, the use of disinfectants, such as bleach, to treat COVID-19 has been widely reported, leading to severe injuries and deaths [12]. False claims that certain drugs, such as hydroxychloroquine, are effective in treating COVID-19 have also been circulated, leading to hoarding and shortages of the drug, which is actually not effective in treating the virus [13]. Rumors about the origins and spread of COVID-19 have also contributed to the misinformation landscape surrounding the disease. For example, the theory that the virus was intentionally created and spread by China has been promoted by some media outlets and politicians, despite the lack of evidence to support this claim [10]. Furthermore, misinformation surrounding vaccines includes unfounded allegations of adverse effects, doubts about their efficacy, and baseless theories about secret agendas. Such misinformation undermines public trust in vaccines, leading to vaccine hesitancy and jeopardizing the achievement of widespread vaccination coverage. The politicization of the pandemic also contributed to the spread of misinformation related to COVID-19. Political leaders and media outlets have often taken polarizing stances on the pandemic, leading to the amplification of misinformation and conspiracy theories [14]. This led to a lack of consensus on the severity of the pandemic and the efficacy of public health measures, further contributing to the spread of misinformation.

Mathematical models have been used extensively to understand the spread of COVID-19 and to inform public health interventions [15]. Several of these models typically involve a set of differential equations that describe the dynamics of the virus spread, including the number of infected individuals, the rate of transmission, and the effectiveness of interventions such as social distancing and vaccination [16,17]. Mathematical models can be used to estimate key parameters of the pandemic, such as the basic reproduction number (R_0), which measures the average number of people infected by each infected individual in a susceptible population [18]. By estimating R_0, mathematical models can be used to predict the potential impact of different interventions and to guide public health policy.

There are many challenges associated with modeling COVID-19, including the uncertainty of key parameters, such as the transmission rate, the duration of infectiousness, and the duration of immunity after recovery or vaccination [19,20]. In addition, the complex dynamics of the pandemic, including the emergence of new variants and changes in public behavior due to misinformation, made it difficult to predict the course of the pandemic accurately. In this study, we will present a mathematical model of the role of misleading information in the transmission dynamics of COVID-19 since misinformed individuals were likely to ignore measures such as social distancing, masking, vaccinations and others, put in place to safeguard the population. The model is based on the assumptions that some proportion of the population is vulnerable to misinformation. We refer to misinformation as false or inaccurate information that is spread, either intentionally or unintentionally, leading to a distortion of facts or a misunderstanding of reality. It can be disseminated through various channels, including social media, news outlets, word-of-mouth, or online platforms. Misinformation can take different forms, such as misleading statements, fabricated data, conspiracy theories, or misinterpretations of legitimate information. It often aims to deceive or manipulate individuals, resulting in harmful consequences, including the erosion of trust, the promotion of fear and panic, the hindrance of informed decision-making, and potentially detrimental effects on public health, social cohesion, and democratic processes.

The purpose of this study is to assess the impact that the amount of misinformation circulating has on the transmission dynamics of COVID-19 in a population. To achieve this goal, we formulated an SIR type model with an extra compartment representing individuals that have been misled through misinformation. Because there was abundant misinformation (whether intended or unintended) associated with COVID-19, the model presented also treats the quantity of misinformation as a state labeled M based on the assumption that it is synthesized by mislead individuals. This approach is similar to target cell models where virions are produced by infected cells [21]. The equilibrium solutions of the model are obtained using computed threshold parameters and their stability established both analytically and numerically. Simulations of the model indicate that COVID-19 will significantly affect a population negatively if misinformation is not properly tackled.

2. Mathematical Model

The model presented below is an SIR model in nature with additional compartments for misinformed individuals, deaths, and the density of misleading information circulating. We assume that as the susceptible population (S) comes into contact with circulating false or inaccurate information (M) spreading among a community, some will be misled at a rate of ρ and move into the misinformed compartment (S_m). Irrespective of the medium through which it is disseminated, we assume that at any point in time (t), circulating misinformation (M) is the sum total of all forms, including misleading statements, fabricated data, conspiracy theories, or misinterpretations of legitimate information. We assume that misinformation is generated at a rate μ_0 by people that are already misinformed or bad actors who we also categorized as misinformed, and we suppose that misleading information density decays at the rate μ_1. The model therefore assumes that misleading information is only generated by individuals in the S_m compartment. We also assume that misled individuals through awareness campaigns or other sources reject false information recover at a rate κ and move back into the susceptible compartment. Interactions between susceptible (S) and misinformed (S_m) individuals with COVID-19 infected individuals (I) lead to infections, respectively, at rates α and σ. We assumed that infected individuals either recovered (R) at the rate β or died (D) from a COVID-19 related complication at the rate γ. It is further assumed new individuals are recruited only as susceptibles at some rate Λ and that the natural death rate in all compartments is θ. A schematic diagram of the model is given in Figure 1, and a summary of parameters and their description is given in Table 1.

Taking into consideration the assumptions stated above and illustrated in the schematic diagram, the system of differential equations that governs the model is as follows:

$$\frac{dS}{dt} = \Lambda + \kappa S_m - \alpha SI - \rho SM - \theta S \tag{1}$$

$$\frac{dS_m}{dt} = \rho SM - \sigma S_m I - \kappa S_m - \theta S_m \tag{2}$$

$$\frac{dI}{dt} = \alpha SI + \sigma S_m I - \beta I - \gamma I - \theta I \tag{3}$$

$$\frac{dR}{dt} = \beta I - \theta R \tag{4}$$

$$\frac{dD}{dt} = \gamma I \tag{5}$$

$$\frac{dM}{dt} = \mu_0 S_m - \mu_1 M, \tag{6}$$

where we have assumed that all the parameter values are positive and the initial conditions $S(0) = S_0$, $S_m(0) = S_{m0}$, $I(0) = I_0$, $R(0) = R_0$, $D(0) = D_0$, and $M(0) = M_0$ are all non-negative. We note that at the start of the outbreak, $S_m(0) \geq 0$ and $R(0) = D(0) = 0$.

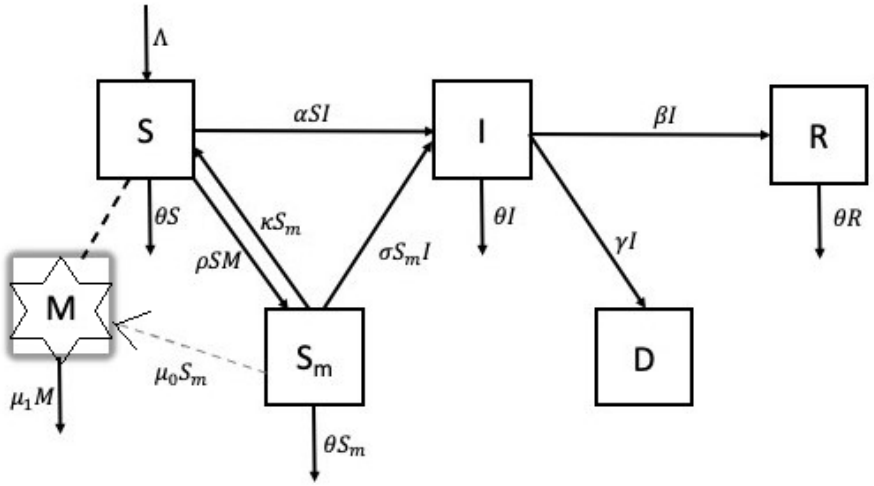


Figure 1. Schematic diagram of COVID-19 transmission with misinformation. Individuals transition from compartments S and S_m to compartment I as a result of their interaction with individuals in compartment I. Individuals transition between compartments S and S_m as a result of interaction with misinform M (either by consuming it or recovering from it). Misinformation density is produced by individuals in compartment S_m.

Table 1. Description of model parameters.

Parameter	Description
Λ	Rate at which susceptibles are recruited
α	Rate at which susceptibles are infected
β	Rate at which infectious do recover
ρ	Rate at which susceptibles are misled
κ	Rate at which misled recover from misinformation
σ	Rate at which misinformed are infected
γ	Death rate of infectious
θ	Natural death rate
μ_0	Growth rate of misleading information
μ_1	Decay rate of misleading information

Before proceeding, we observe that the equation for COVID-19 deaths (D), that is Equation (5), is decoupled from the rest of the system and therefore will be neglected in all subsequent analyses. We start by exploring the basic dynamical features of the model. In particular, we establish the positivity of solutions of the model and an invariance space. The following result shows that the model always admits positive solutions.

Theorem 1. *Let the initial conditions of the model be* $S(0) > 0$, $S_m(0) \geq 0$, $I(0) \geq 0$, $R(0) \geq 0$, *and* $M(0) \geq 0$. *Then, the solutions* $(S(t),\ S_m(t),\ I(t),\ R(t),\ M(t))$ *of the model with positive initial conditions, will remain positive for all time* $t > 0$.

Proof. We start by dividing Equations (1)–(4) and (6), respectively, by the variables S, S_m, I, R, and M. Integrating the resulting differential equations on the time interval $[0,\ t]$, and applying the initial conditions, we obtain the following expressions:

$$S(t) = S_0 \exp \int_0^t \left(\frac{\Lambda}{S(\eta)} + \kappa \frac{S_m(\eta)}{S(\eta)} - \alpha I(\eta) - \rho M(\eta) - \theta \right) d\eta,$$
$$S_m(t) = S_{m0} \exp \int_0^t \left(\rho \frac{S(\eta) M(\eta)}{S_m(\eta)} - \sigma I(\eta) - \kappa - \theta \right) d\eta,$$
$$I(t) = I_0 \exp \int_0^t (\alpha S(\eta) + \sigma S_m(\eta) - \beta - \gamma - \theta) d\eta,$$
$$R(t) = R_0 \exp \int_0^t \left(\beta \frac{I(\eta)}{R(\eta)} - \theta \right) d\eta, \text{ and } M(t) = M_0 \exp \int_0^t \left(\mu_0 \frac{S_m(\eta)}{M(\eta)} - \mu_1 \right) d\eta.$$

Since the initial conditions are non-negative, it is obvious from the above expressions that $S(t)$, $S_m(t)$, $I(t)$, $R(t)$, and $M(t)$ are always positive for all times where $t > 0$. Thus, the solution of the model with stated initial condition is always positive. □

Next, we show that the solution of the model is bounded. With the given condition, we consider the living population (L) defined by $L(t) = S(t) + S_m(t) + I(t) + R(t)$ and show that $L(t)$ and $M(t)$ are bounded in the following result.

Theorem 2. *If the initial conditions of the model satisfy* $L(0) \leq \frac{\Lambda}{\theta}$ *and* $M(0) \leq \frac{\mu_0 \Lambda}{\theta \mu_1}$, *then the solution of the model is bounded.*

Proof. Since $S(t) > 0$, $S_m(t) \geq 0$, $I(t) \geq 0$, and $R(t) \geq 0$ from Theorem 1, it is obvious that $L(t)$ is bounded from below, and therefore, $L(t) \geq 0$. It only remains to establish that $L(t)$ is bounded from above. Differentiating $L(t) = S(t) + S_m(t) + I(t) + R(t)$ and substituting the derivatives from the model, we obtain the differential equation

$$\frac{dL}{dt} = \Lambda - \theta(S + S_m + I + R) - \gamma I.$$

from which we obtain the differential inequality

$$\frac{dL}{dt} \leq \Lambda - \theta L$$

whose solution satisfies

$$L(t) \leq \frac{\Lambda}{\theta} - (\frac{\Lambda}{\theta} - L(0))e^{-\theta t}.$$

It is evident from the above inequality that the living population of the system is always bounded from above. Further, we observe from the inequality that if the initial condition $L(0) \leq \frac{\Lambda}{\theta}$, then we have that $L(t) \leq \frac{\Lambda}{\theta}$ as $t \to \infty$ is a strict upper bound.

To conclude boundedness of the model, we also established that the misinformation density $M(t)$ is bounded. We already showed in Theorem 1 that $M(t)$ is always positive, that is, $M(t) \geq 0$, thus it is bounded from below. For boundedness from above, we consider the case when $L(0) \leq \frac{\Lambda}{\theta}$. Since $L(t) \leq \frac{\Lambda}{\theta}$, we must have that $S_m(t) \leq \frac{\Lambda}{\theta}$. Using the differential equation for misinformation density, we obtain the differential inequality

$$\frac{dM}{dt} \leq \frac{\mu_0 \Lambda}{\theta} - \mu_1 M$$

whose solution $M(t)$ satisfies

$$M(t) \leq \frac{\mu_0 \Lambda}{\mu_1 \theta} - (\frac{\mu_0 \Lambda}{\mu_1 \theta} - M(0))e^{-\mu_1 t}.$$

Clearly, the information density is bounded from above. Similarly, if $M(0) \leq \frac{\mu_0 \Lambda}{\theta \mu_1}$, then we have that $M(t) \leq \frac{\mu_0 \Lambda}{\theta \mu_1}$ as a strict upper bound. □

Lemma 1. *The closed set*

$$\Pi = \left\{ (S(t), S_m(t), I(t), R(t), M(t)) \in \mathbb{R}^5_+;\ L(t) \leq \frac{\Lambda}{\theta},\ M(t) \leq \frac{\mu_0 \Lambda}{\theta \mu_1} \right\}$$

is positively invariant and attracting for the model.

Proof. The proof follows directly from Theorems 1 and 2 where the trivial existence and uniqueness of solution is assumed. □

3. Analysis of the Model

In this section, we compute the equilibria and provide conditions for their existence. We also provide the reproduction number of the model and use it to study the stabilty of the computed equilibria. Note also that the equation for the recovered (R) is decoupled from the rest of the system, in addition to that for deaths (D). We will therefore ignore these equations in the subsequent analysis given that their behavior can easily be deduced given the behavior of the remaining sub-system.

3.1. Existence of Equilibria

By decoupling the equations for R and D, the equilibria of the model is given by the solution of the sub-system

$$\begin{cases} \Lambda + \kappa S_m^* - \alpha S^* I^* - \rho S^* M^* - \theta S^* = 0 \\ \rho S^* M^* - \sigma S_m^* I^* - \kappa S_m^* - \theta S_m^* = 0 \\ \alpha S^* I^* + \sigma S_m^* I^* - \beta I^* - \gamma I^* - \theta I^* = 0 \\ \mu_0 S_m^* - \mu_1 M^* = 0. \end{cases}$$

Since the last equation above only depends on variable S_m^*, M^* can be represented as

$$M^* = \frac{\mu_0 S_m^*}{\mu_1}.$$

From the third equation in the sub-system, we see that either $I = 0$ which leads to the disease-free equilibrium (DFE) of the model, or $\alpha S^* + \sigma S_m^* - \beta - \gamma - \theta = 0$, which leads to the Endemic Equilibrium.

Suppose that $E_0 = (S^*, S_m^*, I^*, M^*)$ denotes a DFE, then further computation yields two DFEs,

$$E_0^1 = \left(\frac{\Lambda}{\theta}, 0, 0, 0\right) \quad \text{and} \quad E_0^2 = \left(\frac{(\theta+\kappa)\mu_1}{\rho\mu_0}, \frac{\Lambda}{\theta} - \frac{(\theta+\kappa)\mu_1}{\rho\mu_0}, 0, \frac{\Lambda\mu_0}{\theta\mu_1} - \frac{(\theta+\kappa)}{\rho}\right).$$

We note here that the second DFE E_0^2 will only exist if $\frac{\Lambda}{\theta} > \frac{(\theta+\kappa)\mu_1}{\rho\mu_0}$.

Before going further into the analysis, we pause and compute the reproduction number, R_0, of the model. The reproduction number is the number of secondary infections that occur when an infectious person is put into a completely susceptible population. Using the next generation matrix method (see for example Duatel et al. [22]) and the DFEs, E_0^1 and E_0^2, respectively, we obtain the reproduction numbers

$$R_0^1 = \frac{\alpha\Lambda}{\theta(\beta+\gamma+\theta)} \quad \text{and} \quad R_0^2 = \frac{1}{\beta+\gamma+\theta}\left(\alpha\left(\frac{(\theta+\kappa)\mu_1}{\rho\mu_0}\right) + \sigma\left(\frac{\Lambda}{\theta} - \frac{(\theta+\kappa)\mu_1}{\rho\mu_0}\right)\right).$$

We remark here that R_0^1 regulates the model in the absence of misinformation, while R_0^2 drives the model when misinformation is persistent.

We now turn our attention to establish the endemic equilibria of the model. Suppose that $EE = (S^*, S_m^*, I^*, M^*)$ denotes an endemic equilibrium, then using equation $\alpha S^* + \sigma S_m^* - \beta - \gamma - \theta = 0$ and the necessary equations in the sub-system, we obtain two endemic equilibria namely,

$$EE_1 = \left(\frac{A}{\alpha}, 0, \frac{\theta}{\alpha}(R_0^1 - 1), 0\right)$$

and

$$EE_2 = \left(\frac{(\theta+\kappa+\sigma I^*)\mu_1}{\rho\mu_0}, \frac{A}{\sigma} - \frac{\alpha(\theta+\kappa+\sigma I^*)\mu_1}{\sigma\rho\mu_0}, I^*, \frac{\mu_0}{\mu_1}\left(\frac{A}{\sigma} - \frac{\alpha(\theta+\kappa+\sigma I^*)\mu_1}{\sigma\rho\mu_0}\right)\right),$$

where $A = \beta + \gamma + \theta$ and where

$$I^* = (R_0^2 - 1)\frac{\theta A}{\sigma}\left(A + \frac{\theta\alpha\mu_1}{\rho\mu_0}(\frac{\sigma}{\alpha} - 1)\right)^{-1}.$$

We also remark here that the first endemic equilibrium EE_1 exists if $R_0^1 > 1$, and the second endemic equilibrium EE_2 exists if $R_0^2 > 1$ and $A - \frac{\alpha(\theta+\kappa+\sigma I^*)\mu_1}{\rho\mu_0} > 0$.

3.2. Stability of Equilibria

In this section, we establish local stability results for the DFEs and endemic equilibria obtained above under appropriate conditions. We begin by noting that the Jacobian matrix, $\mathbb{J}(S, S_m, I, M)$, of the sub-system is given by

$$\mathbb{J}(\cdot) = \begin{bmatrix} -\alpha I - \rho M - \theta & \kappa & -\alpha S & -\rho\, S \\ \rho M & -\sigma I - \kappa - \theta & -\sigma S_m & \rho\, S \\ \alpha I & \sigma I & \alpha S + \sigma S_m - \beta - \gamma - \theta & 0 \\ 0 & \mu_0 & 0 & -\mu_1 \end{bmatrix}.$$

Theorem 3. *If $\mu_1\theta(\theta + \kappa) > \mu_0\rho\Lambda$ and $R_0^1 < 1$, then the DFE E_0^1 is locally asymptotically stable, otherwise it is unstable.*

Proof. Evaluating the Jacobian matrix at the DFE $E_0^1 = \left(\frac{\Lambda}{\theta}, 0, 0, 0\right)$, we obtain

$$\mathbb{J}(E_0^1) = \begin{bmatrix} -\theta & \kappa & -\alpha\frac{\Lambda}{\theta} & -\rho\frac{\Lambda}{\theta} \\ 0 & -\kappa - \theta & 0 & \rho\frac{\Lambda}{\theta} \\ 0 & 0 & \alpha\frac{\Lambda}{\theta} - \beta - \gamma - \theta & 0 \\ 0 & \mu_0 & 0 & -\mu_1 \end{bmatrix}$$

whose characteristic polynomial is

$$\begin{aligned} P(\lambda) &= \det(\mathbb{J}(E_0^1) - \lambda I) \\ &= (\alpha\frac{\Lambda}{\theta} - \beta - \gamma - \theta - \lambda)[(-\theta - \lambda)(-\theta - \kappa - \lambda)(-\mu_1 - \lambda) - \mu_0(\rho\frac{\Lambda}{\theta})(-\theta - \lambda)], \end{aligned}$$

yeilding the eigenvalues

$$\begin{aligned} \lambda_1 &= -\theta \\ \lambda_2 &= \alpha\frac{\Lambda}{\theta} - \beta - \gamma - \theta \\ \lambda_3 &= \frac{-(\theta + \kappa + \mu_1) - \sqrt{(\mu_1 + \theta + \kappa)^2 - 4((\theta + \kappa)\mu_1 - \mu_0\rho\frac{\Lambda}{\theta})}}{2} \\ \lambda_4 &= \frac{-(\theta + \kappa + \mu_1) + \sqrt{(\mu_1 + \theta + \kappa)^2 - 4((\theta + \kappa)\mu_1 - \mu_0\rho\frac{\Lambda}{\theta})}}{2}. \end{aligned}$$

From the above, it is obvious that λ_1 and λ_3 have negative real parts, since $R_0^1 = \frac{\alpha\Lambda}{\theta(\beta+\gamma+\theta)}$ and λ_2 can be rewritten as

$$\begin{aligned} \lambda_2 &= \alpha\frac{\Lambda}{\theta} - (\beta + \gamma + \theta) \\ &= R_0^1(\beta + \gamma + \theta) - (\beta + \gamma + \theta) \\ &= (R_0^1 - 1)(\beta + \gamma + \theta). \end{aligned}$$

Thus, if $R_0^1 < 1$, then $\lambda_2 < 0$. λ_4 has a negative real part if

$$\begin{aligned}
\theta + \kappa + \mu_1 &> \sqrt{(\mu_1 + \theta + \kappa)^2 - 4((\theta + \kappa)\mu_1 - \mu_0\rho\frac{\Lambda}{\theta})} \\
(\theta + \kappa + \mu_1)^2 &> (\mu_1 + \theta + \kappa)^2 - 4((\theta + \kappa)\mu_1 - \mu_0\rho\frac{\Lambda}{\theta}) \\
0 &> -4((\theta + \kappa)\mu_1 - \mu_0\rho\frac{\Lambda}{\theta}) \\
\mu_1\theta(\theta + \kappa) &> \mu_0\rho\Lambda
\end{aligned}$$

holds true. Therefore, if $\mu_1\theta(\theta + \kappa) > \mu_0\rho\Lambda$ and $R_0^1 < 1$, then the DFE E_0^1 is locally asymptotically stable, otherwise it is unstable if $R_0^1 > 1$. □

We now conclude our analysis of the first DFE E_0^1 by establishing a global stability result. For this purpose, we recall establishing in Section 2 that at any time t, the living population $L(t) = S(t) + S_m(t) + I(t) + R(t)$ is bounded above and satisfies

$$L(t) \leq \frac{\Lambda}{\theta} - (\frac{\Lambda}{\theta} - L_0)e^{-\theta t},$$

where $L_0 = L(0)$. The following results are a consequence of the above inequality.

Lemma 2. *The living population $L(t) \leq L_0$ at all times t.*

The following result provides global stability for the first DFE E_0^1, that is, when there is no COVID-19 and there is also no misinformation.

Theorem 4. *In view of Lemma 2, if $L_0 = \frac{\alpha}{\sigma}\frac{\Lambda}{\theta}$, and if $R_0^1 < 1$, then the first disease-free equilibrium, E_0^1 is globally asymptotically stable, otherwise it is unstable.*

Proof. In order to prove Theorem 4, we consider Lyapunov function candidate $V = \frac{1}{2}I^2$. Taking its derivative we obtain,

$$\begin{aligned}
\frac{dV}{dt} &= I\frac{dI}{dt} \\
&= (\alpha SI + \sigma S_m I - \beta I - \gamma I - \theta I)I \\
&= [\alpha S + \sigma S_m - (\beta + \gamma + \theta)]I^2 \\
&= (\beta + \gamma + \theta)(\frac{\alpha S + \sigma S_m}{\beta + \gamma + \theta} - 1)I^2.
\end{aligned}$$

Since the model assumes that $\alpha < \sigma$, then the above equation becomes

$$\frac{dV}{dt} \leq (\beta + \gamma + \theta)(\frac{\sigma(S + S_m)}{\beta + \gamma + \theta} - 1)I^2.$$

If $L_0 = \frac{\alpha}{\sigma}\frac{\Lambda}{\theta}$ then by Lemma 2, we obtain that $S + S_m \leq \frac{\alpha}{\sigma}\frac{\Lambda}{\theta}$. Thus, we obtain the differential inequality

$$\begin{aligned}
\frac{dV}{dt} &\leq (\beta + \gamma + \theta)(\frac{\alpha\Lambda}{\theta(\beta + \gamma + \theta)} - 1)I^2 \\
&\leq (\beta + \gamma + \theta)(R_0^1 - 1)I^2.
\end{aligned}$$

It is clear that if $R_0^1 < 1$, then $\frac{dV}{dt} < 0$ and therefore, by LaSalle's invariance principle the disease-free equilibrium E_0^1 is globally asymptotically stable [23]. □

Next, we consider the local stability of the second DFE E_0^2 and state the following result.

Theorem 5. *If $\rho\Lambda\mu_0 > \theta(\theta+\kappa)\mu_1$ and $R_0^2 < 1$, then the DFE E_0^2 is locally asymptotically stable, otherwise it is unstable.*

Proof. Evaluating the Jacobian matrix at the the second DFE, that is, $E_0^2 = (\frac{(\theta+\kappa)\mu_1}{\rho\mu_0}, \frac{\Lambda}{\theta} - \frac{(\theta+\kappa)\mu_1}{\rho\mu_0}, 0, \frac{\Lambda\mu_0}{\theta\mu_1} - \frac{(\theta+\kappa)}{\rho})$, we obtain the Jacobian matrix

$$\mathbb{J}(E_0^2) = \begin{bmatrix} -\frac{\rho\Lambda\mu_0}{\theta\mu_1} + \kappa & \kappa & -\alpha\frac{(\theta+\kappa)\mu_1}{\rho\mu_0} & -\rho\frac{(\theta+\kappa)\mu_1}{\rho\mu_0} \\ \frac{\rho\Lambda\mu_0}{\theta\mu_1} - \theta - \kappa & -\kappa-\theta & -\sigma(\frac{\Lambda}{\theta} - \frac{(\theta+\kappa)\mu_1}{\rho\mu_0}) & \rho\frac{(\theta+\kappa)\mu_1}{\rho\mu_0} \\ 0 & 0 & \alpha\frac{(\theta+\kappa)\mu_1}{\rho\mu_0} + \sigma(\frac{\Lambda}{\theta} - \frac{(\theta+\kappa)\mu_1}{\rho\mu_0}) - \beta - \gamma - \theta & 0 \\ 0 & \mu_0 & 0 & -\mu_1 \end{bmatrix}.$$

Clearly, we see from the third row of the above matrix that

$$\begin{aligned} \lambda_1 &= \alpha\frac{\theta\mu_1}{\rho\mu_0} + \sigma(\frac{\Lambda}{\theta} - \frac{\theta\mu_1}{\rho\mu_0}) - \beta - \gamma - \theta \\ &= R_0^2(\beta+\gamma+\theta) - (\beta+\gamma+\theta) \\ &= (R_0^2 - 1)(\beta+\gamma+\theta) < 0 \text{ if } R_0^2 < 1. \end{aligned}$$

The remaining factor of the characteristic polynomial is of degree three and is given by

$$\lambda^3 + b_1\lambda^2 + b_2\lambda + b_3 = 0,$$

where

$$\begin{aligned} b_1 &= \mu_1 + \frac{\rho\Lambda\mu_0}{\theta\mu_1} + \theta \\ b_2 &= \frac{\rho\Lambda\mu_0}{\theta} + \frac{\rho\Lambda\mu_0}{\mu_1} - \kappa\mu_1 \\ b_3 &= \rho\Lambda\mu_0 - (\theta+\kappa)\mu_1\theta. \end{aligned}$$

To establish that the remaining eigenvalues (that is the zeros of the above cubic equation) have negative real parts, we use Routh–Hurwitz criterion. It is obvious from the above coefficients b_1, b_2, and b_3 that if the condition $\rho\Lambda\mu_0 > (\theta+\kappa)\mu_1\theta$ is satisfied then $b_1 > 0$, $b_2 > 0$, and $b_3 > 0$. Further, we also see that the inequality

$$b_1b_2 = (\mu_1 + \frac{\rho\Lambda\mu_0}{\theta\mu_1} + \theta)(\frac{\rho\Lambda\mu_0}{\theta} + \frac{\rho\Lambda\mu_0}{\mu_1} - \kappa\mu_1) > \rho\Lambda\mu_0 - (\theta+\kappa)\mu_1\theta = b_3$$

holds true. Since b_1, b_2, and b_3 are all positive and $b_1b_2 > b_3$, we conclude using Routh–Hurwitz criterion that the remaining eigenvalues have negative real parts. Therefore, if $R_0^2 < 1$, the DFE, E_0^2, is locally asymptotically stable, otherwise, it is unstable. □

Remark 1. *Just as a reminder, we recall here that the condition $\rho\Lambda\mu_0 > (\theta+\kappa)\mu_1\theta$ was imposed to ensure that the DFE E_0^2 is valid.*

Theorem 6. *If $R_0^1 > 1$, then the Endemic Equilibrium, EE_1, is locally asymptotically stable, otherwise it is unstable.*

Proof. Evaluating the Jacobian matrix at the endemic equilibrium EE_1, that is, at $EE_1 = (\frac{A}{\alpha}, 0, \frac{\theta}{\alpha}(R_0^1 - 1), 0)$ where $A = \beta + \gamma + \theta$, we obtain the Jacobian matrix

$$\mathbb{J}(EE_1) = \begin{bmatrix} -\theta R_0^1 & \kappa & -A & -\rho\frac{A}{\alpha} \\ 0 & -\sigma(\frac{\theta}{\alpha}(R_0^1 - 1)) - \kappa & 0 & \rho\frac{A}{\alpha} \\ \theta R_0^1 & \sigma(\frac{\theta}{\alpha}(R_0^1 - 1)) & 0 & 0 \\ 0 & \mu_0 & 0 & -\mu_1 \end{bmatrix}.$$

The characteristic polynomial is

$$\begin{aligned} P(\lambda) &= \det(\mathbb{J}(EE_1) - \lambda I) \\ &= (-\sigma(\frac{\theta}{\alpha}(R_0^1 - 1)) - \lambda)[(-\mu_1 - \lambda)(\theta R_0^1 \lambda + \lambda^2 + A\theta R_0^1)], \end{aligned}$$

and the eigenvalues are

$$\begin{aligned} \lambda_1 &= -\sigma(\frac{\theta}{\alpha}(R_0^1 - 1)) < 0 \text{ if } R_0^1 > 1 \\ \lambda_2 &= -\mu_1 < 0 \\ \lambda_3 &= \frac{-\theta R_0^1 - \sqrt{(\theta R_0^1)^2 - 4A\theta R_0^1}}{2} \\ \lambda_4 &= \frac{-\theta R_0^1 + \sqrt{(\theta R_0^1)^2 - 4A\theta R_0^1}}{2}. \end{aligned}$$

From the results above, the eigenvalues λ_3 and λ_4 clearly are always negative real numbers or have negative real parts, since $\sqrt{(\theta R_0^1)^2 - 4A\theta R_0^1} < \theta R_0^1$. Therefore, we can conclude that if $R_0^1 > 1$, then the endemic equilibrium, EE_1, is locally asymptotically stable and unstable if $R_0^1 \leq 1$. □

Remark 2. *We note here that in a similar manner to Theorem 6, it can be established that if $R_0^2 > 1$, then the endemic equilibrium*

$$EE_2 = \left(S^*, \frac{A - \alpha S^*}{\sigma}, I^*, \frac{\mu_0}{\mu_1}(\frac{A - \alpha S^*}{\sigma})\right),$$

where $S^* = \frac{(\theta + \kappa + \sigma I^*)\mu_1}{\rho\mu_0}$, $I^* = (R_0^2 - 1)\frac{\theta A}{\sigma}\left(A + \frac{\theta\alpha\mu_1}{\rho\mu_0}(\frac{\sigma}{\alpha} - 1)\right)^{-1}$, and $A = \beta + \gamma + \theta$ is locally stable.

4. Results and Discussion

In this section, we present numerical simulations of the model under different circumstances focusing on the population of the United States. We simulate the number of infections for different rates related to misinformation and in some instances the number of cumulative deaths. The parameter values used in the simulations were chosen based on values from the published literature for COVID-19 [20,24–26] where available and other parameter values were assumed to explore the behavior of the model. The initial conditions were chosen to reflect the United States population of approximately 335 million by setting $S(0) = 3.35 \times 10^8$. Considering that the true population of the US is estimated to range between 331 million and 341 million, we carried out several simulations with small perturbations in the initial conditions and the long term behavior of the model's predictions did not change.

In the previous section, we established several qualitative results relating to the equilibria of the model. We start by confirming these results numerically in Figures 2–4.

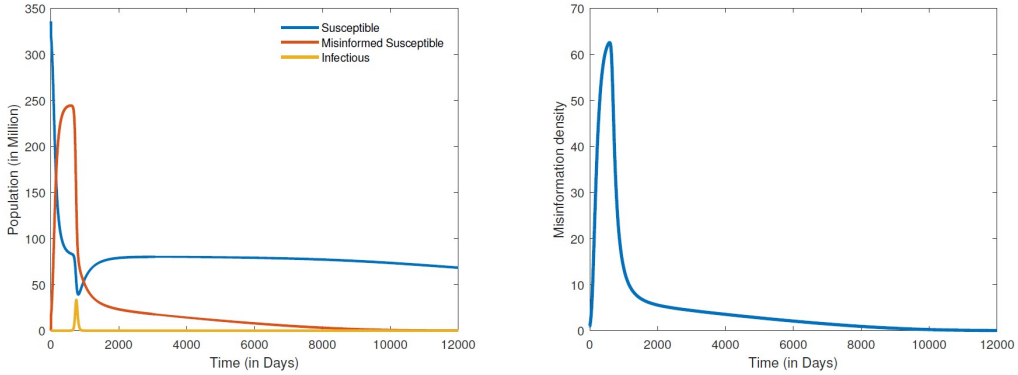


Figure 2. Numerical simulation of the first DFE: populations distribution (**left**) and information density (**right**), for parameter values $\Lambda = 0$, $\alpha = 0.25 \times 10^{-4}$, $\beta = 1/14$, $\rho = 0.1$, $\kappa = 20\rho$, $\sigma = 0.5 \times 10^{-3}$, $\gamma = 10^{-4}$, $\theta = 4 \times 10^{-5}$, $\mu_0 = 0.0025$, and $\mu_1 = 0.01$.

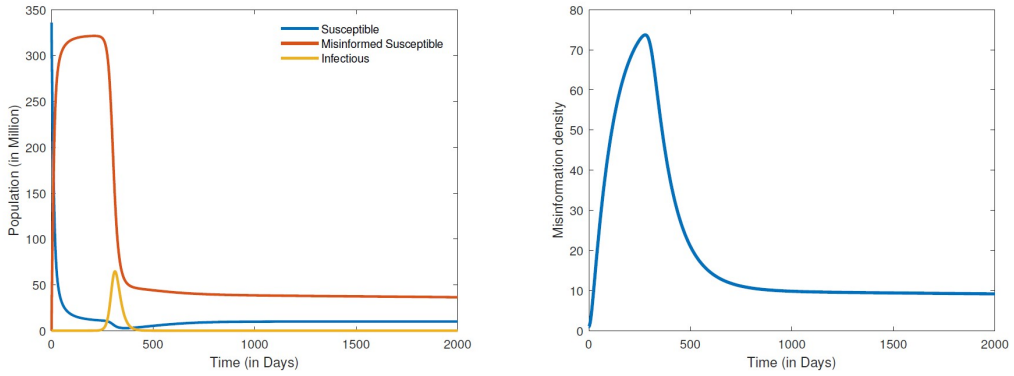


Figure 3. Numerical simulation of the second DFE: populations distribution (**left**) and information density (**right**), for parameter values $\Lambda = 0$, $\alpha = 0.25 \times 10^{-4}$, $\beta = 1/14$, $\rho = 0.1$, $\kappa = 2.5\rho$, $\sigma = 0.5 \times 10^{-3}$, $\gamma = 10^{-4}$, $\theta = 4 \times 10^{-5}$, $\mu_0 = 0.0025$, and $\mu_1 = 0.01$.

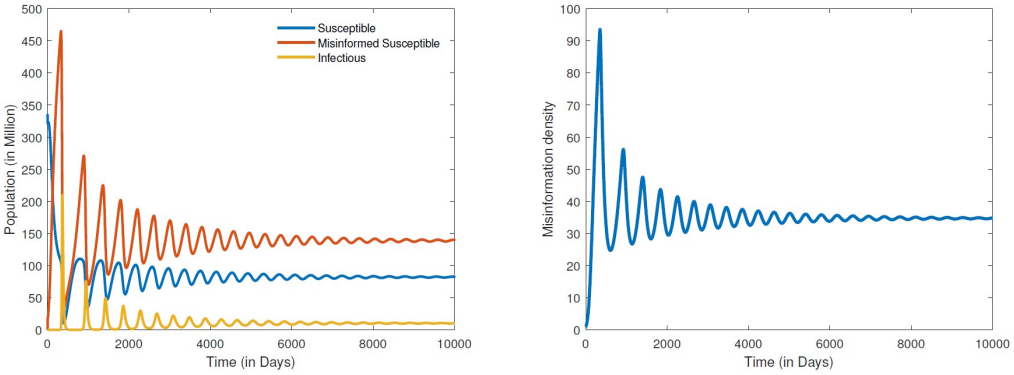


Figure 4. Numerical simulation of a stable endemic equilibrium: populations distribution (**left**) and information density (**right**), for parameter values $\Lambda = 0.75$, $\alpha = 0.25 \times 10^{-4}$, $\beta = 1/14$, $\rho = 0.1$, $\kappa = 20\rho$, $\sigma = 0.5 \times 10^{-3}$, $\gamma = 10^{-4}$, $\theta = 4 \times 10^{-5}$, $\mu_0 = 0.0025$, and $\mu_1 = 0.01$.

The results in Figure 2 demonstrate the global stability of the first DFE. The simulations show the disease, the misinformed and the density of misinformation eventually dying out with the passage of time. It is important to observe from these results that though COVID-19 may die out much earlier, it takes much longer for both the misinformed and misinformation density to be eradicated. We note here that the numerical computations for the first DFE were obtained by setting the recovery rate of misinformed individuals (κ)

to about 20 times the rate at which susceptible people are misled (ρ). When the recovery rate of misinformed individuals is set at about two times the rate at which the susceptible become misled, we obtained the simulations given in Figure 3. Clearly, the results in Figure 3 demonstrate the second DFE that we also established in the previous section. These results show that even if COVID-19 completely dies out, misinformation about the disease will never go away. Conspiracy theories such as the origin of COVID-19, the disease being a bioweapon, misconceptions about COVID-19 vaccines, and COVID-19 deniers are likely to persist for a long while given the ease and speed at which misinformation is currently disseminated.

The results for the disease-free equilibria discussed above were obtained by assuming a closed population ($\Lambda = 0$), that is, no new individuals are entering the community. This is a scenario where there is either strict lockdown or restriction of movement. This was indeed the case with the early variants of COVID-19 where cities/countries were able to restrain the spread of the disease through strict lockdowns; however, misinformation could not be contained. By shifting the paradigm from a closed population to an open one where there is freedom of movement with people entering and leaving the community, the dynamics quickly shift to simulations of an endemic equilibrium. The results in Figure 4 show a globally stable endemic equilibrium. Recall that setting the recovery rate (κ) of misinformed individuals to about 20 times the rate at which susceptible people are misinformed (ρ) led to a DFE. However, for an open population where the recruitment rate $\Lambda = 0.75$ and $\kappa = 20\rho$, we obtain an endemic equilibrium Figure 4. A further increase in the recovery rate of misinformed individuals say by a multiple of 50 times the rate at which susceptible people are misinformed, that is, $\kappa = 50\rho$, leads to a hopf bifurcation that renders the endemic equilibrium unstable as demonstrated in Figure 5. It is worth noting that the bifurcation here transforms a stable equilibrium to an unstable one containing a very large amplitude of oscillations. We can therefore infer from the results in Figures 4 and 5 that it is not enough to refute misleading information through fact checking by the media. Refuting and also correctly educating the public about COVID-19 should be carried out simultaneously. Media literacy is also essential in combating misinformation related to COVID-19. Increased media literacy can help individuals identify and distinguish between credible sources of information and misinformation. An educated public that is less susceptible to being misled will lead to a better outcome in eradicating COVID-19 or controlling the spread of the disease in an endemic state. We remark here that all simulations that we carried out for a positive recruitment rate, that is $\Lambda > 0$ (irrespective of how small the value of Λ), led to an endemic equilibrium. As a consequence, considering the fact that we live in an open population with freedom of movement, COVID-19 will become endemic.

Next, we turn our attention to simulate the effect that varying the recovery rate of misinformed individuals (κ) will have on the number of COVID-19 cases and cumulative deaths. The results for the number of COVID-19 cases for a closed population are depicted in Figure 6. We observe here that an increase in the recovering rate of the misinformed (κ) leads to fewer cases of COVID-19. The larger the recovery rate (κ), the further it takes for the number of cases to peak, and the magnitude of the peak is also smaller. We also note that because it takes longer to peak given a larger recovery rate (κ), precious time is afforded that can be used by public health officials to implement other measures so as to further reduce the impact of the disease on the entire population. The results in Figure 7 show the cumulative deaths from COVID-19 for different misinformed recovery rates for a closed population. It is also evident from these simulations that a higher recovery rate κ will lead to a lower number of cumulative COVID-19 deaths. According to the WHO [6], there were about 1,127,252 COVID-19 deaths reported in the United States between 3 January 2020 and 21 September 2023. We observe from the cumulative death simulations in Figure 7 that the rate at which misinformed people recover for the United States is less than five times the rate at which people were misled. Considering that misinformed people contribute to the growth of misinformation, the United States recovery rate of misled individuals appears to be small.

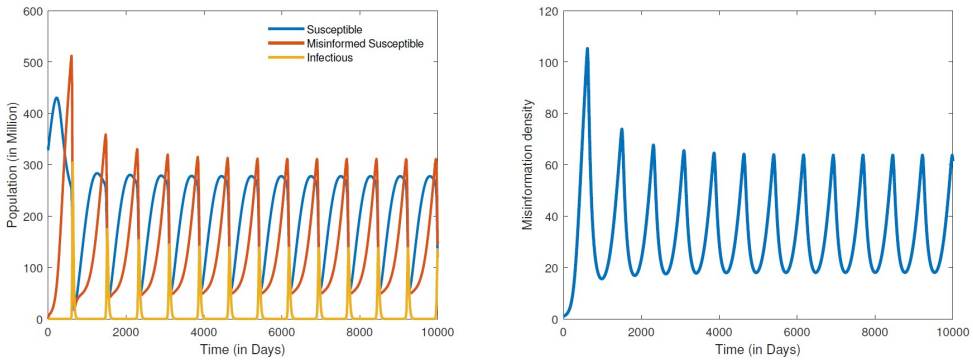


Figure 5. Numerical simulation of an unstable endemic equilibrium: populations distribution (**left**) and information density (**right**), for parameter values $\Lambda = 0.75$, $\alpha = 0.25 \times 10^{-4}$, $\beta = 1/14$, $\rho = 0.1$, $\kappa = 50\rho$, $\sigma = 0.5 \times 10^{-3}$, $\gamma = 10^{-4}$, $\theta = 4 \times 10^{-5}$, $\mu_0 = 0.0025$, and $\mu_1 = 0.01$.

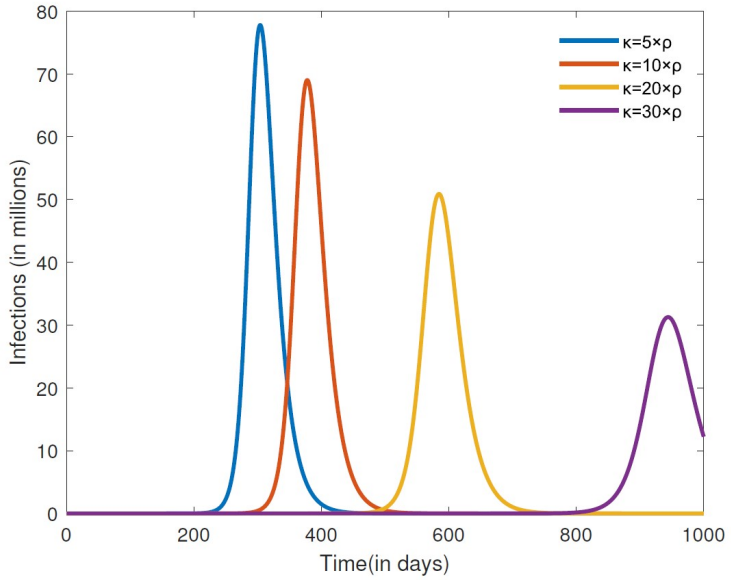


Figure 6. Numerical simulation of the number of infections for different misinformed recovery rates κ. The other parameter values are fixed at $\Lambda = 0$, $\alpha = 0.25 \times 10^{-4}$, $\beta = 1/14$, $\rho = 0.1$, $\sigma = 0.5 \times 10^{-3}$, $\gamma = 10^{-4}$, $\theta = 4 \times 10^{-5}$, $\mu_0 = 0.0025$, and $\mu_1 = 0.01$.

Finally, we consider the effects of misinformation density growth and decay rates on the number of COVID-19 cases, focusing on a closed population. The results in Figure 8 are for different growth rates μ_0, while those in Figure 9 are for different decay rates μ_1. We can clearly see from the results in Figure 8 that lowering the growth rate of misinformation density leads to fewer cases of COVID-19. Indeed, keeping the growth rate low also leads to a smaller infection peak that occurs further down the road, thus affording time that can be used to implement other containment strategies and consequently minimize the disease burden. Promoting accurate information is essential in combating misinformation related to COVID-19. Public health officials and credible sources of information should provide clear and concise messaging about the virus and public health measures, including vaccination. Positive messaging slows down the growth rate of misinformation. On the other hand, increasing the decay rate of misinformation likewise leads to a smaller infection peak that occurs much later in time. Fact-checking mechanisms can be implemented to combat misinformation related to COVID-19. TV, radio, and social media platforms can implement fact-checking mechanisms to flag and remove misinformation related to the disease.

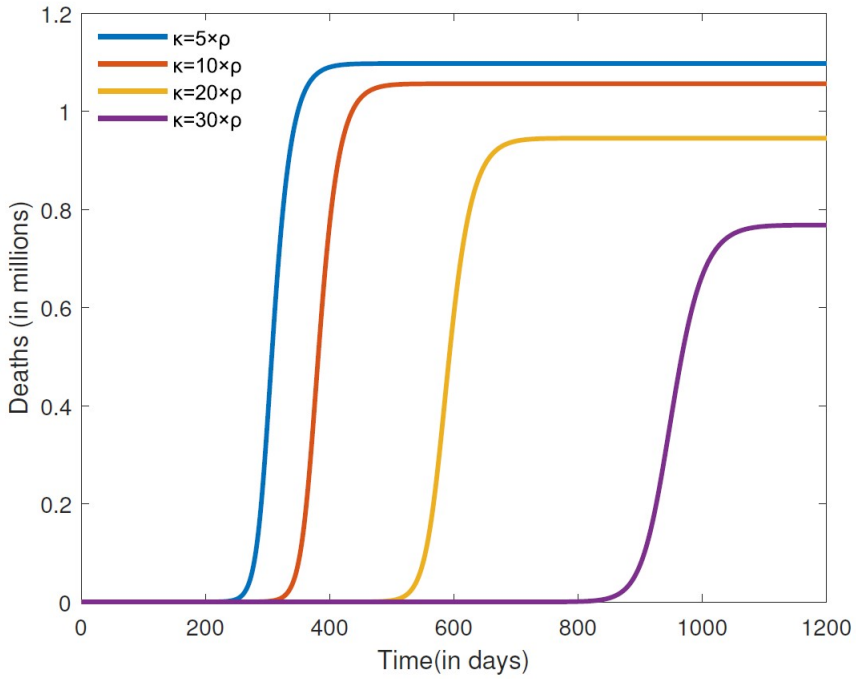


Figure 7. Numerical simulation cumulative deaths for different recovery rates κ. The other parameter values are fixed at $\Lambda = 0$, $\alpha = 0.25 \times 10^{-4}$, $\beta = 1/14$, $\rho = 0.1$, $\sigma = 0.5 \times 10^{-3}$, $\gamma = 10^{-4}$, $\theta = 4 \times 10^{-5}$, $\mu_0 = 0.0025$, and $\mu_1 = 0.01$.

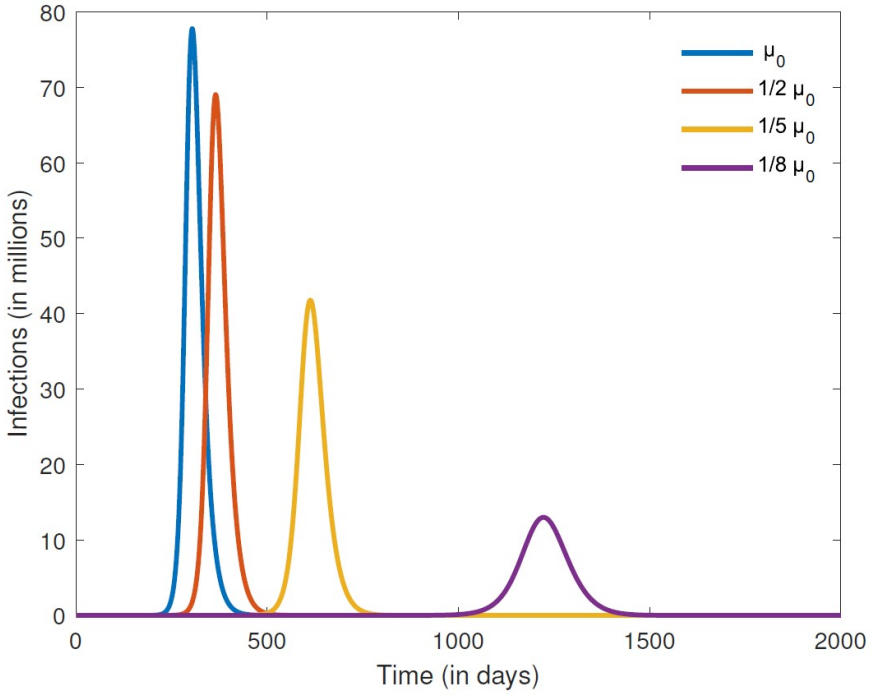


Figure 8. Numerical simulation of the number of infections for different misinformation growth rates μ_0. The other parameter values are fixed at $\Lambda = 0$, $\alpha = 0.25 \times 10^{-4}$, $\beta = 1/14$, $\rho = 0.1$, $\kappa = 5\rho$, $\sigma = 0.5 \times 10^{-3}$, $\gamma = 10^{-4}$, $\theta = 4 \times 10^{-5}$, and $\mu_1 = 0.01$.

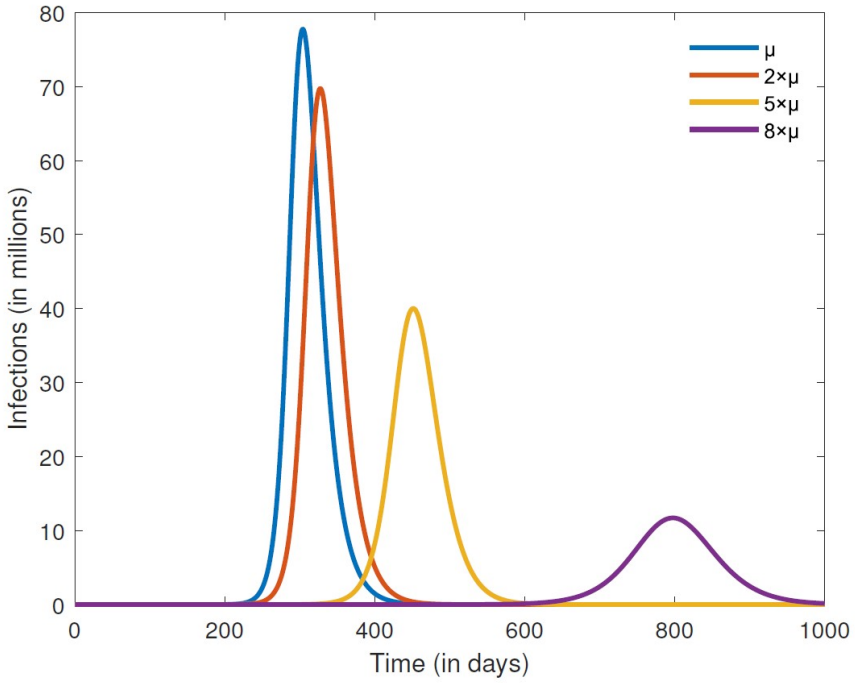


Figure 9. Numerical simulation of the number of infections for different misinformation decay rates $\mu_1 = \mu$. The other parameter values are fixed at $\Lambda = 0$, $\alpha = 0.25 \times 10^{-4}$, $\beta = 1/14$, $\rho = 0.1$, $\kappa = 5\rho$, $\sigma = 0.5 \times 10^{-3}$, $\gamma = 10^{-4}$, $\theta = 4 \times 10^{-5}$, and $\mu_0 = 0.0025$.

5. Conclusions

The COVID-19 pandemic has brought the issue of misinformation to the forefront, highlighting the significant impact on public health. Misinformation can undermine efforts to control the spread of the virus and thus lead to a range of negative consequences, including increased vaccine hesitancy, non-compliance with public health measures, and mistrust of public health officials. In this paper, we formulated a model that captures the transmission dynamics of COVID-19 simultaneously with the spread of misinformation in a community. Analysis of the model shows that it has two disease-free equilibria, one containing misinformation and the other does not. The DFE with no misinformation yields a reproduction number R_0 and is both locally and globally stable if $R_0 < 1$. The DFE containing misinformation yields a threshold parameter that renders the equilibrium stable if the value of the parameter is less than one.

Factors contributing to the spread of misinformation related to COVID-19 include the rapid spread of information through social media, the echo chamber effect, and the politicization of the pandemic. When the model is applied to the population of the United States, simulations of cumulative deaths suggest that the United States did a poor job at combating misinformation related to COVID-19. Numerical simulations of the model suggest that a multi-faceted approach is needed to combat the spread of misinformation and promote accurate information about the pandemic. Strategies to combat misinformation related to COVID-19 include promoting accurate information, increasing media literacy, and implementing fact-checking mechanisms. Further, simulations of the model predict that as long as people are able to enter and exit a community, the COVID-19 virus will remain endemic in such a community.

In conclusion, we have shown that mathematical modeling can play an important role in understanding the spread of misinformation related to COVID-19 and developing effective strategies for combating misinformation. Combating COVID-19 misinformation is essential for promoting public health and controlling the spread of the virus. Further, addressing the underlying causes of the spread of misinformation is essential, as is promoting accurate information, increasing media literacy, and implementing fact-checking mechanisms.

Author Contributions: Conceptualization, Z.S. and E.A.; methodology, Z.S. and E.A.; software, Z.S. and E.A.; validation, Z.S. and E.A.; formal analysis, Z.S. and E.A.; investigation, Z.S. and E.A.; resources, Z.S. and E.A.; data curation, Z.S. and E.A.; writing—original draft preparation, Z.S. and E.A.; writing—review and editing, Z.S. and E.A.; visualization, Z.S. and E.A.; supervision, E.A.; project administration, E.A.; funding acquisition, E.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. WHO Director-General's Opening Remarks at the Media Briefing on COVID-19—11 March 2020. Available online: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-COVID-19---11-march-2020 (accessed on 1 January 2023).
2. Cucinotta, D.; Vanelli, M. WHO Declares COVID-19 a Pandemic. *Acta Biomed.* **2020**, *91*, 157–160. [CrossRef] [PubMed]
3. Centers for Disease Control and Prevention. Available online: https://www.cdc.gov/coronavirus/2019-ncov/science/science-briefs/sars-cov-2-transmission.html (accessed on 1 January 2023).
4. Centers for Disease Control and Prevention. Available online: https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html (accessed on 1 January 2023).
5. da Rosa Mesquita, R.; Francelino Silva Junior, L.C.; Santos Santana, F.M.; Farias de Oliveira, T.; Campos Alcântara, R.; Monteiro Arnozo, G.; Rodrigues da Silva Filho, E.; Galdino dos Santos, A.G.; Oliveira da Cunha, E.J.; Salgueiro de Aquino, S.H.; et al. Clinical manifestations of COVID-19 in the general population: Systematic review. *Wien Klin Wochenschr.* **2021**, *133*, 377–382. [CrossRef] [PubMed]
6. World Health Organization. Who Coronavirus (COVID-19) Dashboard. Available online: https://covid19.who.int/ (accessed on 1 October 2023).
7. Suarez-Lledo, V.; Alvarez-Galvez, J. Prevalence of Health Misinformation on Social Media: Systematic Review. *J. Med. Internet Res.* **2021**, *23*, e17187. [CrossRef] [PubMed]
8. Greene, C.M.; Murphy, G. Quantifying the Effects of Fake News on Behavior: Evidence From a Study of COVID-19 Misinformation. *J. Exp. Psychol. Appl. Adv. Online Publ.* **2021**, *27*, 773. [CrossRef]
9. Agley, J.; Xiao, Y. Misinformation about COVID-19: Evidence for differential latent profiles and a strong association with trust in science. *BMC Public Health* **2021**, *21*, 89. [CrossRef]
10. Lew owsky, S.; Jacobs, P.; Neil, S. The Lab-Leak Hypothesis Made It Harder for Scientists to Seek the Truth. *Sci. Am.* **2022**, *326*, 72–77.
11. Flaherty, E.; Sturm, T.; Farries, E. The conspiracy of COVID-19 and 5G: Spatial analysis fallacies in the age of data democratization. *Soc. Sci. Med.* **2022**, *293*, 114546. [CrossRef] [PubMed]
12. Litman, L.; Rosen, Z.; Hartman, R.; Rosenzweig, C.; Weinberger-Litman, S.L.; Moss, A.J.; Robinson, J. Did people really drink bleach to prevent COVID-19? A guide for protecting survey data against problematic respondents. *PLoS ONE* **2023**, *18*, e0287837 [CrossRef] [PubMed]
13. Coman, C.; Bularca, M.C.; Repanovici, A.; Rogozea, L. Misinformation about medication during the COVID- 19 pandemic: A perspective of medical staff. *PLoS ONE* **2022**, *17*, e0276693. [CrossRef] [PubMed]
14. Hart, P.S.; Chinn, S.; Soroka, S. Politicization and Polarization in COVID-19 News Coverage. *Sci. Commun.* **2020**, *42*, 679–697. [CrossRef] [PubMed]
15. Enserink, M.; Kupferschmidt, K. With COVID-19, modeling takes on life and death importance. *Science* **2020**, *367*, 1414–1415. [CrossRef] [PubMed]
16. Musa, S.; Qureshi, S.; Zhao, S.; Yusuf, A.; Mustapha, U.T.; He, D. Mathematical modeling of COVID-19 epidemic with effect of awareness programs. *Infect. Dis. Model.* **2021**, *6*, 448–460. [CrossRef] [PubMed]
17. Moore, S.; Hill, E.M.; Tildesley, M.J.; Dyson, L.; Keeling, M.J. Vaccination and non-pharmaceutical interventions for COVID-19: A mathematical modelling study. *Lancet Infect. Dis.* **2021**, *21*, 793–802. [CrossRef] [PubMed]
18. Achaiah, N.C.; Subbarajasetty, S.B.; Shetty, R.M. R_0 and R_e of COVID-19: Can We Predict When the Pandemic Outbreak will be Contained? *Indian J. Crit. Care Med.* **2020**, *24*, 1125–1127. [CrossRef] [PubMed]
19. Olivares, A.; Staffetti, E. Uncertainty quantification of a mathematical model of COVID-19 transmission dynamics with mass vaccination strategy. *Chaos Solitons Fractals* **2021**, *146*, 110895. [CrossRef] [PubMed]
20. Zhu, M.; Agyingi, E. Modeling COVID-19 Breakthrough Infections in a Vaccinated Population. *Wseas Trans. Syst.* **2023**, *22*, 584–592. [CrossRef]

21. Nguyen, V.K.; Hernandez-Vargas, E.A. Parameter Estimation in Mathematical Models of Viral Infections Using R. In *Influenza Virus. Methods in Molecular Biology, Volume 1836*; Yamauchi, Y., Ed.; Humana Press: New York, NY, USA, 2018.
22. Dautel, K.; Agyingi, E. Modeling the impact of educational campaign on the transmission dynamics of Ebola. *J. Biol. Syst.* **2020**, *29*, 1–23. [CrossRef]
23. LaSalle, J.P. *The Stability of Dynamical Systems*; SIAM: Philadelphia, PA, USA, 1976.
24. Eikenberry, S.E.; Mancuso, M.; Iboi, E.; Phan, T.; Eikenberry, K.; Kuang, Y.; Kostelich, E.; Gumel, A.B. To mask or not to mask: Modeling the potential for face mask use by the general public to curtail the COVID-19 pandemic. *Infect. Dis. Model.* **2020**, *5*, 293–308. [CrossRef] [PubMed]
25. Tang, B.; Bragazzi, N.L.; Li, Q.; Tang, S.; Xiao, Y.; Wu, J. An updated estimation of the risk of transmission of the novel coronavirus (2019nCov). *Infect. Dis. Model.* **2020**, *5*, 248255.
26. Xu, C.; Pang, Y.; Liu, Z.; Shen, J.; Liao, M.; Li, P. Insights into COVID-19 stochastic modelling with effects of various transmission rates: Simulations with real statistical data from UK, Australia, Spain, and India. *Phys. Scr.* **2024**, *99*, 025218 [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

AppliedMath

MDPI

Article

Approximating a Function with a Jump Discontinuity—The High-Noise Case

Qusay Muzaffar [1,*], David Levin [2] and Michael Werman [1]

[1] Department of Computer Science, The Hebrew University of Jerusalem, Jerusalem 9190401, Israel; michael.werman@mail.huji.ac.il
[2] Department of Applied Mathematics, Tel Aviv University, Tel Aviv 6997801, Israel; levin@tauex.tau.ac.il
* Correspondence: qusay.muzaffar@mail.huji.ac.il

Abstract: This paper presents a novel deep-learning network designed to detect intervals of jump discontinuities in single-variable piecewise smooth functions from their noisy samples. Enhancing the accuracy of jump discontinuity estimations can be used to find a more precise overall approximation of the function, as traditional approximation methods often produce significant errors near discontinuities. Detecting intervals of discontinuities is relatively straightforward when working with exact function data, as finite differences in the data can serve as indicators of smoothness. However, these smoothness indicators become unreliable when dealing with highly noisy data. In this paper, we propose a deep-learning network to pinpoint the location of a jump discontinuity even in the presence of substantial noise.

Keywords: approximation theory; deep learning; piecewise smooth function approximation; highly noisy data

MSC: 65D15

check for updates

Citation: Muzaffar, Q.; Levin, D.; Werman, M. Approximating a Function with a Jump Discontinuity—The High-Noise Case. *AppliedMath* **2024**, *4*, 561–569. https://doi.org/10.3390/appliedmath4020030

Academic Editor: Claude Chaudet

Received: 29 February 2024
Revised: 2 April 2024
Accepted: 15 April 2024
Published: 2 May 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Jump discontinuities occur when a function experiences a sudden, finite change in its value at a specific point. Functions containing jump discontinuities are prevalent in various fields, including physics, electrical engineering, computer vision, geology, biology, and other applied domains [1–3]. Consequently, the task of approximating these functions from sampled data frequently arises. Estimating the location of the discontinuity is crucial as standard approximation methods often introduce significant errors due to the loss of regularity near jump discontinuities.

Several papers have investigated algorithms to approximate piecewise smooth univariate functions with a particular focus on detecting the interval of discontinuities, including [4,5]. In a more recent paper [6], a function of the differences in the data, referred to as the signature, was used to approximate the function. However, the aforementioned methods fail when dealing with noisy data.

Alternative algorithms designed for noisy data have been proposed and discussed in other papers, such as [7]. However, these methods rely on prior knowledge of the function, on finding a type of function, called an averaging function, and on regularization parameters. In addition, no information about the amount or the type of the additive noise is given. In [8], another method is presented based on the numerical second derivative, albeit limited to samples on a uniform grid. It requires knowing the supremum of the second derivative of the function and is limited to noise that is less than a sixth of the length of the jump. In [9], a third method is described using convolution with a filter function. However, a comparative analysis demonstrates a lower success rate when contrasted with the proposed method.

Deep learning is a methodology in machine learning where artificial neural networks with multiple layers of processing are used to extract progressively higher level features from data between the input and output. Deep learning models are composed of multiple layers of neurons that hierarchically process the input data. Each layer learns to transform the data into a more abstract representation, with the final layer providing the model's output (e.g., classification, regression, generation). The learning process involves iteratively adjusting the parameters of the neural network (e.g., weights and biases) using optimization algorithms such as gradient descent [10].

Deep learning models can be trained on large datasets to automatically learn patterns and features directly from the data, eliminating the need for manual feature engineering. This ability to learn complex representations from raw data has gained widespread acceptance as the preferred approach in various applications, including speech recognition, visual object recognition, object detection, natural language processing, drug discovery, and genomics [11].

In this paper, we propose a novel fully convolutional neural network [10] to estimate the position of the jump discontinuity of a univariate function from a set of noisy samples. Subsequently, we employ two cubic splines to approximate the function, one for the data on the left side of the jump discontinuity and the other for the right side. Our method is shown to be superior to previous methods demonstrated through a series of experiments.

2. Our Approach

In this section, we present our approach for approximating the location of the discontinuity of a univariate piecewise smooth function f from a sample with *iid* additive noise $\{(x_i, f(x_i) + \omega_i)\}_{i=0}^{N}$, where

$$\{\omega_j\} \sim \mathcal{N}(0, \sigma^2)\,.$$

2.1. The Model

Our deep learning model consists of a total of ten convolutional layers, as illustrated in Figure 1. Through experimentation, we determined that a ten-layer model strikes the best balance between simplicity in terms of learnable features and output accuracy. These ten layers consist of one input layer, one output layer, and eight intermediate layers. Each layer is equipped with two convolution operators that utilize filter sizes of 3 and 5. We choose two different filter sizes so that the model can capture the function's behavior through its samples in two different ways. It is worth mentioning that we choose filters with odd numbers so that the size of the output after the convolution operation is the same as the input after padding. For example, if we have an input of size N, we add zeros at the beginning and at the end of the input, and then, we apply a convolution with a filter size of 3, and we retain the original size of the input, which is N. If we were to choose a filter of size of 2, then the output size would be $N + 1$ in this case. Retaining the size of the input helps with capturing the details of the model. The output channels of these two operators are concatenated as shown in Equation (1). The initial layer accepts 8 input channels, Equation (3), and outputs 64 channels. For the i-th intermediate layer, the input is $64 \times i$ channels, and it generates 64 channels. The output of the i-th intermediate layer is concatenated with its input channels and is passed to the $(i+1)$-th layer.

$$\left(\boxed{v * \mathcal{C}_{f_3}} \quad \boxed{v * \mathcal{C}_{f_5}} \right). \tag{1}$$

The last layer has an input of 64 channels and outputs a probability vector.

$$v(x) = \begin{cases} x, & 0 \le x < 1 \\ 2 - x, & 1 \le x \le 2 \end{cases} \tag{2}$$

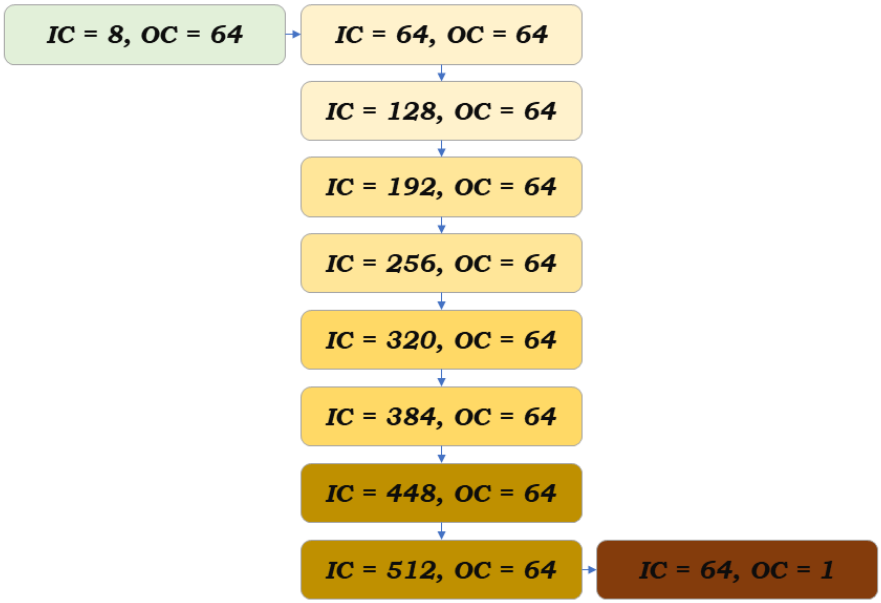
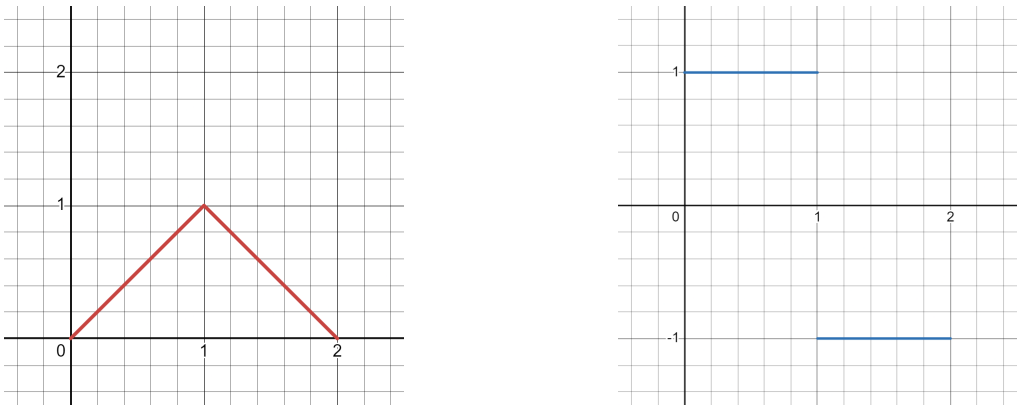


Figure 1. The model architecture. Each rectangle represents a layer, and the two numbers indicate the number of input and the number of output channels. In each layer, there are two convolutional operations as in Equation (1). In the intermediate layers, each output is concatenated to the input of the same layer and is passed to the next layer.

The activation function, Equation (2) (see Figure 2), is applied to the outputs of the two convolution operators after they are concatenated before being passed to the next layer.



(**a**) (**b**)

Figure 2. The graph of the activation function (**a**) and the graph of its derivative (**b**).

The samples x and y of size $N + 1$, where N is arbitrary, are part of the input data for the model. We augment the input data with first-order, second-order, and third-order forward differences as presented in Equation (3).

$$(x, y) \to \left(x, y, \Delta_x, \Delta_y, \Delta_x^2, \Delta_y^2, \Delta_x^3, \Delta_y^3\right) \tag{3}$$

The output of the model is a vector of probabilities, of length N, where the kth entry is the probability that there is a discontinuity between the kth and $k + 1$th sample. The loss function is cross-entropy [10] as the problem is classification.

$$\mathcal{L}(x) = -\sum_{n=1}^{N}\sum_{c=1}^{C} \log \frac{e^{x_{n,c}}}{\sum_{i=1}^{C} e^{x_{n,i}}} \tag{4}$$

Let $\mathcal{M}$ be the model, and let $p = \mathcal{M}\left(\left(x, y, \Delta_x, \Delta_y, \Delta_x^2, \Delta_y^2, \Delta_x^3, \Delta_y^3\right)\right)$. Then, $\tilde{i} = \arg\max\{p\}$ is the index of the closest sample to the discontinuity point from the left side. We define

$$\tilde{\xi} = \frac{x_{\tilde{i}} + x_{\tilde{i}+1}}{2} \tag{5}$$

as the approximated discontinuity point.

2.2. Training Data

In this section, we outline the process of generating synthetic data for training the model, both as input and output. We recall that a random number between a and b from a uniform distribution is denoted as

$$c \sim \mathcal{U}([a, b])$$

We start by generating univariate piecewise functions, each containing a jump discontinuity, from which we randomly sample points. To create a realistic scenario, we introduce white noise to these samples and use them as input data. The output data consist of the index of the closest sample to the left side of the jump discontinuity. Let

$$t = \left\{ -3\Delta_t + i\Delta_t \right\}_{i=0}^{T} \tag{6}$$

be the set of knots where $\Delta_t \sim \mathcal{U}([\frac{3}{10}, \frac{5}{10}])$ and $T = \left\lfloor \Delta_t^{-1} \right\rfloor + 7$, so there are enough knots inside and outside the interval, and let

$$g(x) = \begin{cases} 1 & , x < 0 \\ 0 & , 0 \leq x \end{cases}.$$

Now, we let

$$f(x) = ag(x - \xi) + bxg(x - \xi) + \sum_{i=0}^{T-k-1} c_i B_{i,k,t}(x). \tag{7}$$

Here, $B_{i,k,t}$ are cubic B-splines basis functions ($k = 3$). The coefficients of the cubic B-splines $\{c_i\}_{i=0}^{T-k-3}$ are selected randomly from a uniform distribution:

$$\{c_i\}_{i=0}^{T-k-3} \sim \mathcal{U}([-2, 2]).$$

Similarly, the coefficients a and b are selected.

$$a, b \sim \mathcal{U}\left(\left[-2, -\frac{4}{10} \right] \cup \left[\frac{4}{10}, 2 \right] \right)$$

Finally, the location of the discontinuity point is as follows:

$$\xi \in \left[\frac{6}{N}, 1 - \frac{5}{N} \right]$$

The input data consist of points sampled from a function $f(x)$ of the form as in Equation (7):

$$\mathcal{S} = \left\{ (x_i, f(x_i) + \omega_i) \right\}_{i=0}^{N}, \tag{8}$$

where

$$\omega_i \sim \mathcal{U}\left(\left[0, \frac{a}{3} \right] \right).$$

Here, $\{\omega_i\}_{i=0}^{N}$ is the additive noise. The output is a unit vector of length $N-1$ where the 1 is located at the closet sample from the left of the jump discontinuity. We train the model with the ADAM [12] stochastic gradient descent algorithm.

2.3. Approximating the Function from Its Samples

After locating the discontinuity interval and approximating the discontinuity point's location, we utilize two cubic spline approximations to approximate the function with its jump discontinuity. We use the cubic spline approximation because approximating with cubic splines is advantageous due to their smoothness, flexibility, accuracy, ease of implementation, and numerical stability. These properties make cubic splines a versatile and widely used tool for data interpolation, curve fitting, and modeling in various fields, including mathematics, engineering, computer graphics, and scientific computing [13]. First, we split the data into the left groups and the right groups as follows:

$$\begin{aligned} x_\ell &= \left\{ x \text{ is a component of, } (x,y) \in \mathcal{S} : x < \tilde{\xi} \right\} \\ x_r &= \left\{ x \text{ is a component of, } (x,y) \in \mathcal{S} : x > \tilde{\xi} \right\} \\ y_\ell &= \left\{ y \text{ is a component of, } (x,y) \in \mathcal{S} : x < \tilde{\xi} \right\} \\ y_r &= \left\{ y \text{ is a component of, } (x,y) \in \mathcal{S} : x > \tilde{\xi} \right\} \end{aligned} \tag{9}$$

where $\mathcal{S}$ is defined as in Equation (8) and $\tilde{\xi}$ is from Equation (5).

Then, by using a least squares approximation, we find two cubic splines, $\Gamma_\ell(x)$, which approximates the function from the left side of the jump discontinuity by using the data of x_ℓ and y_ℓ and second one, $\Gamma_r(x)$, which approximates the function from the right side of the jump discontinuity by using the data of x_r and y_r. Before proceeding, we begin by constructing the knots for both splines. Let

$$\Delta_\ell = \left\lfloor \frac{\tilde{\xi} * N}{5} \right\rfloor^{-1} \text{ and } \quad \Delta_r = \left\lfloor \frac{(1-\tilde{\xi}) * N}{5} \right\rfloor^{-1},$$

and let $T_\ell = \Delta_\ell^{-1} + 7$ and $T_r = \Delta_r^{-1} + 7$. Then, the two sets of knots for the left and the right approximations are

$$t_\ell = \left\{ -3\Delta_\ell + i\Delta_\ell \right\}_{i=0}^{T_\ell} \quad \text{and} \quad t_r = \left\{ -3\Delta_r + \xi + j\Delta_r \right\}_{j=0}^{T_r}$$

We denote the set of the coefficients of $\Gamma_\ell(x)$ and $\Gamma_r(x)$ by c_ℓ and c_r, respectively. To find the optimal sets of coefficients c_ℓ and c_r, we find a closed-form solution for the following equation:

$$\underset{\substack{c_\ell \in \mathbb{R}^{T_\ell - 4} \\ c_r \in \mathbb{R}^{T_r - 4}}}{\arg\min} \left\| \left(\begin{array}{c|c} M_\ell & \mathbf{0} \\ \hline \mathbf{0} & M_r \end{array} \right) \begin{pmatrix} c_\ell \\ c_r \end{pmatrix} - \begin{pmatrix} y_\ell \\ y_r \end{pmatrix} \right\|_{L_2}^2 + \left\| \left(\begin{array}{c|c} \lambda_\ell D_\ell & \mathbf{0} \\ \hline \mathbf{0} & \lambda_r D_r \end{array} \right) \begin{pmatrix} c_\ell \\ c_r \end{pmatrix} \right\|_{L_2}^2.$$

Here, the matrices of $M_\ell \in \mathbb{R}^{\|x_\ell\| \times \|c_\ell\|}$ and $M_r \in \mathbb{R}^{\|x_r\| \times \|c_r\|}$ are the data matrices and are defined as follows:

$$M_\ell = \begin{bmatrix} \ddots & \vdots & \iddots \\ \dots & B_{j,k=3,t_\ell}(x_{\ell,i}) & \dots \\ \iddots & \vdots & \ddots \end{bmatrix}, \quad M_r = \begin{bmatrix} \ddots & \vdots & \iddots \\ \dots & B_{j,k=3,t_r}(x_{r,i}) & \dots \\ \iddots & \vdots & \ddots \end{bmatrix}$$

The matrices of $D_\ell \in \mathbb{R}^{\left(\|c_\ell\|-2\right) \times \|c_\ell\|}$ and $D_r \in \mathbb{R}^{\left(\|c_r\|-2\right) \times \|c_r\|}$ are the second difference matrices. Those matrices are used in the above equation so that the optimal solution is the one with a minimal second-derivative.

3. Numerical Results

In this section, we illustrate the two aspects of our approach: approximating the location of a jump discontinuity in a given function and approximating the function itself. Furthermore, we conducted a statistical test to show the effectiveness of our method compared to two alternative methods. Lastly, we performed another statistical test to present the average error of our final approximation.

3.1. Detecting the Interval of the Discontinuity and Approximating the Function

In this section, we present our approach to approximating the discontinuity point of a randomly generated function and subsequently approximating the function itself. To accomplish this, we generated two functions following the format described in Equation (7). We took 40 random samples from $[0, 1]$. These samples were then subjected to the addition of white noise with a variance equal to one-third of the jump's height.

Next, we employed our model to determine the index of the closest sample from the left of the discontinuity. Subsequently, we utilized the procedure outlined in Section 2.3 to derive an approximation of the function. The results of these experiments are presented in Figure 3.

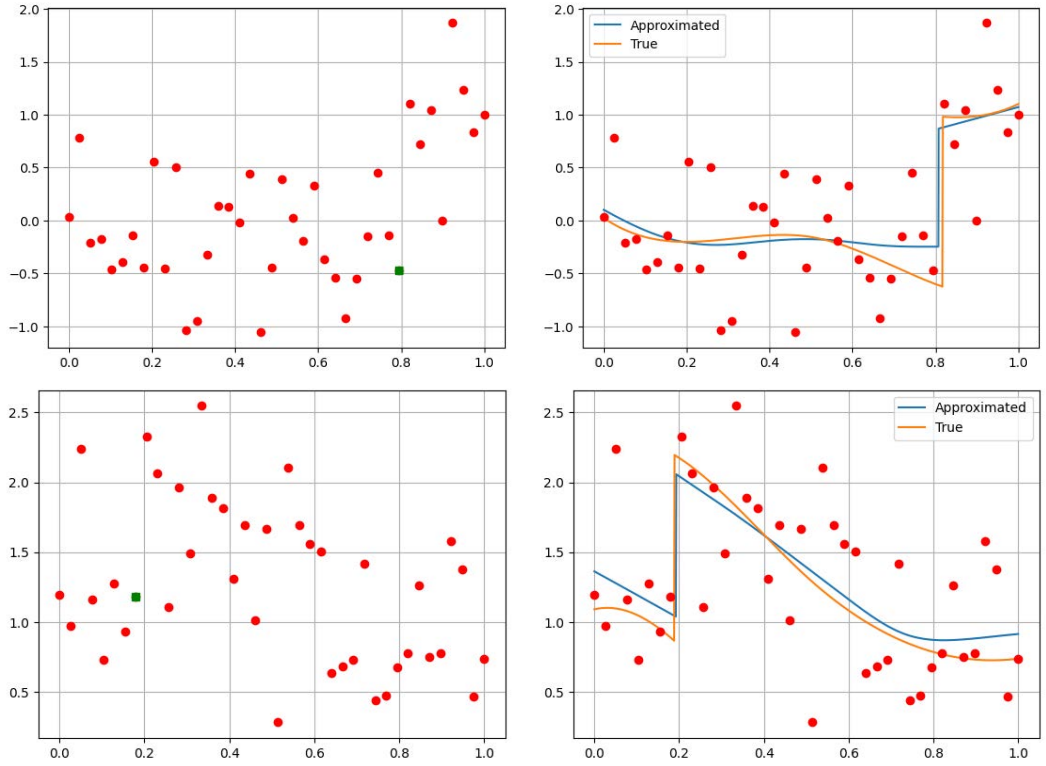


Figure 3. We demonstrate our method visually by approximating the location of the discontinuity point of two univariate piecewise smooth functions from their samples, to which white noise was added. The upper-left and lower-left figures show the samples with the added noise. All samples are red-colored except for the point that indicates the closest sample from the right to the jump discontinuity, which is green-colored. Our approach found the green point. The upper-right and the lower-right figures show the approximations for the two functions, which were achieved using two cubic splines for each one of the functions.

3.2. Comparing Our Approach with Two Other Approaches

In [6], the signature, of degree k, of g, is $\sigma_k(g) \in \mathbb{R}^{N+k}$, which is the forward kth-order difference of the samples of g padded by k zeros on each side. In [9], a method was developed based on the use of a filter function with which a convolution is computed for detecting the discontinuity point. In this experiment, we compare our method with

the other two aforementioned methods on a large number of samples with various noise levels. The samples are taken from univariate piecewise smooth functions with several discontinuity jump heights. Tables 1–3 each represent the results of one of the approaches. The columns represent the height of the jump, and the rows represent the variance in the additive white noise. As shown in the tables, our approach has an increase of between twofold and tenfold compared to the other methods.

Table 1. The probability of detecting the interval of the discontinuity location using the method in [6]. Columns and rows represent jump height and noise level, respectively. Each cell contains the average success of 1000 trials.

	0.30	0.60	0.90	1.20	1.50	1.80	2.10	2.40
0.00	0.38	0.47	0.50	0.49	0.48	0.49	0.48	0.49
0.10	0.10	0.28	0.45	0.50	0.50	0.53	0.48	0.51
0.20		0.10	0.21	0.29	0.42	0.46	0.46	0.50
0.30			0.12	0.14	0.24	0.29	0.37	0.42
0.40				0.11	0.14	0.19	0.25	0.31
0.50					0.11	0.13	0.17	0.22
0.60						0.11	0.13	0.17
0.70							0.10	0.12
0.80								0.12

Table 2. The probability of detecting the interval of the discontinuity location using the method in [9]. Columns and rows represent jump height and noise level, respectively. Each cell contains the average success of 1000 trials.

	0.30	0.60	0.90	1.20	1.50	1.80	2.10	2.40
0.00	0.24	0.24	0.28	0.24	0.26	0.25	0.24	0.23
0.10	0.30	0.27	0.25	0.25	0.27	0.24	0.25	0.27
0.20		0.30	0.26	0.28	0.24	0.25	0.28	0.25
0.30			0.33	0.28	0.29	0.26	0.28	0.28
0.40				0.34	0.30	0.30	0.26	0.30
0.50					0.34	0.32	0.30	0.28
0.60						0.32	0.29	0.33
0.70							0.34	0.31
0.80								0.33

Table 3. The probability of detecting the interval of the discontinuity location using our method. Columns and rows represent jump height and noise level, respectively. Each cell is the average success of 1000 trials.

	0.30	0.60	0.90	1.20	1.50	1.80	2.10	2.40
0.00	0.89	0.99	1.00	1.00	1.00	1.00	1.00	1.00
0.10	0.53	0.93	1.00	1.00	1.00	1.00	1.00	1.00
0.20		0.67	0.92	0.98	1.00	1.00	1.00	1.00
0.30			0.73	0.92	0.97	0.99	1.00	1.00
0.40				0.76	0.90	0.95	0.98	0.99
0.50					0.79	0.89	0.94	0.97
0.60						0.79	0.88	0.95
0.70							0.79	0.88
0.80								0.79

3.3. Error Measurement of Approximations of Functions

In this experiment, we compare our approach with the approach of using the signature [6] in approximating a univariate piecewise smooth function from its noisy samples. As before, we ran a test on many samples with various noise levels. We measure the error of the approximation by using the L_∞ metric on the set $\left[0, \min\left(\xi, \tilde{\xi}\right)\right] \cup \left[\max\left(\xi, \tilde{\xi}\right), 1\right]$, where ξ and $\tilde{\xi}$ are the real and the approximated point of discontinuity. Additionally, we measure

$\left|\tilde{\xi} - \xi\right|$. We created two graphs, in Figures 4 and 5, for the two approaches and the two measurements. The x-axis represents the variance of the additive white noise, whereas the y axis represents the height of the jump. Each point in the graph represents the average error of 100 trials. In the first test, the error of our approach was less than half of the error of the other approach, and in the second test, the error of our approach was fifth.

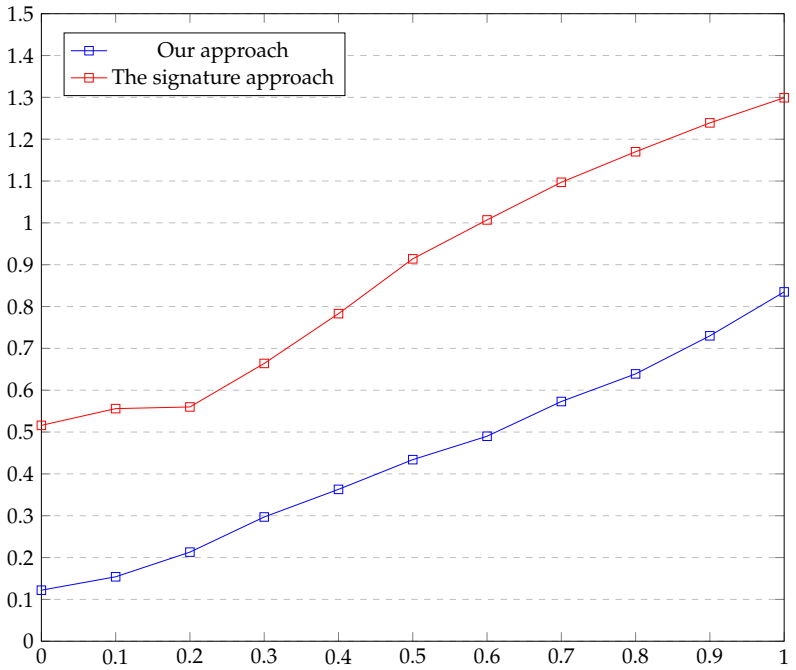


Figure 4. This graph shows the L_∞-error of the approximation on the set $\left[0, \min\left(\xi, \tilde{\xi}\right)\right] \cup \left[\max\left(\xi, \tilde{\xi}\right), 1\right]$, where ξ and $\tilde{\xi}$ are the real and the approximated point of discontinuity. Here, ξ and $\tilde{\xi}$ are the real and the approximated point of discontinuity.

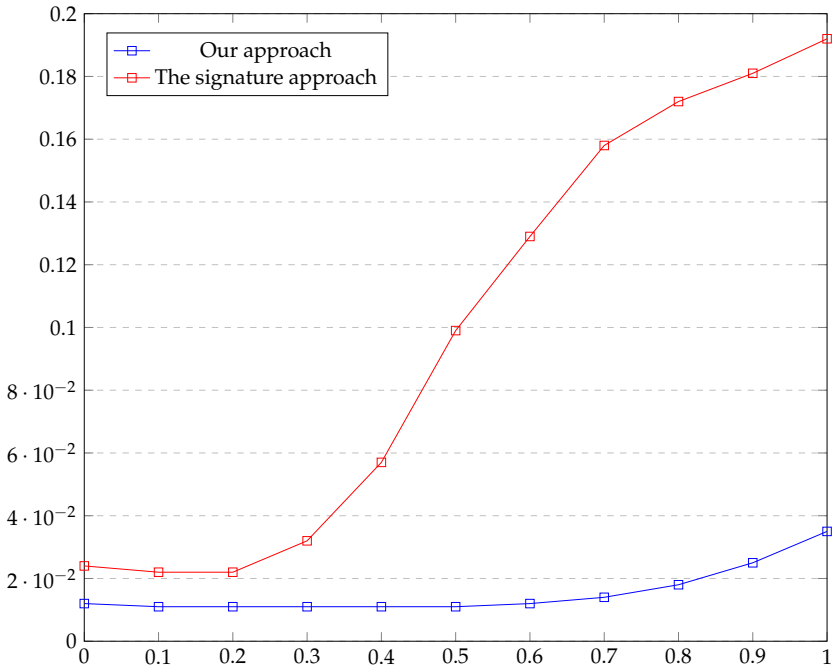


Figure 5. This graph shows the error of $\left|\tilde{\xi} - \xi\right|$. Here, ξ and $\tilde{\xi}$ are the real and the approximated point of discontinuity.

4. Conclusions

We presented a new approach, which is based on deep learning, to approximate the location of the point of jump discontinuity from given samples, with highly additive noise, of a univariate piecewise smooth function with a jump discontinuity. We compared our approach against two other methods and illustrated the significantly higher success rate of our method against them. In addition, we demonstrated our approach to approximating the location of the discontinuity in synthetic data of generated piecewise smooth univariate functions. Then, we approximated these functions and showed the error measurements. Our future work will deal with extending this work to bivariate piecewise smooth functions, in which the jump does not occur at one point but rather on a one-dimensional curve. In case that the two-dimensional data are equally distanced along the y-axis, we can utilize this approach to locate the nearest points to the one-dimensional curve. However, approximating the one-dimensional curve is not intuitive and will be a subject for future work.

Author Contributions: Methodology, Q.M.; investigation, Q.M.; writing—original draft preparation, Q.M.; writing—review and editing, M.W. and Q.M.; supervision, M.W. and D.L. All authors have read and agreed to the published version of the manuscript.

Funding: ISF Israeli Science Foundation.

Data Availability Statement: There are no specific data used in this paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ageev, A.L.; Antonova, T.V. On a new class of ill-posed problems. *Izv. Ural. Gos. Univ. Mat. Mekh. Inform.* **2008**, *58*, 24–42.
2. Cherepashchuk, A.M.; Goncharskii, A.V.; Yagola, A.G. Inverse problems of astrophysics. In *Ill-Posed Problems in Natural Sciences, Proceedings of the International Conference Held in Moscow, August 19–25. 1991*; De Gruyter: Berlin, Germany, 1992; pp. 472–481.
3. Terebizh, V.Y. *Introduction to the Statistical Theory of Inverse Problems*; Fimzmatlit: Moscow, Russia, 2005.
4. Aràndiga, F.; Donat, R. Nonlinear multiscale decompositions: The approach of A. Harten. *Numer. Algorithm* **2000**, *23*, 175–216. [CrossRef]
5. Shu, C.-W. Essentially non-oscillatory and weighted essentially non-oscillatory schemes for hyperbolic conservation laws. In *Advanced Numerical Approximation of Nonlinear Hyperbolic Equations: Lectures Given at the 2nd Session of the Centro Internazionale Matematico Estivo (C.I.M.E.) Held in Cetraro, Italy, June 23–28. 1997*; Springer: Berlin/Heidelberg, Germany, 1998; pp. 325–432.
6. Amat, S.; Levin, D.; Ruiz-Álvarez, J. A two-stage approximation strategy for piecewise smooth functions in two and three dimensions. *IMA J. Numer. Anal.* **2021**, *42*, 3330–3359. [CrossRef]
7. Antonova, T.V. New methods for localizing discontinuities of a noisy function, *Numer. Anal. Appl.* **2010**, *3*, 306–316.
8. Mbakop, E. *Reconstruction of Discontinuities in Noisy Data*; Citeseer: Pennsylvania, USA. Available online: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=0c1b69c38e1df16c7c939ad07705ee6bb53a6bd2 (accessed on 31 March 2024).
9. Lee, D. Detection, Classification, and Measurement of Discontinuities. *SIAM J. Sci. Stat. Comput.* **1991**, *12*, 311–341. [CrossRef]
10. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
11. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
12. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
13. Chapra, S.C.; Canale, R.P. *Numerical Methods for Engineers*, 8th ed.; McGraw-Hill Education: New York, NY, USA, 2021.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

AppliedMath

MDPI

Article

Quantum Mixtures and Information Loss in Many-Body Systems

Diana Monteoliva [1], Angelo Plastino [2,*] and Angel Ricardo Plastino [3]

[1] UNLP-Comisión de Investigaciones Científicas, Provincia de Buenos Aires, La Plata 1900, Argentina; monteoli@fisica.unlp.edu.ar
[2] Instituto de Física La Plata–CCT-CONICET, Universidad Nacional de La Plata, C.C. 727, La Plata 1900, Argentina
[3] CeBio-Departamento de Ciencias Básicas, Universidad Nacional del Noroeste, Prov. de Buenos Aires (UNNOBA), CONICET, Junin 1988, Argentina; arplastino@unnoba.edu.ar
* Correspondence: plastino@fisica.unlp.edu.ar

Abstract: In our study, we investigate the phenomenon of information loss, *as measured by the Kullback–Leibler divergence* , in a many-fermion system, such as the Lipkin model. Information loss is introduced as the number N of particles increases, particularly when the system is in a mixed state. We find that there is a significant loss of information under these conditions. However, we observe that this loss nearly disappears when the system is in a pure state. Our analysis employs tools from information theory to quantify and understand these effects.

Keywords: many-fermion systems; mixture degree; finite temperature; SU2 symmetry;

MSC: 81-82

1. Introduction

In this paper we study, using information theory tools, peculiarities of quantum mixtures [1–5] in the context of an exactly solvable model of N interacting fermions of mass m called the Lipkin model. Our focus is on information losses, as measured by the Kullback–Leibler divergence (KL), that take place when N augments from a previous, lower, reference fermion number N_o.

We present below the main contents of our study.

The Lipkin model [6–16] is a well-known model of N interacting fermions of mass m. It is well known in nuclear physics and for quantum many-body systems, particularly in the study of the interplay between pairing correlations and quantum phase transitions [6]. Using information theory tools to explore correlations between the degree of quantum mixture (QMx) and some model traits is an interesting and valid research direction. Here, we pay special attention to the QMx–particle number N relationship. Information theory provides a formalism for quantifying and understanding correlations, entropy, and information content in physical systems. Key concepts include entropy, mutual information, and conditional entropy. These tools can be applied to analyze the relationships between different parameters in a quantum system and provide insights into the complexity, correlations, and entanglement present in the system. Such an interdisciplinary approach, combining the Lipkin model, information theory, and quantum statistical mechanics, can provide a rich framework for exploring the intricate connections between model traits and special statistical quantifiers.

1.1. Quantum Mixtures

In this contribution, the concept of a quantum mixture (QMx) [1–4] is studied from the viewpoint of interacting many-fermions systems. QMx is important in quantum mechanics because it reflects the inherent probabilistic nature of quantum systems and provides a way to describe their statistical behavior. Quantum mechanics is fundamentally different

check for updates

Citation: Monteoliva, D.; Plastino, A.; Plastino, A.R. Quantum Mixtures and Information Loss in Many-Body Systems. *AppliedMath* **2024**, *4*, 570–579. https://doi.org/10.3390/appliedmath4020031

Received: 4 March 2024
Revised: 13 April 2024
Accepted: 20 April 2024
Published: 2 May 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

from classical mechanics and one of its key features is the existence of superposition states and the associated concept of quantum entanglement. The concept of quantum mixtures is important because [1–5]:

- In quantum mechanics, systems can exist in multiple states simultaneously, a phenomenon known as superposition. A quantum mixture is a probabilistic combination of these states, and it allows us to describe the overall state of a system when it is not in a pure state;
- Probabilistic predictions: quantum mechanics predicts outcomes probabilistically. The state of a system is described by a wave function, and the probabilities of different outcomes are determined by the squared magnitudes of probability amplitudes. A quantum mixture encapsulates these probabilistic predictions for composite systems;
- Measurement and observables: when a measurement is made on a quantum system, the system typically collapses into one of its possible states, and the outcome of the measurement is probabilistic. Quantum mixtures provide a way to express the statistical distribution of outcomes for measurements on an ensemble of identically prepared systems;
- Quantum information theory: quantum mixture concepts are fundamental in quantum information theory, where the manipulation and transmission of quantum information are studied. Understanding how mixed states evolve and interact is essential for developing quantum algorithms and quantum communication protocols;
- Statistical mechanics: in the context of statistical mechanics, quantum mixtures are used to describe the statistical ensembles of quantum systems. The grand canonical ensemble, for example, involves a mixture of states with different particle numbers;
- In summary [1–5], the concept of a quantum mixture is a fundamental aspect of quantum mechanics that allows us to handle the probabilistic and statistical nature of quantum systems. It is a crucial tool for making predictions, understanding correlations, and developing quantum technologies.

The above considerations make it abundantly clear that mixed states play a crucial role in quantum mechanics, and that their importance lies in providing a more complete and realistic description of physical systems compared to pure states alone. As stated above, mixed state represents a statistical ensemble of pure states. This ensemble may include different pure states with certain probabilities. Mathematically, a mixed state is described by a density matrix, which is a positive semi-definite, Hermitian operator acting on Hilbert's space. Mixed states are essential when dealing with statistical ensembles, thermal equilibrium, or open quantum systems that are subject to interactions with their environment. In many practical situations, we may not have complete knowledge about the exact state of a quantum system. Instead, we might have statistical information about the probabilities of different pure states that the system could be in. Mixed states provide a natural way to describe such statistical ensembles. In quantum information theory, mixed states play a central role in characterizing the performance of diverse quantum algorithms and quantum communication protocols. Indeed, it must be emphasized that mixed states are fundamental to a comprehensive understanding of quantum systems, especially when considering the statistical nature of measurements, thermal effects, and the impact of interactions with the environment. They provide a bridge between the idealized concept of pure states and the practicalities of real-world quantum systems [1–5]. Finally, it is interesting for our present purposes to note that the loss of quantum characteristic behaviors as the number of particles becomes large is a well-known phenomenon [5]. Indeed, we know that when the number of fermions in a collective becomes very large, certain quantum systems can exhibit behavior that approximates classical behavior. This is often referred to as the correspondence principle. The statistical comportment of a large number of quantum particles can then resemble classical behavior on a macroscopic scale [5].

1.2. Nature of Our Results

Let us emphasize that in the present contribution, every result is exact. We use an easily solvable many-fermion model called the Lipkin model (LM) [6] that is of significant relevance in the field of quantum physics and condensed matter physics [6–16]. This model, while simplified, offers valuable insights into the behavior of complex quantum systems and provides simplified settings in which fundamental quantum principles can be explored analytically. Researchers often use the LM to develop and test new theoretical frameworks, such as many-body techniques and quantum field theory [6–16].

We will use here very low temperatures. This is a well-known statistical mechanics technique to approximate the ground-state properties of a system. It results in a common and powerful approach in condensed matter physics and quantum mechanics. This approach exploits the relationship between low-temperature properties and the ground state of a system [15,16]. We take as our low temperatures T the values corresponding to the inverse temperature $\beta = 20$ or 10, with $\beta = 1/kT$ and k being the Boltzmann constant.

1.3. Present Goal and Organization

It is our goal to investigate some clues regarding the precedent considerations in the context of a celebrated nuclear physics model called the Lipkin Model [7]. There are hundreds of pertinent references that one could cite. Of course, there is no space for such a task. We content ourselves with [6–16]. The model offers us a purely quantum environment. We wish to investigate those of the model's traits that depict quantum mixtures. The rest of the paper is organized as follows. We begin in Section 2 by describing the mathematics of the Lipkin model [6–16]. Section 3 recalls materials relevant to the Kullback–Leibler divergence, one of the two main information tools that we employ herein. The second tool is advanced in this effort in the context of the Likpin model (it is well known elsewhere), being addressed in Section 4, where we show that it constitutes a valid quantifier for assessing the mixture degree. Our central topic regarding information loss is discussed in Section 5 and conclusions are drawn in Section 6.

2. The Lipkin Formalism [15,16]

We consider N *interacting fermions of mass* m and define $\Omega = N/2$. The Lipkin model [6–16] consists of $N = 2\Omega$ fermions of mass m that accommodate themselves into two distinct single-particle (sp) energy levels, each of them N-fold degenerate. An energy gap separates the two levels. This gap is called ϵ. Thus we face a total of 4Ω s.p. microstates, labeled by two quantum numbers (denominated μ and p). The first quantum number μ attains the values $\mu = -1$ (lower level) and $\mu = +1$ (upper level). The quantum number p, called the quasi spin, pertains to a $2N$-fold degeneracy. This pair p, μ is customarily viewed as a "site". This site can be occupied (by a fermion) or be empty. Lipkin sets [6]

$$N = 2J, \tag{1}$$

where J acts as a sort of angular momentum. In addition, Lipkin [6] employs particular angular-momentum-like operators denominated by quasi-spin operators, which are

$$J_z = \sum_{p,\mu} \mu \, C^+_{p,\mu} C_{p,\mu}, \tag{2}$$

$$J_+ = \sum_{p} C^+_{p,+} C_{p,-}, \tag{3}$$

$$J_- = \sum_{p} C^+_{p,-} C_{p,+}, \tag{4}$$

together with the Casimir operator

$$J^2 = J_z^2 + \frac{1}{2}(J_+J_- + J_-J_+). \tag{5}$$

The eigenvalues of J^2 take the form $J(J+1)$ and the Lipkin Hamiltonian takes the form (v is a coupling constant) [6]

$$H = \epsilon J_z + \frac{v}{4}(J_+^2 + J_-^2). \tag{6}$$

The matrix of our Hamiltonian reads [6]

$$\begin{aligned} \langle n'|H_L|n\rangle &= \left\{\frac{N}{2} - n + 1 - \left(Nn - \frac{N}{2} - n^2 + 2n - 1\right)\omega\right\}\delta_{n',n} \\ &-\frac{v}{2}\sqrt{(N-n)(N-n+1)(n+1)n}\,\delta_{n',n+2} \\ &-\frac{v}{2}\sqrt{(N-n')(N-n'+1)(n'+1)n'}\,\delta_{n',n-2} \end{aligned} \tag{7}$$

with $n = 0, 1, \ldots, N$ for $J = N/2$. A diagonalization (numerical) gives the energetic eigenvalues $E_n(v, J)$.

2.1. System's External Environment: Gibb's Canonical Ensemble Heat Bath at Temperature T

Our Lipkin structure is supposed to interact with a heat reservoir. With the above eigenvalues [15,16] we construct a canonical ensemble partition function Z [15] from which we obtain all the associated thermal quantifiers that emerge [15,16].

We build up Z employing probabilities entering Z [16]. Our thermal indicators and Z derive from probability distributions [16] $p_n(v, J, \beta)$. β is, as we saw above, the inverse temperature. The pertinent formulas are found in [16]. If we denominate the mean energy U, the entropy S, and the free energy F, we obtain [15,16]

$$p_n(v, J, \beta) = \frac{1}{Z(v, J, \beta)} \mathrm{e}^{-\beta E_n(v,J)}, \tag{8}$$

$$Z(v, J, \beta) = \sum_{n=0}^{N} \mathrm{e}^{-\beta E_n(v,J)}, \tag{9}$$

$$\begin{aligned} U(v, J, \beta) &= \langle E\rangle = -\frac{\partial ln Z(v, J, \beta)}{\partial \beta} \\ &= \sum_{n=0}^{N} E_n(v, J) P_n(v, J, \beta) \\ &= \frac{1}{Z(v, J, \beta)} \sum_{n=0}^{N} E_n(v, J) \mathrm{e}^{-\beta E_n(v,J)}, \end{aligned} \tag{10}$$

$$S(v, J, \beta) = -\sum_{n=0}^{N} P_n(v, J, \beta) \ln[P_n(v, J, \beta)] \tag{11}$$

$$F(v, J, \beta) = U(v, J, \beta) - T\, S(v, J, \beta). \tag{12}$$

These thermal quantifiers yield much more information than that derived via the quantum quantifiers of zero temperature T [16]. As stated above, using a low enough T, our indicators provide a reasonably good representation of the $T = 0$ panorama [16]. Below, we often take $\beta = 10$ or even lower.

2.2. Lipkin's Degree of Mixture for a Quantum State ρ

Purity and mixing are basic notions that play a significant role in describing quantum systems like the Lipkin model. The two notions are very important in the analysis of

quantum information, computation, entanglement, and measurements. Understanding the difference between pure states and mixed ones requires an understanding of the coherence, superposition, and statistical behavior. As is well known, the degree of mixture C_f of a state ρ is, in terms of the probabilities p_i [17],

$$C_f = 1 - Tr\rho^2 = 1 - \sum_i p_i^2, \tag{13}$$

where $Tr\rho^2$ is the purity P_y (upper case P). For a pure state, one has $C_f = 0$ and $P_y = 1$. C_f is a very important quantifier in this study. $P_y = \sum_{n=0}^{N}(P_n(v,J,\beta))^2$ and $C_f = S_2 = 1 - P_y^2$.

We reiterate also, at the risk of redundancy, that for a Lipkin's state at the fixed temperature T, one uses, for the probabilities and partition function, all the important quantities

$$p_n(v,J,\beta) = \frac{1}{Z(v,J,\beta)} \mathrm{e}^{-\beta E_n(v,J)} Z(v,J,\beta) = \sum_{n=0}^{N} \mathrm{e}^{-\beta E_n(v,J)}. \tag{14}$$

3. The Kullback–Leibler Divergence (KL) [18]

This is an important tool of information theory. The Kullback–Leibler divergence (KL) quantifies the distinguishability between two states. For quantum states, the KL divergence can capture differences related to interference and entanglement, providing valuable information about the quantum system's properties. One finds KL applications in quantum information theory, quantum computing, and quantum communication, where understanding and quantifying it yields a measure of the difference between two probability distributions. In classical information theory, the KL divergence between probability distributions $P(x)$ and $Q(x)$ is given by:

$$DKL(P||Q) = \sum_x P(x) log(P(x)/Q(x)). \tag{15}$$

The KL divergence quantifies how much information is lost when using Q to approximate P. It is non-negative and equal to zero if and only if the two distributions are the same. In the context of quantum mechanics, the concept of KL divergence can be extended to compare two quantum states. For two quantum states described by density operators ρ and σ, the quantum Kullback–Leibler FF divergence is given by

$$D(\rho||\sigma) = Tr(\rho log(\rho) - \rho log(\sigma)). \tag{16}$$

4. New Quantum-Thermal Indicator for the Lipkin Model

This may sound strange at first, but consider the quantity (with length units)

$$\lambda = \frac{2\pi\hbar}{\sqrt{2\pi m k_B T}}, \tag{17}$$

called the de Broglie thermal length. Of course, there is no dimension of length in the Lipkin model, but our scenario contains fermions of mass m (system's mass Nm) at the temperature T so that a λ value can be computed and is up to us to show its relevance to the Lipkin model, which we do below. According to de Broglie convention, very small λ's should indicate classicality. This happens obviously for high T and large N as indicated by (17), as should intuitively be expected.

The proton mass is $m_p = 1.67262192 \times 10^{-27}$ kg. The Hydrogen atom's radius r_H is $\sim 10^{-10}$ m. One can plot λ versus Nm_p. Since $\hbar = 6.582119569 \times 10^{-16}$ Js, one obtains

$$\lambda = \sqrt{\frac{2\pi\hbar^2}{m_p}}\sqrt{\frac{\beta}{N}} = 0.0403419\sqrt{\frac{\beta}{N}} = 1.0857 \times 10^{20}\sqrt{\frac{1}{NT}}\ r_H, \tag{18}$$

where the center considers $k_B = 1$ and is expressed in meters, while the last equality is expressed in r_H units and we took $\beta = 1/k_B T$ with $k_B = 1.380648 \times 10^{-23} J/K$.

4.1. Relationship between λ and N

Figure 1 depicts λ versus the fermion number N. The latter diminishes, as expected, when N grows. We expect this behavior because it is well known that classicality begins to insinuate itself for large N [5].

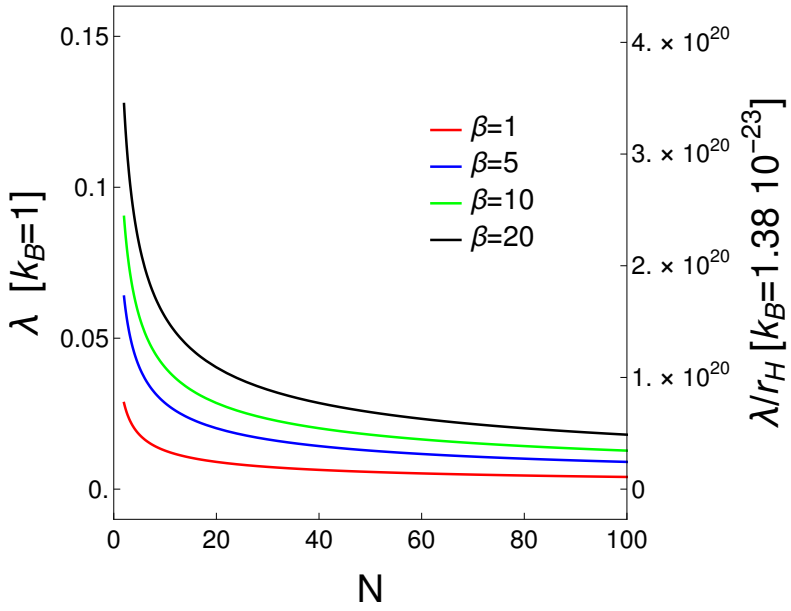


Figure 1. Thermal length quantifier λ *vs.* N for several values of $\beta = 1, 5, 10$, and 20 (in meters for the left vertical scale and in terms of the Hydrogen atom radius in the right vertical scale) taking $k_B = 1$ (**left**) and $kB = 1.380648 \times 10^{-23}$ (**right**). We see that the thermal length decreases as the number of fermions in the system grows, at rates regulated by the temperature. The rates increase as the temperature decreases. The length becomes smaller as N grows, which can be read as a vestige of classicality. That the effects are more noticeable for lower temperatures than for higher ones, is something that one should intuitively expect.

Let us repeat that the graph clearly indicates that our quantifier λ does work in the fashion one would expect. Indeed, as mentioned above, the well-known emergence of classicality as the number of particles N in a system grows is a fascinating aspect of quantum physics [5]. We will NOT encounter this phenomenon further in this work.

4.2. Relationship between Purity, Temperature, and λ

Figure 2 plots $P_y = 1 - C_f$ versus λ.

We see that, as promised above, the thermal length displays the expected properties of growing when the purity is large and diminishing when the mixture degree is large. Thus, λ is validated as a quantum indicator.

The limitation we observe in the Lipkin model, in which the degree of quantum mixture cannot exceed one-half likely arises from the combination of the Pauli exclusion principle and the specific characteristics of degenerate levels in the system. The Pauli exclusion principle restricts the occupancy of sites by fermions. No two fermions can occupy the same site simultaneously. In our case, each quasi-spin $p-$site can accommodate two fermions (one with spin up and one with spin down).

In the Lipkin model, each level can host multiple fermions. However, the maximum number of fermions that can occupy the pertinent degenerate levels is $2N$. This is because there are $2N$ available distinct sites.

When the system is half-filled, as is the case here, we have N fermions in total. This situation corresponds to one fermion in each of the twice degenerate p-spins, and it represents a scenario of maximum quantum mixture.

The Pauli exclusion principle still applies independently to each spin site–state, allowing for a maximum of two fermions (spin up and spin down) in each degenerate p spin.

In summary, the Pauli exclusion principle, combined with the degeneracy of the levels, results in a maximum quantum mixture when the system is half-filled by N fermions. This situation corresponds to one fermion in each twice degenerate p quasi-spin, leading to a state of maximum superposition. The specific structure of the Lipkin model, considering quais-spin degeneracy, determines the constraints on the occupancy and the resulting quantum mixture.

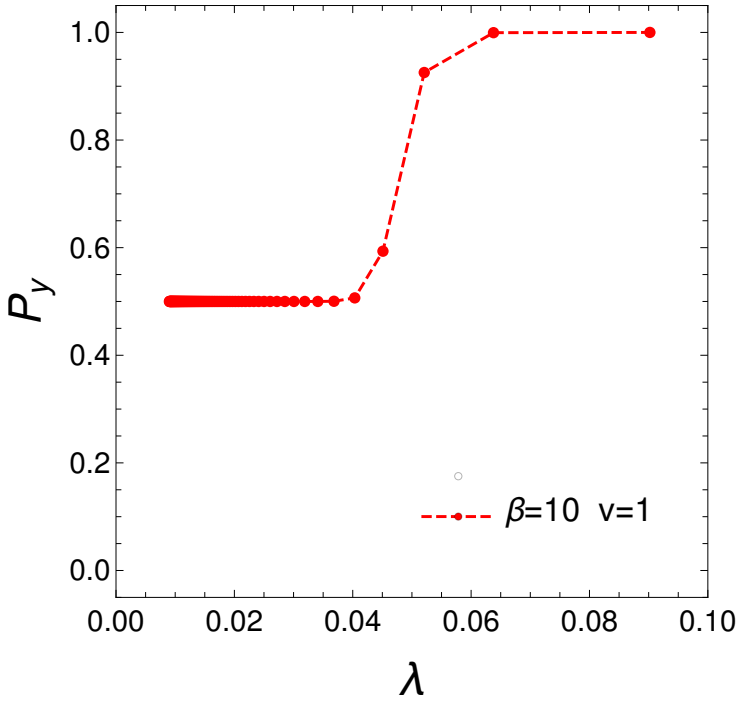


Figure 2. Purity versus de Broglie thermal wavelength (in meters) for $\beta = 10$, $N = 20$, and $v = 1$. When the purity is small, the wavelength is also small, and vice versa when the purity approaches unity. One detects a minimum purity equal to one-half (see text).

5. Information Loss and λ

5.1. The KL–Information Loss Relationship

We reframe here our Kullback–Leibler material in terms of the number of fermions N that our Lipkin system contains. The Kullback–Leibler (KL) divergence between two probability distributions P_1 and P_2 provides, as we saw, a measure of how one probability distribution diverges from another. Specifically, $KL(P_1||P_2)$ quantifies the information lost when P_2 is used to approximate P_1.In other words, the KL divergence measures the expected additional amount of information needed to encode data from P_1 using a probability distribution P_2 that is, it quantifies how much information is lost when the true distribution P_1 is approximated by P_2 In this work, the first distribution will be that describing an N-fermion system while P_2 does the same for an identical system but containing $2N$ particles.

5.2. The Degree of Information Loss as N changes

We tackle now an issue regarding the degree of information loss as N changes, using KL as a measure of how one probability distribution (PD) diverges from a second PD in the context of comparing systems with different particle numbers. We will find that the KL divergence diminishes when comparing a system with a relatively small particle number N to one with a larger N for quantum mixed systems. The diminution effect is drastically

smaller if the systems are in a pure state. This issue is addressed in Figure 3, which relates λ (in meters) to the KL divergence between N and $2N$ fermions ($N = 100$). Information is lost when N increases, but it does so in a very different manner if the state is mixed or pure. This fact constitutes the main finding of the present work.

The graph shows that this loss process also occurs at larger thermal lengths, the larger v becomes. The main feature here is that the degree of information loss is rather big for small thermal lengths. We contend that the loss continues to exist at large lengths. We do not see this occur in our graph for scale reasons. What becomes clear is that the larger the purity, the smaller the information loss.

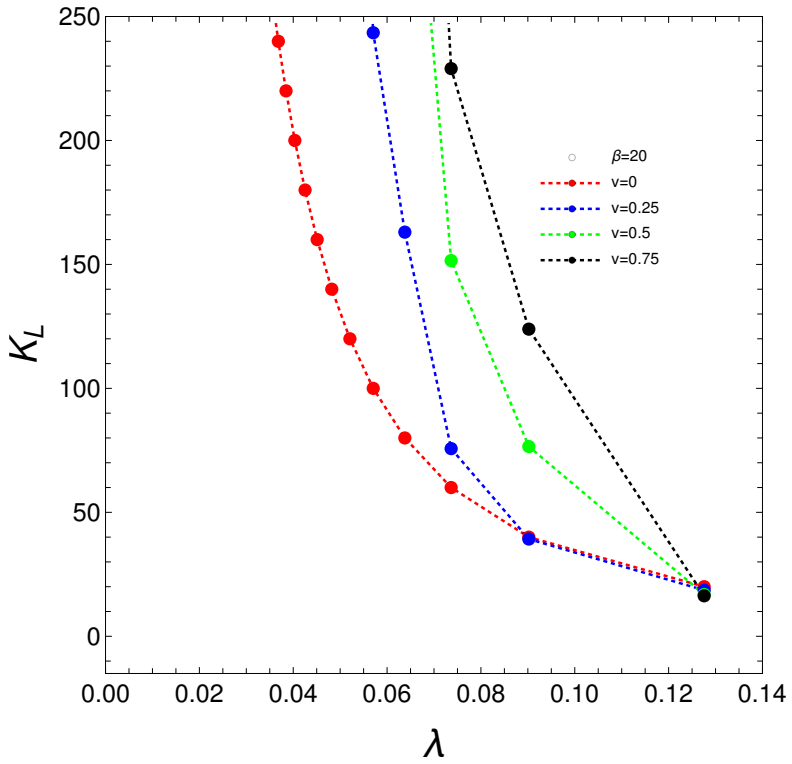


Figure 3. *KL vs.* λ for several v values, obtained by comparing probability distributions pertaining to N versus $2N$ scenarios. These KL values measure our information loss. We take $\beta = 20$. When we reach $\lambda \sim 13$ cm, we are in the pure-state range of our indicator, as we saw above. In this pure-state regime, the information loss is so small that it cannot be seen in the vertical scale we have chosen here. Differences are smaller than the dots used there. We see that information loss when we augment N is immensely larger for mixed states than for pure ones.

5.3. Possible Interpretations

In quantum mechanics, entanglement is a unique feature that can exist in pure states. When a system is in a pure state, the entanglement between particles can lead to more complex and correlated probability distributions. In contrast, mixed states arise from ensembles of quantum systems, potentially with reduced entanglement. The presence of mixed states might lead to a more gradual change in the probability distribution with increasing particle number, resulting in a smaller KL divergence.

Mixed states can be viewed as statistical ensembles of pure states. When comparing a system of small particle number to one with a large particle number in a mixed-state scenario, you are effectively considering an average over a variety of possible pure states within the ensemble. This averaging might lead to a smoother and less pronounced change in the probability distribution, contributing to a smaller KL divergence.

Quantum mixtures can exhibit reduced correlations between particles compared to entangled pure states. As the particle number increases, the statistical independence

between particles might become more pronounced, leading to a less complex and more predictable probability distribution.

Mixed states often have higher entropy than pure states. The increased entropy may contribute to a more uniform and less informative probability distribution, particularly as the particle number increases. This can result in a smaller KL divergence.

6. Conclusions

We have encountered different ways of visualizing the difference between mixed and pure states and also introduced a new quantifier for the quantum mixture degree, called λ, that can be cast as a length.

For the systems considered in this work, λ-values larger than $\sim$12 cm indicate purity equal to unity while smaller values indicate increasing degrees of state mixing (see Figure 3). It should be noted that, according to Equation (18), λ values smaller than, say, 0.001 mm, might be evidence of the vestiges of classical behavior. Thus, we can properly call our λ a pseudo thermal length that is a helpful purity indicator for the Lipkin model. For high purity, it yields large values, and vice versa for a large degree of of state mixing.

We saw that, at a finite temperature T, the system purity P_y diminishes in a faster fashion when the degree of fermion–fermion interaction v becomes larger.

We discovered that, in the Lipkin model at a finite temperature T, the growth of N increases the degree of mixture. If one augments N, information, as measured by the KL divergence, is lost. The larger the mixture degree, the greater the accompanying information loss. In the case of pure states (at zero temperature), the loss becomes vanishingly small.

Author Contributions: Investigation, D.M., A.P. and A.R.P.; Project administration, A.P.; Writing—original draft, D.M., A.P. and A.R.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially supported by CONICET (National Scientific and Technical Research Council of Argentina).

Data Availability Statement: All data are contained in the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Bengtsson, I.; Zyczkowsi, K. *Geometry of Quantum States: An Introduction to Quantum Entanglement*; Cambridge University Press: Cambridge, MA, USA, 2006.
2. Jaeger, G. *Quantum Information: An Overview*; Springer: Berlin/Heidelberg, Germany, 2007.
3. Giulini, D.; Joos, E.; Kiefer, C.; Kupsch, J.; Stamatescu, I.-O.; Zeh, H.D. *Decoherence and the Appearance of a Classical World in Quantum Theory*; Springer: Berlin/Heidelberg, Germany, 1996.
4. Facchi, P.; Mariano, A.; Pascazio, S. Decoherence versus entropy in neutron interferometry. *Phys. Rev. A* **2001**, *63*, 052108. [CrossRef]
5. Zurek, W.H. Decoherence and the Transition from Quantum to Classical—Revisited. *Phys. Today* **1991**, *44*, 1–31. [CrossRef]
6. Lipkin, H.J.; Meshkov, N.; Glick, A.J. Validity of many-body approximation methods for a solvable model: (I) Exact solutions and perturbation theory. *Nucl. Phys.* **1965**, *62*, 188. [CrossRef]
7. Lipkin, H.J.; Meshkov, N.; Glick, A.J. Validity of many-body approximation methods for a solvable model: (II). Linearization procedures. *Nucl. Phys.* **1965**, *62*, 199. [CrossRef]
8. Lipkin, H.J.; Meshkov, N.; Glick, A.J. Validity of many-body approximation methods for a solvable model: (III). Exact solutions and perturbation theory. *Nucl. Phys.* **1965**, *62*, 188. [CrossRef]
9. Lipkin, H.J.; Meshkov, N.; Glick, A.J. Validity of many-body approximation methods for a solvable model: (IV). The deformed Hartree-Fock solution. *Nucl. Phys.* **1965**, *62*, 220. [CrossRef]
10. Rowe, D.J. The Lipkin model. *Rev. Mod. Phys.* **1968**, *40*, 153. [CrossRef]
11. Schmid, K.W.; Bauer, A. The Lipkin model. *Phys. Rep.* **1982**, *88*, 251–334.
12. Isacker, P.V.; Heinze, S. The Lipkin model: A paradigm for a self-consistent theory of shape transitions. *Phys. Rep.* **1991**, *199*, 67.
13. Dukelsky, J.; Pittel, S.; Sierra, G. Quantum phase transitions in the Lipkin-Meshkov-Glick model. *Rev. Mod. Phys.* **2004**, *76*, 643. [CrossRef]
14. Li, J. Quantum phase transition and phase diagram in the Lipkin-Meshkov-Glick model. *Phys. Rev. B* **2017**, *95*, 075139.

15. Plastino, A.R.; Monteoliva, D.; Plastino, A. Information-theoretic features of many fermion systems: An exploration based on exactly solvable models. *Entropy* **2021**, *23*, 1488. [CrossRef] [PubMed]
16. Monteoliva, D.; Plastino, A.; Plastino, A.R. Statistical Quantifiers Resolve a Nuclear Theory Controversy. *Q. Rep.* **2022**, *4*, 127. [CrossRef]
17. Ring, P.; Schuck, P. *The Nuclear Many-Body Problem*; Springer: Berlin/Heidelberg, Germany, 1980.
18. Sanchez-Moreno, P.; Zarzo, A.; Dehesa, J.S. Jensen divergence based on Fisher's information. *J. Phys. A Math. Theor.* **2012**, *45*, 125305. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

AppliedMath

MDPI

Article

A G-Modified Helmholtz Equation with New Expansions for the Earth's Disturbing Gravitational Potential, Its Functionals and the Study of Isogravitational Surfaces

Gerassimos Manoussakis

Department of Mathematics, School of Applied Mathematical and Physical Sciences, National Technical University of Athens, Iroon Polytechneiou 9, 15780 Zografos, Greece; gmanous@math.ntua.gr

Abstract: The G-modified Helmholtz equation is a partial differential equation that enables us to express gravity intensity g as a series of spherical harmonics having radial distance r in irrational powers. The Laplace equation in three-dimensional space (in Cartesian coordinates, is the sum of the second-order partial derivatives of the unknown quantity equal to zero) is used to express the Earth's gravity potential (disturbing and normal potential) in order to represent other useful quantities—which are also known as functionals of the disturbing potential—such as gravity disturbance, gravity anomaly, and geoid undulation as a series of spherical harmonics. We demonstrate that by using the G-modified Helmholtz equation, not only gravity intensity but also disturbing potential and its functionals can be expressed as a series of spherical harmonics. Having gravity intensity represented as a series of spherical harmonics allows us to create new Global Gravity Models. Furthermore, a more detailed examination of the Earth's isogravitational surfaces is conducted. Finally, we tabulate our results, which makes it clear that new Global Gravity Models for gravity intensity g will be very useful for many geophysical and geodetic applications.

Keywords: gravity; gravity anomaly; gravity disturbance; vertical gradient of gravity; disturbing potential; spherical harmonics; isogravitational surfaces

MSC: 35A09; 35C10; 35J25; 33C05; 33C75



Citation: Manoussakis, G. A G-Modified Helmholtz Equation with New Expansions for the Earth's Disturbing Gravitational Potential, Its Functionals and the Study of Isogravitational Surfaces. *AppliedMath* **2024**, *4*, 580–595. https://doi.org/10.3390/appliedmath4020032

Academic Editor: Thomas Woolley

Received: 7 March 2024
Revised: 24 April 2024
Accepted: 26 April 2024
Published: 4 May 2024



Copyright: © 2024 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Spherical harmonics are important in many theoretical and practical applications (electromagnetic fields, electron configurations, gravitational fields, geoids, computer graphics, cosmic microwave background radiation, etc.). Some of these applications are briefly mentioned in this work. Spherical harmonics have been extensively studied and are widely applicable due to their numerous convenient properties. In computer graphics [1], they are quite useful since they share many of the same strengths as Haar wavelets, and there are techniques that bridge the gap between them. This enables the representation of low-frequency and high-frequency functions. Additionally, in [2–4], they are used in the formation of a realistic global lightning model.

Gravity potential is a quantity expressed as a series of spherical harmonics. This series is utilized for formulating Global Gravity Models and Local Gravity Models. A Global Gravity Model [5], also referred to as a Global Geopotential Model, serves as a mathematical representation of the external gravitational potential of a celestial body. Here, we focus on the scenario where the Earth is the attracting body. The use of Global Gravity Models enables us to determine all the related gravity field functionals—for example, gravity vectors and gravity.

It is important to note that these models [6–9] are crucial as they offer valuable insights into Earth's features, including geoid undulation, gravity anomalies, gravity disturbances,

and vertical deflection. It is worth mentioning that these models have wide-ranging applications, including orbit determination, inertial navigation for airplanes and missiles, oceanographic studies of ocean circulation, and geophysical research into density distributions.

The Earth's magnetic vector field [10] is related to the scalar geomagnetic potential that obeys the Laplace equation and can be represented as a series of spherical harmonics. The applications of the geomagnetic potential [11,12] are incredibly vast; they range from studying hydromagnetic wave propagation and ionospheric currents to investigating tides, monitoring changes in tropospheric weather, and exploring the Earth's crust.

The intensity of gravity and electrostatic forces can be mathematically expressed as a series of spherical harmonics. This series arises from the solution of the G-modified Helmholtz equation. The intensity of electrostatics plays a critical role in a diverse array of applications [13], ranging from heat and mass transfer to the synthesis of nanomaterials and the reduction inCO_2 emissions. Furthermore, spheroids [14–16] have been employed to address electromagnetic scattering problems and analyze electrostatic interactions between spheroidal particles in different scenarios. The intensity of gravity is very valuable in various applications. It plays a crucial role [17,18] in detecting seismic faults, estimating simple spherical structures within the Earth, determining geoid and quasi-geoid (which define the physical shape of the Earth and reference surface of physical heights), [19] analyzing sea surface topography, understanding the structure of the lithosphere, inferring the thickness of floating ice, and [20] even detecting archaeological structures during exploration.

In this study, we present a novel method for determining the disturbing potential that is independent of the Laplace equation. By expressing [21] gravity anomaly and gravity disturbance as series of spherical harmonics using the modified Helmholtz equation (with the radial distance, r, raised to irrational powers), we can effortlessly determine the disturbing potential through a simple partial differential equation. This new approach allows us to represent relative quantities associated with gravitational potential (i.e., its functionals) in a new series of spherical harmonics. In addition, by expressing gravity intensity as a series of spherical harmonics, we can conduct a more thorough analysis of isogravitational surfaces. It is important to emphasize that, in this context, the Earth is regarded as a non-rotating body.

2. Laplace Equation, Disturbing Potential, and Relative Quantities

Focusing on the case of the Earth, the disturbing potential T [22] (which is the difference between Earth's actual gravity potential W and the normal potential U generated by a suitable ellipsoid of revolution) is a solution of the Laplace equation.

$$\nabla^2 T = 0 \tag{1}$$

The concept of disturbing potential holds significant importance, as it is vital for calculating the geoid undulation N, which represents the distance between the geoid's surface and the surface of a suitable ellipsoid of revolution. Additionally, it plays a crucial role in determining other important quantities such as gravity disturbance δg, gravity anomaly Δg, and deflections of the vertical. There are three classical problems [5] for determining the disturbing potential T from boundary values on a sphere: (a) the Poisson problem, (b) the Hotine problem, and (c) the Stokes problem.

The first problem is a Dirichlet problem on a sphere with a radius of R, where the values of T are given on the sphere. The second problem, as a spherical approximation, involves solving a Neumann problem. In this problem, we are provided with the values of gravity disturbance δg (representing the radial derivative of the disturbing potential T). The third problem (as a spherical approximation) is a Robin problem in which the values of gravity anomaly Δg are given on the geoid fulfilling the following relation:

$$\frac{\partial T}{\partial r} + \frac{2}{R} T = -\Delta g \tag{2}$$

Instead of solving boundary value problems to determine the disturbing potential T, we use Global Geopotential Models (GGMs). A Global Geopotential Model [5,23] is a mathematical expression that enables the calculation of the disturbing potential through spherical harmonic expansion. The harmonic series of a GGM is finite and also provides a collection of mathematical expressions, numerical values, and algorithms. With these tools, users can easily perform computations to determine numerical values of quantities associated with the disturbing potential (such as gravity disturbance and gravity anomaly). Additionally, these computations enable the evaluation of specific errors that are linked to these quantities.

We present some finite series expressions [22–24], in spherical coordinates (r, θ, λ) (θ is measured from the z-axis, (normalized coefficients).

$$\begin{aligned} T(r,\theta,\lambda) &= \sum_{n=2}^{N_{\max}} \left(\tfrac{R}{r}\right)^{n+1} T_n(\theta,\lambda) = \\ &= \tfrac{GM}{r} \sum_{n=2}^{N_{\max}} \sum_{m=0}^{n} \left(\tfrac{R}{r}\right)^{n+1} (\overline{C}_{nm}\cos m\lambda + \overline{S}_{nm}\sin m\lambda)\overline{P}_{nm}(\cos\theta) \end{aligned} \tag{3}$$

$$\delta g(r,\theta,\lambda) = -\frac{\partial T}{\partial r} = \frac{GM}{r^2} \sum_{n=0}^{N_{\max}} (n+1)\left(\frac{R}{r}\right)^{n+1} T_n(\theta,\lambda) \tag{4}$$

$$\Delta g(r,\theta,\lambda) = \frac{GM}{r^2} \sum_{n=0}^{N_{\max}} (n-1)\left(\frac{R}{r}\right)^{n+1} T_n(\theta,\lambda) \tag{5}$$

Geoid undulation is calculated by

$$N(\theta,\lambda) = \frac{1}{\gamma_{ellispoid}(\theta,\lambda)} T(R,\theta,\lambda) \tag{6}$$

where γ is the value of normal gravity [22] on the ellipsoid, G is the gravitational constant and M is the mass of the Earth. Components of the deflection of the vertical [23] are

$$\xi(r,\theta,\lambda) = -\frac{1}{r\gamma(r,\theta,\lambda)} \frac{\partial T}{\partial \theta} \tag{7}$$

$$\eta(r,\theta,\lambda) = -\frac{1}{r\gamma(r,\theta,\lambda)\sin\theta} \frac{\partial T}{\partial \lambda} \tag{8}$$

$$\frac{\partial T}{\partial \theta} = \frac{GM}{r} \sum_{n=2}^{N_{\max}} \sum_{m=0}^{n} \left(\frac{R}{r}\right)^{n+1} (\overline{C}_{nm}\cos m\lambda + \overline{S}_{nm}\sin m\lambda)(\overline{P}_{nm+1}(\cos\theta) - m\tan\theta \cdot \overline{P}_{nm}(\cos\theta)) \tag{9}$$

$$\frac{\partial T}{\partial \lambda} = -\frac{GM}{r} \sum_{n=2}^{N_{\max}} \sum_{m=0}^{n} \left(\frac{R}{r}\right)^{n+1} m(\overline{C}_{nm}\sin m\lambda - \overline{S}_{nm}\cos m\lambda)\overline{P}_{nm}(\cos\theta) \tag{10}$$

In Equations (7) and (8) normal gravity [22] is determined on a suitable point which is located on an appropriate equipotential surface of the normal gravity field. The disturbing potential, T, is a quantity that cannot be directly measured. Therefore, other measurable quantities are required to determine its value. However, we will demonstrate an alternative method for determining T in the following paragraph.

3. G-Modified Helmholtz Equation, Disturbing Potential, and Relative Quantities

Assuming that the gravity disturbance δg is a continuous function on the surface of the geoid, we can determine it by solving the Dirichlet problem that involves the G-modified Helmholtz equation.

$$\begin{aligned} &\nabla^2(\delta g) - \tfrac{2}{r^2}\delta g = 0 \\ &\delta g|_S = f_1(\theta\prime,\lambda) \end{aligned} \tag{11}$$

The expression of gravity disturbance in spherical harmonics [21] is as follows:

$$\delta g(r,\theta\prime,\lambda) = \sum_{n=0}^{+\infty}\sum_{m=0}^{n} r^{-\frac{1+\sqrt{9+4n(n+1)}}{2}}[a_{nm}^{\delta g}P_{nm}(\sin\theta\prime)\cos m\lambda + b_{nm}^{\delta g}P_{nm}(\sin\theta\prime)\sin m\lambda] \quad (12)$$

According to [22], on the surface of the geoid, which, as a spherical approximation, corresponds to a sphere of radius R, we can derive

$$f_1(\theta\prime,\lambda) = \sum_{n=0}^{+\infty}\sum_{m=0}^{n} R^{-\frac{1+\sqrt{9+4n(n+1)}}{2}}[a_{nm}^{\delta g}P_{nm}(\sin\theta\prime)\cos m\lambda + b_{nm}^{\delta g}P_{nm}(\sin\theta\prime)\sin m\lambda] \quad (13)$$

Putting

$$A_{nm}^{\delta g} = R^{-\frac{1+\sqrt{9+4n(n+1)}}{2}} a_{nm}^{\delta g} \quad (14)$$

$$B_{nm}^{\delta g} = R^{-\frac{1+\sqrt{9+4n(n+1)}}{2}} b_{nm}^{\delta g} \quad (15)$$

The series in Equation (13) becomes

$$f_1(\theta\prime,\lambda) = \sum_{n=0}^{+\infty}\sum_{m=0}^{n}[A_{nm}^{\delta g}P_{nm}(\sin\theta\prime)\cos m\lambda + B_{nm}^{\delta g}P_{nm}(\sin\theta\prime)\sin m\lambda] \quad (16)$$

Thus,

$$\begin{aligned}&\int_0^{2\pi}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} f_1(\theta\prime,\lambda)P_{jk}(\sin\theta\prime)\sin\theta\prime d\lambda d\theta\prime = \\ &= \int_0^{2\pi}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}}\left\{\sum_{n=0}^{+\infty}\sum_{m=0}^{n}[A_{nm}^{\delta g}P_{nm}(\sin\theta\prime)\cos m\lambda + B_{nm}^{\delta g}P_{nm}(\sin\theta\prime)\sin m\lambda]\right\}\cdot \\ &\quad \cdot P_{jk}(\sin\theta\prime)\sin\theta\prime d\lambda d\theta\prime\end{aligned} \quad (17)$$

and the coefficients [22] of series (17) are equal to

$$A_{n0}^{\delta g} = \frac{2n+1}{4\pi}\int_0^{2\pi}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} f_1(\theta\prime,\lambda)P_n(\sin\theta\prime)\sin\theta\prime d\lambda d\theta\prime \quad (18)$$

$$A_{nm}^{\delta g} = \frac{2n+1}{4\pi}\frac{(n-m)!}{(n+m)!}\int_0^{2\pi}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} f_1(\theta\prime,\lambda)P_{nm}(\sin\theta\prime)\cos m\lambda\sin\theta\prime d\lambda d\theta\prime \quad , \quad m \neq 0 \quad (19)$$

$$B_{nm}^{\delta g} = \frac{2n+1}{4\pi}\frac{(n-m)!}{(n+m)!}\int_0^{2\pi}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} f_1(\theta\prime,\lambda)P_{nm}(\sin\theta\prime)\sin m\lambda\sin\theta\prime d\lambda d\theta\prime \quad (20)$$

Hence, the desirable expression for gravity disturbance is

$$\delta g(r,\theta\prime,\lambda) = \sum_{n=0}^{+\infty}\sum_{m=0}^{n}\left(\frac{R}{r}\right)^{\frac{1+\sqrt{9+4n(n+1)}}{2}}[A_{nm}^{\delta g}P_{nm}(\sin\theta\prime)\cos m\lambda + B_{nm}^{\delta g}P_{nm}(\sin\theta\prime)\sin m\lambda] \quad (21)$$

Gravity anomaly can also be identified through a similar Dirichlet problem if the values of gravity anomaly Δg are known on the geoid's surface.

$$\begin{aligned}&\nabla^2(\Delta g) - \tfrac{2}{r^2}\Delta g = 0 \\ &\Delta g_S = f_2(\theta\prime,\lambda)\end{aligned} \quad (22)$$

The expression of gravity anomaly is

$$\Delta g(r,\theta\prime,\lambda)=\sum_{n=0}^{+\infty}\sum_{m=0}^{n}\left(\frac{R}{r}\right)^{\frac{1+\sqrt{9+4n(n+1)}}{2}}[A_{nm}^{\Delta g}P_{nm}(\sin\theta\prime)\cos m\lambda+B_{nm}^{\Delta g}P_{nm}(\sin\theta\prime)\sin m\lambda] \tag{23}$$

Now, we are ready to find the disturbing potential. The function T will be determined [22] from the following partial differential equation:

$$\begin{aligned}&\tfrac{\partial T}{\partial r}=-\delta g\Leftrightarrow\\&\Leftrightarrow\tfrac{\partial T}{\partial r}=-\sum_{n=0}^{+\infty}\sum_{m=0}^{n}\left(\tfrac{R}{r}\right)^{\frac{1+\sqrt{9+4n(n+1)}}{2}}[A_{nm}^{\delta g}P_{nm}(\sin\theta\prime)\cos m\lambda+B_{nm}^{\delta g}P_{nm}(\sin\theta\prime)\sin m\lambda]\end{aligned} \tag{24}$$

The solution of this partial differential equation is

$$\begin{aligned}&T(r,\theta\prime,\lambda)=\\&=\sum_{n=0}^{+\infty}\tfrac{2R}{\sqrt{9+4n(n+1)}-1}\left(\tfrac{R}{r}\right)^{\frac{\sqrt{9+4n(n+1)}-1}{2}}\\&\left\{\sum_{m=0}^{n}[A_{nm}^{\delta g}P_{nm}(\sin\theta\prime)\cos m\lambda+B_{nm}^{\delta g}P_{nm}(\sin\theta\prime)\sin m\lambda]\right\}+f_{\delta g}(\theta\prime,\lambda)\end{aligned} \tag{25}$$

The function $f_{\delta g}$ is an arbitrary function, but it can be determined from the following boundary condition [22]:

$$\delta g=\Delta g+\frac{2}{R}T \tag{26}$$

Substituting the necessary quantities from Equations (21), (23), and (25) we have

$$\begin{aligned}&\sum_{n=0}^{+\infty}\sum_{m=0}^{n}[A_{nm}^{\delta g}P_{nm}(\sin\theta\prime)\cos m\lambda+B_{nm}^{\delta g}P_{nm}(\sin\theta\prime)\sin m\lambda]=\\&=\sum_{n=0}^{+\infty}\sum_{m=0}^{n}[A_{nm}^{\Delta g}P_{nm}(\sin\theta\prime)\cos m\lambda+B_{nm}^{\Delta g}P_{nm}(\sin\theta\prime)\sin m\lambda]+\\&+\sum_{n=0}^{+\infty}\tfrac{4}{\sqrt{9+4n(n+1)}-1}\sum_{m=0}^{n}[A_{nm}^{\delta g}P_{nm}(\sin\theta\prime)\cos m\lambda+B_{nm}^{\delta g}P_{nm}(\sin\theta\prime)\sin m\lambda]+\\&+\tfrac{2}{R}f_{\delta g}(\theta\prime,\lambda)\end{aligned} \tag{27}$$

Therefore, the function $f_{\delta g}$ is expressed as a series of spherical harmonics.

$$f_{\delta g}(\theta\prime,\lambda)=\sum_{n=0}^{+\infty}\sum_{m=0}^{n}[A_{nm}P_{nm}(\sin\theta\prime)\cos m\lambda+B_{nm}P_{nm}(\sin\theta\prime)\sin m\lambda] \tag{28}$$

The coefficients of the above series are equal to

$$A_{nm}=\frac{R}{2}\left[\frac{\sqrt{9+4n(n+1)}-5}{\sqrt{9+4n(n+1)}-1}A_{nm}^{\delta g}-A_{nm}^{\Delta g}\right] \tag{29}$$

$$B_{nm}=\frac{R}{2}\left[\frac{\sqrt{9+4n(n+1)}-5}{\sqrt{9+4n(n+1)}-1}B_{nm}^{\delta g}-B_{nm}^{\Delta g}\right] \tag{30}$$

Finally, the disturbing potential T is given as

$$\begin{aligned}&T(r,\theta\prime,\lambda)=\\&=\sum_{n=0}^{+\infty}\tfrac{2R}{\sqrt{9+4n(n+1)}-1}\left(\tfrac{R}{r}\right)^{\frac{\sqrt{9+4n(n+1)}-1}{2}}\\&\left\{\sum_{m=0}^{n}[A_{nm}^{\delta g}P_{nm}(\sin\theta\prime)\cos m\lambda+B_{nm}^{\delta g}P_{nm}(\sin\theta\prime)\sin m\lambda]\right\}+\\&+\sum_{n=0}^{+\infty}\sum_{m=0}^{n}\tfrac{R}{2}\left[\left(\tfrac{\sqrt{9+4n(n+1)}-5}{\sqrt{9+4n(n+1)}-1}A_{nm}^{\delta g}-A_{nm}^{\Delta g}\right)P_{nm}(\sin\theta\prime)\cos m\lambda+\right.\\&\left.+\left(\tfrac{\sqrt{9+4n(n+1)}-5}{\sqrt{9+4n(n+1)}-1}B_{nm}^{\delta g}-B_{nm}^{\Delta g}\right)P_{nm}(\sin\theta\prime)\sin m\lambda\right]\end{aligned} \tag{31}$$

The procedure of finding Equation (31) can be named the alternative gravimetric determination of the disturbing potential.

The expression for the disturbing potential given by Equation (31) can be replaced with a finite series of spherical harmonics in order to develop a new Global Geopotential Model. Following geodetic terminology, in this model, the disturbing potential is a functional of gravity disturbances. For n = N_{max}, Equation (31) becomes a finite series and allows us to determine the third-order partial derivatives of the disturbing potential in spherical coordinates. These derivatives can be expressed in different Cartesian systems [25] and effectively used in various geophysical studies and explorations.

In Equation (31), for r = R, the disturbing potential is determined on the surface of the geoid. The result is as follows:

$$T(R,\theta\prime,\lambda) \equiv T(\theta\prime,\lambda) = \frac{R}{2}\Big\{[A_{nm}^{\delta g}P_{nm}(\sin\theta\prime)\cos m\lambda + B_{nm}^{\delta g}P_{nm}(\sin\theta\prime)\sin m\lambda] - \\ -[A_{nm}^{\Delta g}P_{nm}(\sin\theta\prime)\cos m\lambda + B_{nm}^{\Delta g}P_{nm}(\sin\theta\prime)\sin m\lambda]\Big\} \tag{32}$$

The above formula can be written as

$$T(\theta\prime,\lambda) = \frac{R}{2}[\delta g(\theta\prime,\lambda) - \Delta g(\theta\prime,\lambda)] \tag{33}$$

Hence, on the surface of the geoid, the disturbing potential is equal to the difference between gravity disturbance and gravity anomaly, multiplied by R/2.Using Equation (6), the geoid undulation can be expressed in a new alternative form.

$$N(\theta\prime,\lambda) = \frac{R}{2\gamma(\theta\prime,\lambda)}[\delta g(\theta\prime,\lambda) - \Delta g(\theta\prime,\lambda)] \tag{34}$$

Equations (33) and (34) also represent the significance of gravity disturbance and gravity anomaly (determined by measurements). An interesting theoretical result occurs if δg and Δg are substituted with their equals.

$$T(r,\theta\prime,\lambda) = \frac{R}{2}[\gamma_{ellepsoid}(\theta\prime,\lambda) - \gamma_{geoid}(r,\theta\prime,\lambda)] \tag{35}$$

$$N(r,\theta\prime,\lambda) = \frac{R}{2}\left(\frac{\gamma_{ellepsoid}(\theta\prime,\lambda) - \gamma_{geoid}(r,\theta\prime,\lambda)}{\gamma_{ellepsoid}(\theta\prime,\lambda)}\right) \tag{36}$$

Therefore, if the value of normal gravity γ is known on the surface of the geoid, then the disturbing potential (on this surface) and the geoid's undulation are known (again as a series of spherical harmonics).

The intensity of gravity g is a highly significant quantity that can be expressed as a series of spherical harmonics.

$$g(r,\theta\prime,\lambda) = \sum_{n=0}^{+\infty}\sum_{m=0}^{n}\left(\frac{R}{r}\right)^{\frac{1+\sqrt{9+4n(n+1)}}{2}}[A_{nm}^{g}P_{nm}(\sin\theta\prime)\cos m\lambda + B_{nm}^{g}P_{nm}(\sin\theta\prime)\sin m\lambda] \tag{37}$$

This formula is much more manageable than the typical formula.

$$g(r,\theta\prime,\lambda) = \|gradV_E\| \tag{38}$$

where V_E stands for the Earth's Newtonian gravity potential. Equation (33) is also valuable for various conversions between different types of heights, such as orthometric height correction and dynamical correction.

As a spherical approximation, the vertical gradient of gravity is equal to the radial derivative of gravity intensity; thus [22] (we remind that we are dealing with a non-rotating Earth),

$$\frac{\partial g}{\partial r} = -2gJ \tag{39}$$

where J is the mean curvature of the relative equipotential surface.

$$\frac{\partial g}{\partial r} = \sum_{n=0}^{+\infty} \sum_{m=0}^{n} -\frac{1+\sqrt{9+4n(n+1)}}{2r} \left(\frac{R}{r}\right)^{\frac{1+\sqrt{9+4n(n+1)}}{2}} \left[A_{nm}^{g} P_{nm}(\sin\theta\prime)\cos m\lambda + B_{nm}^{g} P_{nm}(\sin\theta\prime)\sin m\lambda\right] \quad , \quad r > R \tag{40}$$

By combining Equations (39) and (40), we can determine the mean curvature of the equipotential surfaces of the Earth's gravity field.

4. Study of the Earth's Isogravitational Surfaces

Equation (37) enables us to investigate the isogravitational surfaces of the Earth's gravity field. On these surfaces, the value of gravity remains constant. As indicated in [26], isogravitational surfaces closely resemble spheres. In order to conduct a localized study of the isogravitational surfaces, we initiate the following transformation:

$$\begin{array}{l} X = r\cos\theta\prime\cos\lambda \\ Y = r\cos\theta\prime\sin\lambda \\ Z = r\sin\theta\prime \end{array} \quad , \quad \theta\prime = \frac{\pi}{2} - \theta \tag{41}$$

The Cartesian system (X, Y, Z) [22] has its center at the Earth's center of gravity, the Z-axis coincides with the Earth's mean axis of rotation, the X-axis is on the meridian plane of Greenwich pointing outwards, and the Y-axis makes the system right-handed. Then,

$$\frac{\partial(X,Y,Z)}{\partial(r,\theta\prime,\lambda)} = \begin{bmatrix} \cos\theta\prime\cos\lambda & \cos\theta\prime\sin\lambda & \sin\theta\prime \\ -r\sin\theta\prime\cos\lambda & -r\sin\theta\prime\sin\lambda & r\cos\theta\prime \\ -r\cos\theta\prime\sin\lambda & r\cos\theta\prime\cos\lambda & 0 \end{bmatrix} \tag{42}$$

The unit normal vectors are

$$\bar{\varepsilon}_1 = (\cos\theta\prime\cos\lambda, \cos\theta\prime\sin\lambda, \sin\theta\prime) \tag{43}$$

$$\bar{\varepsilon}_2 = (-\sin\theta\prime\cos\lambda, -\sin\theta\prime\sin\lambda, \cos\theta\prime) \tag{44}$$

$$\bar{\varepsilon}_3 = (-\sin\lambda, \cos\lambda, 0) \tag{45}$$

Let P an arbitrary point with spherical coordinates (r_P, $\theta_P{}'$, λ_P). We define a second Cartesian system (x, y, z), such that its center is at point P, and the z–axis has the same direction asthe vector ε_1 (Equation (43)). The x-axis is tangent to the meridian of a sphere of radius r_P and center the Earth's center of gravity, and it has the same direction as the vector ε_2. The y-axis makes the system right-handed. The transformation between (X, Y, Z) and (x, y, z) [27] is

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} -\sin\theta\prime\cos\lambda & \sin\lambda & \cos\theta\prime\cos\lambda \\ -\sin\theta\prime\sin\lambda & -\cos\lambda & \cos\theta\prime\sin\lambda \\ \cos\theta\prime & 0 & \sin\theta\prime \end{bmatrix}_P \begin{bmatrix} x \\ y \\ z + r_P \end{bmatrix} \tag{46}$$

The first- and second-order partial derivatives of gravity intensity expressed in the (x, y, z) system [27] are given as follows:

$$\frac{\partial g}{\partial x} = -\frac{1}{r}\frac{\partial g}{\partial \theta\prime} \tag{47}$$

$$\frac{\partial g}{\partial y} = -\frac{1}{r\cos\theta\prime}\frac{\partial g}{\partial \lambda} \tag{48}$$

$$\frac{\partial g}{\partial z} = \frac{\partial g}{\partial r} \tag{49}$$

$$\frac{\partial^2 g}{\partial x^2} = \frac{1}{r}\frac{\partial g}{\partial r} + \frac{1}{r^2}\frac{\partial^2 g}{\partial \theta\prime^2} \tag{50}$$

$$\frac{\partial^2 g}{\partial x \partial y} = \frac{1}{r^2\cos\theta\prime}\frac{\partial^2 g}{\partial \theta\prime \partial \lambda} - \frac{\sin\theta\prime}{r^2\cos^2\theta\prime}\frac{\partial g}{\partial \lambda} \tag{51}$$

$$\frac{\partial^2 g}{\partial x \partial z} = \frac{1}{r^2}\frac{\partial g}{\partial \theta\prime} - \frac{1}{r}\frac{\partial g}{\partial r \partial \theta\prime} \tag{52}$$

$$\frac{\partial^2 g}{\partial y^2} = \frac{1}{r}\frac{\partial g}{\partial r} + \frac{1}{r^2\cot\theta\prime}\frac{\partial g}{\partial \theta\prime} + \frac{1}{r^2\cos^2\theta\prime}\frac{\partial^2 g}{\partial \lambda^2} \tag{53}$$

$$\frac{\partial^2 g}{\partial y \partial z} = \frac{1}{r^2\cos\theta\prime}\frac{\partial g}{\partial \lambda} - \frac{1}{r\cos\theta\prime}\frac{\partial g}{\partial r \partial \lambda} \tag{54}$$

$$\frac{\partial^2 g}{\partial z^2} = \frac{\partial^2 g}{\partial r^2} \tag{55}$$

The above equations (Equation (47) to Equation (55)) are referred to as point P. Due to the non-zero partial derivative of g with respect to r, a parameterization of the isogravitational surface [21] for a small region near point P is

$$\bar{s} : (-\varepsilon, \varepsilon) \times (-\varepsilon, \varepsilon) \to \Re^3 : (x, y) \to \bar{s}(x, y) = (x, y, z(x, y)) \tag{56}$$

The function z is not known, but its partial derivatives can be determined. Hence, the coordinate vectors are

$$\frac{\partial \bar{s}}{\partial x} = \left(1, 0, \frac{\partial z}{\partial x}\right) = \left(1, 0, -\frac{\frac{\partial g}{\partial x}}{\frac{\partial g}{\partial z}}\right) \tag{57}$$

$$\frac{\partial \bar{s}}{\partial y} = \left(1, 0, \frac{\partial z}{\partial y}\right) = \left(1, 0, -\frac{\frac{\partial g}{\partial y}}{\frac{\partial g}{\partial z}}\right) \tag{58}$$

The normal vector is

$$\overline{N}_C = \frac{\partial \bar{s}}{\partial x} \times \frac{\partial \bar{s}}{\partial y} = \left(\frac{\frac{\partial g}{\partial x}}{\frac{\partial g}{\partial z}}, \frac{\frac{\partial g}{\partial y}}{\frac{\partial g}{\partial z}}, 1\right) \tag{59}$$

The unit normal vector is

$$\overline{N} = \left|\frac{\partial g}{\partial z}\right| \frac{1}{\|grad g\|} \overline{N}_C \tag{60}$$

Fundamental elements [21] of the first kind are

$$E = \frac{\left(\frac{\partial g}{\partial x}\right)^2 + \left(\frac{\partial g}{\partial z}\right)^2}{\left(\frac{\partial g}{\partial z}\right)^2} \tag{61}$$

$$F = \frac{\frac{\partial g}{\partial x}\frac{\partial g}{\partial y}}{\left(\frac{\partial g}{\partial z}\right)^2} \tag{62}$$

$$G = \frac{\left(\frac{\partial g}{\partial y}\right)^2 + \left(\frac{\partial g}{\partial z}\right)^2}{\left(\frac{\partial g}{\partial z}\right)^2} \tag{63}$$

Fundamental elements of the second kind are

$$L = \langle \overline{N}, \frac{\partial^2 \overline{s}}{\partial x^2} \rangle = \left| \frac{\partial g}{\partial z} \right| \frac{-\frac{\partial^2 g}{\partial x^2}\left(\frac{\partial g}{\partial z}\right)^2 + 2\frac{\partial^2 g}{\partial x \partial z}\frac{\partial g}{\partial x}\frac{\partial g}{\partial z} - \frac{\partial^2 g}{\partial z^2}\left(\frac{\partial g}{\partial x}\right)^2}{\left(\frac{\partial g}{\partial z}\right)^3 \|gradg\|} \tag{64}$$

$$M = \langle \overline{N}, \frac{\partial^2 \overline{s}}{\partial x \partial y} \rangle = \left| \frac{\partial g}{\partial z} \right| \frac{\frac{\partial^2 g}{\partial x \partial z}\frac{\partial g}{\partial y}\frac{\partial g}{\partial z} + \frac{\partial^2 g}{\partial y \partial z}\frac{\partial g}{\partial x}\frac{\partial g}{\partial z} - \frac{\partial^2 g}{\partial x \partial y}\left(\frac{\partial g}{\partial z}\right)^2 - \frac{\partial^2 g}{\partial z^2}\frac{\partial g}{\partial x}\frac{\partial g}{\partial y}}{\left(\frac{\partial g}{\partial z}\right)^3 \|gradg\|} \tag{65}$$

$$N = \langle \overline{N}, \frac{\partial^2 \overline{s}}{\partial y^2} \rangle = \left| \frac{\partial g}{\partial z} \right| \frac{-\frac{\partial^2 g}{\partial y^2}\left(\frac{\partial g}{\partial z}\right)^2 + 2\frac{\partial^2 g}{\partial y \partial z}\frac{\partial g}{\partial y}\frac{\partial g}{\partial z} - \frac{\partial^2 g}{\partial z^2}\left(\frac{\partial g}{\partial y}\right)^2}{\left(\frac{\partial g}{\partial z}\right)^3 \|gradg\|} \tag{66}$$

Let P be a point far away from the Earth, i.e., r_P takes a large value. As we slowly approach Earth along the polar axis $r = r_P$, g_x and g_y are small quantities. Thus, Equation (61) to Equation (66) become

$$E_P = 1 \tag{67}$$

$$F_P = 0 \tag{68}$$

$$G_P = 1 \tag{69}$$

$$L_P = \left. \frac{\frac{\partial^2 g}{\partial x^2}}{\|gradg\|} \right|_P = \left. -\frac{\frac{\partial^2 g}{\partial x^2}}{\frac{\partial g}{\partial z}} \right|_P = -\frac{1}{r_P} - \frac{1}{r_P^2}\left. \frac{\frac{\partial^2 g}{\partial \theta \prime^2}}{\frac{\partial g}{\partial r}} \right|_P \tag{70}$$

$$M_P = \left. \frac{\frac{\partial^2 g}{\partial x \partial y}}{\|gradg\|} \right|_P = \left. -\frac{\frac{\partial^2 g}{\partial x \partial y}}{\frac{\partial g}{\partial z}} \right|_P = -\frac{1}{r_P^2 \cos\theta_P \prime}\left. \frac{\frac{\partial^2 g}{\partial \theta \prime \partial \lambda}}{\frac{\partial g}{\partial r}} \right|_P + \frac{\sin\theta_P \prime}{r_P^2 \cos^2\theta_P \prime}\left. \frac{\frac{\partial g}{\partial \lambda}}{\frac{\partial g}{\partial r}} \right|_P \tag{71}$$

$$N_P = \left. \frac{\frac{\partial^2 g}{\partial y^2}}{\|gradg\|} \right|_P = \left. -\frac{\frac{\partial^2 g}{\partial y^2}}{\frac{\partial g}{\partial z}} \right|_P = -\frac{1}{r_P} + \frac{1}{r_P^2 \cot\theta_P \prime}\left. \frac{\frac{\partial g}{\partial \theta \prime}}{\frac{\partial g}{\partial r}} \right|_P + \frac{1}{r_P^2 \cos^2\theta_P \prime}\left. \frac{\frac{\partial^2 g}{\partial \lambda^2}}{\frac{\partial g}{\partial r}} \right|_P \tag{72}$$

At point P, the principal curvatures, k_1 and k_2, of the isogravitational surface [28] are the roots of the following equation:

$$\begin{aligned} &(EG - F^2)_P k - (EN - 2FM + GL)_P k + (LN - M^2)_P = 0 \Rightarrow \\ &\Rightarrow k^2 - (L + N)_P k + (LN - M^2)_P = 0 \end{aligned} \tag{73}$$

Hence,

$$k_1 = \left. \frac{L + N + \sqrt{(L - N)^2 - 4M^2}}{2} \right|_P \tag{74}$$

$$k_2 = \left. \frac{L + N - \sqrt{(L - N)^2 - 4M^2}}{2} \right|_P \tag{75}$$

The mean and Gaussian curvatures of the isogravitational surface at point P are equal to

$$H = \frac{L+N}{2}\Big|_P = \frac{1}{2}\left[-\frac{1}{r_P} - \frac{1}{r_P^2}\frac{\frac{\partial^2 g}{\partial \theta'^2}}{\frac{\partial g}{\partial r}}\Big|_P - \frac{1}{r_P} + \frac{1}{r_P^2 \cot\theta_{P'}}\frac{\frac{\partial g}{\partial \theta'}}{\frac{\partial g}{\partial r}}\Big|_P + \frac{1}{r_P^2 \cos^2\theta_{P'}}\frac{\frac{\partial^2 g}{\partial \lambda^2}}{\frac{\partial g}{\partial r}}\Big|_P\right] \quad (76)$$

$$K_G = \frac{LN-M^2}{EG-F^2}\Big|_P = \left(-\frac{1}{r_P} - \frac{1}{r_P^2}\frac{\frac{\partial^2 g}{\partial \theta'^2}}{\frac{\partial g}{\partial r}}\Big|_P\right)\left(-\frac{1}{r_P} + \frac{1}{r_P^2 \cot\theta_{P'}}\frac{\frac{\partial g}{\partial \theta'}}{\frac{\partial g}{\partial r}}\Big|_P + \frac{1}{r_P^2 \cos^2\theta_{P'}}\frac{\frac{\partial^2 g}{\partial \lambda^2}}{\frac{\partial g}{\partial r}}\Big|_P\right) - \left(-\frac{1}{r_P^2 \cos\theta_{P'}}\frac{\frac{\partial^2 g}{\partial \theta' \partial \lambda}}{\frac{\partial g}{\partial r}}\Big|_P + \frac{\sin\theta_{P'}}{r_P^2 \cos^2\theta_{P'}}\frac{\frac{\partial g}{\partial \lambda}}{\frac{\partial g}{\partial r}}\Big|_P\right)^2 \quad (77)$$

If point P is far away from the Earth, Equations (76) and (77) become

$$H = -\frac{1}{r_P} \quad (78)$$

$$K_G = \frac{1}{r_P^2} \quad (79)$$

Therefore, for large values of r, the isogravitational surfaces are spheres of radius r_P. Approaching the Earth along polar axis $r = r_P$, the mean curvature is given by Equation (76) and the Gauss curvature is derived from the following relation:

$$K_G^{r>} = \frac{1}{r_P^2} + \frac{1}{r_P^3}\left(-\frac{1}{\cot\theta_{P'}}\frac{\frac{\partial g}{\partial \theta'}}{\frac{\partial g}{\partial r}}\Big|_P - \frac{1}{\cos^2\theta_{P'}}\frac{\frac{\partial^2 g}{\partial \lambda^2}}{\frac{\partial g}{\partial r}}\Big|_P + \frac{\frac{\partial^2 g}{\partial \theta'^2}}{\frac{\partial g}{\partial r}}\Big|_P\right) \quad (80)$$

The Gauss curvature at point P characterizes the shape of the isogravitational surface around that point. The second term of Equation (80) describes the deviation of the isogravitational surface's shape from that of a sphere. As point P gradually approaches Earth (with a small δr), isogravitational surfaces form closed surfaces with a positive Gauss curvature. Around point P, the shape of the isogravitational surface is an elliptic paraboloid.

Figure 1 shows the shape of the isogravitational surface around point P which is a paraboloid. The shape of isogravitational surfaces becomes more complicated when point P is close to the Earth. Assuming the geoid represents the true shape of the Earth, let P be a point on the geoid, particularly at the North Pole ($\theta' = \pi/2$). The isogravitational surface passing through point P cannot intersect the geoid. Hence, both surfaces are tangent at point P, i.e., they have common tangent plane. Around point P, the z - coordinates of the neighboring points have a positive sign.

Let γ be a curve on the isogravitational surface that is formed by the intersection of the isogravitational surface and a meridian plane. Since isogravitational surfaces are closed surfaces, this means that as we move further from point P along the curve γ, we reach a point A where the Gauss curvature is positive and the neighboring points of the isogravitational surface have negative z-coordinates. Consequently, on the curve γ between point P and A, there is a point, Q, where the Gauss curvature is equal to zero and the z-coordinates of the neighboring points are also equal to zero. This means that point Q is a singular point, i.e., $L_Q = M_Q = N_Q = 0$.

The isogravitational surface that passes through point P intersects the equatorial plane and lies outside the geoid, without any singular points. The set of points that have a zero Gauss curvature is defined as a singular ring of the isogravitational surface. Close to the surface of the geoid, each isogravitational surface has at least one singular ring.

Figure 2 shows a simplified representation of a singular ring. As point P moves along the z-axis, r_P increases. We have demonstrated that as the polar distance r reaches high values, the isogravitational surfaces become spheres. This indicates that the Gauss curvature at point P approaches zero. Simultaneously, as the singular rings slowly shrink,

they reach a critical point when the Gauss curvature at point P becomes zero. Point P represents the location of the singular ring on this surface. If r_P surpasses the critical value, the isogravitational surfaces will exhibit a positive Gauss curvature throughout, resulting in the absence of singular rings. Finally, as r_P takes on large values, isogravitational surfaces become spheres with a radius of r_P.

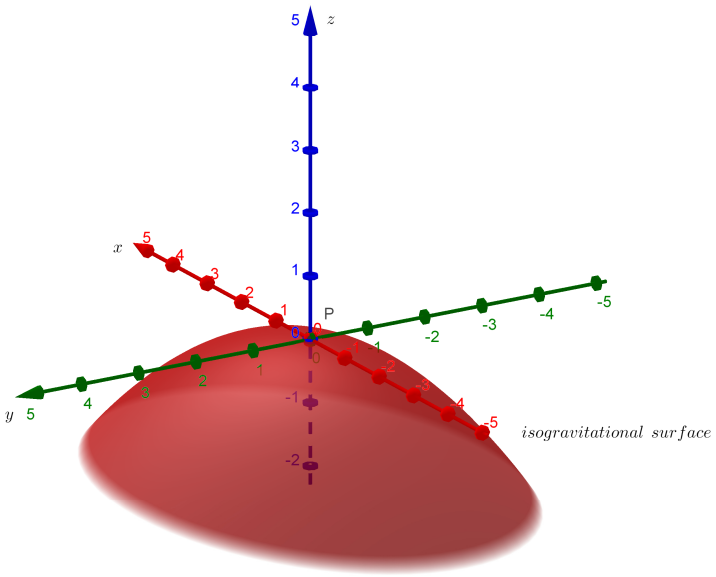


Figure 1. Isogravitational Surface.

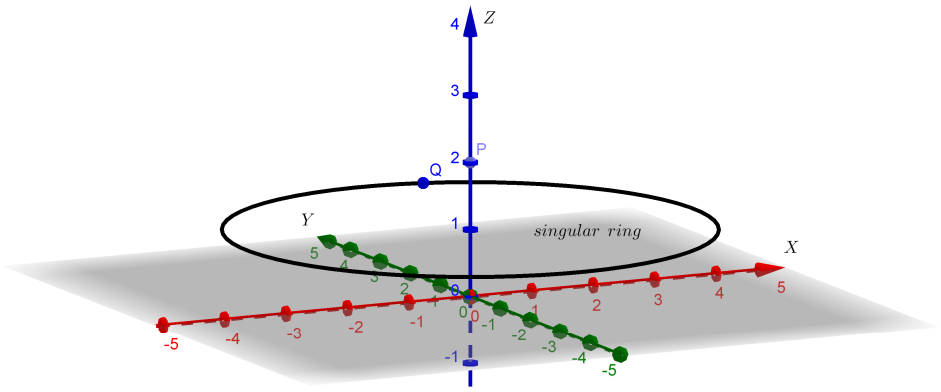


Figure 2. Singular Ring.

Figure 3 shows a simplified image of a family of singular rings.

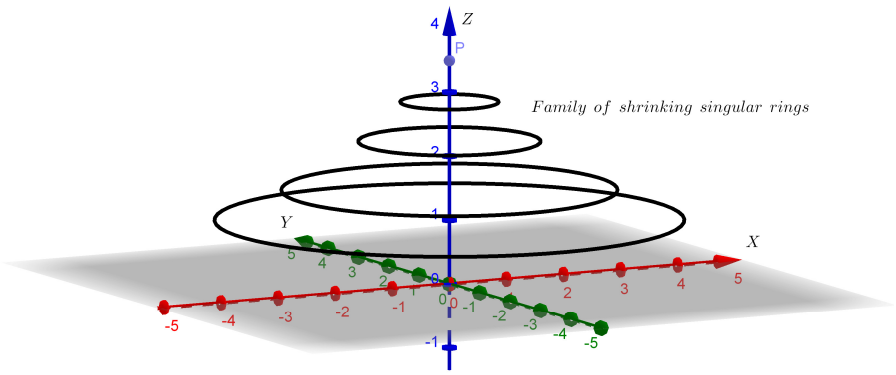


Figure 3. Family of Singular Rings.

5. Disturbing Isogravitational Surfaces, Anomalistic Isogravitational Surfaces, and a Comparison between the Two Methods of Determining Disturbing Potential

A disturbing isogravitational surface is characterized by a constant value of gravity disturbance δg. Similarly, an anomalistic isogravitational surface is characterized by a constant value of gravity anomaly Δg. This local study is similar to that of the isogravitational surfaces, where the mean and Gaussian curvatures are given by Equations (76) and (77), respectively, with g replaced by δg (or Δg).

In the table below, we present a summary of the findings from Sections 3 and 4. The first column lists the gravitational quantities, while the second column indicates whether each quantity is expressed as a series of spherical harmonics provided by the Laplace equation. We use "yes" to indicate when the quantity is explicitly expressed as such and "no" when it is not. In the third column, "yes" is selected when the quantity is represented as a series of spherical harmonics obtained from the G-modified Helmholtz equation. The term "direct" is used to describe a quantity that satisfies the chosen partial differential equation and remains either measurable or constant on the boundary. Conversely, we utilize the term "indirect" when the conditions mentioned above are not met. In addition, the symbols "Ve" and "V_E" represent the Earth's and ellipsoid's Newtonian gravitational potential, respectively.

From Table 1, we can observe that the Laplace equation offers an advantage in the analysis of gravitational potential, whereas the G-modified Helmholtz equation has an edge in the study of gravity intensity, gravity disturbance, and gravity anomaly. It is worth noting that the process of determining the disturbing potential using the G-modified Helmholtz equation is quite straightforward, as it only involves solving a simple partial differential equation (see Equation (24)).

Table 1. Comparison between Laplace and G – modified Helmholtz equation.

Quantity	Laplace Equation Spherical Harmonics, Radial Distance in Rational Powers	G-Modified Helmholtz Equation, Spherical Harmonics, Radial Distance in Irrational Powers
Disturbing Potential T	Yes, indirect determination	Yes, indirect determination
Ellipsoid's gravity potential V_e	Yes, direct determination *	No
Earth's gravity potential V_E	Yes, indirect determination	No
Gravity disturbance δg	Yes, indirect determination	Yes, direct determination
Gravity anomaly Δg	Yes, indirect determination	Yes, direct determination
Gravity intensity g	No	Yes, direct determination
Ellipsoid's gravity intensity γ	No	Yes, direct determination
Vertical gradient of g	No	Yes, indirect determination
Vertical gradient of γ	No	Yes, indirect determination
Component ξ	Yes, indirect determination	Yes, indirect determination
Component η	Yes, indirect determination	Yes, indirect determination
Geoid undulation N	Yes, indirect determination	Yes, indirect determination

* For details, we refer to [22].

6. Summation of the Results

In this section, we tabulate the new results that were presented in Section 3.

Table 2 represents the significance of the G–modified Helmholtz equation, which enables us to study gravitational quantities independently of gravity potential and express them as a new kind of series of spherical harmonics. We chose to express the most significant formulae in detail, which are related to gravity g, gravity disturbance δg, gravity anomaly Δg, disturbing potential T, geoid undulation N, vertical gradient of gravity, and normal

gravity on the geoid (marked in bold). Moreover, Table 2 also includes the components of the deflection of the vertical.

Table 2. New spherical harmonic expressions occured from G—modified Helmholtz equation.

Quantity	Expression
Gravity g	$g(r,\theta\prime,\lambda)=\sum_{n=0}^{+\infty}\sum_{m=0}^{n}\left(\frac{R}{r}\right)^{\frac{1+\sqrt{9+4n(n+1)}}{2}}[A_{nm}^{g}P_{nm}(\sin\theta\prime)\cos m\lambda+B_{nm}^{g}P_{nm}(\sin\theta\prime)\sin m\lambda]$
Gravity disturbance δg	$\delta g(r,\theta\prime,\lambda)=\sum_{n=0}^{+\infty}\sum_{m=0}^{n}\left(\frac{R}{r}\right)^{\frac{1+\sqrt{9+4n(n+1)}}{2}}[A_{nm}^{\delta g}P_{nm}(\sin\theta\prime)\cos m\lambda+B_{nm}^{\delta g}P_{nm}(\sin\theta\prime)\sin m\lambda]$
Gravity anomaly Δg	$\Delta g(r,\theta\prime,\lambda)=\sum_{n=0}^{+\infty}\sum_{m=0}^{n}\left(\frac{R}{r}\right)^{\frac{1+\sqrt{9+4n(n+1)}}{2}}[A_{nm}^{\Delta g}P_{nm}(\sin\theta\prime)\cos m\lambda+B_{nm}^{\Delta g}P_{nm}(\sin\theta\prime)\sin m\lambda]$
Disturbing potential T in three-dimensional space	$T(r,\theta\prime,\lambda)=\sum_{n=0}^{+\infty}\frac{2R}{\sqrt{9+4n(n+1)}-1}\left(\frac{R}{r}\right)^{\frac{\sqrt{9+4n(n+1)}-1}{2}}\{\left\{\sum_{m=0}^{n}[A_{nm}^{\delta g}P_{nm}(\sin\theta\prime)\cos m\lambda+B_{nm}^{\delta g}P_{nm}(\sin\theta\prime)\sin m\lambda]\right\}+\sum_{n=0}^{+\infty}\sum_{m=0}^{n}\frac{R}{2}\left[\left(\frac{\sqrt{9+4n(n+1)}-5}{\sqrt{9+4n(n+1)}-1}A_{nm}^{\delta g}-A_{nm}^{\Delta g}\right)P_{nm}(\sin\theta\prime)\cos m\lambda+\left(\frac{\sqrt{9+4n(n+1)}-5}{\sqrt{9+4n(n+1)}-1}B_{nm}^{\delta g}-B_{nm}^{\Delta g}\right)P_{nm}(\sin\theta\prime)\sin m\lambda\right]$
Disturbing potential on the geoid	$T(\theta\prime,\lambda)=\frac{R}{2}[A_{nm}^{\delta g}P_{nm}(\sin\theta\prime)\cos m\lambda+B_{nm}^{\delta g}P_{nm}(\sin\theta\prime)\sin m\lambda]-\left.-[A_{nm}^{\Delta g}P_{nm}(\sin\theta\prime)\cos m\lambda+B_{nm}^{\Delta g}P_{nm}(\sin\theta\prime)\sin m\lambda]\right\}=$ $=\frac{R}{2}[\delta g(\theta\prime,\lambda)-\Delta g(\theta\prime,\lambda)]\quad,\quad r=R$
Geoid undulation N	$N(\theta\prime,\lambda)=\frac{R}{2\gamma_{ellispoid}(\theta\prime,\lambda)}[\delta g(\theta\prime,\lambda)-\Delta g(\theta\prime,\lambda)]$ or $N(r,\theta\prime,\lambda)=\frac{R}{2}\left(\frac{\gamma_{ellepsoid}(\theta\prime,\lambda)-\gamma_{geoid}(r,\theta\prime,\lambda)}{\gamma_{ellepsoid}(\theta\prime,\lambda)}\right)$
Vertical gradient of gravity	$\frac{\partial g}{\partial r}=\sum_{n=0}^{+\infty}\sum_{m=0}^{n}-\frac{1+\sqrt{9+4n(n+1)}}{2r}\left(\frac{R}{r}\right)^{\frac{1+\sqrt{9+4n(n+1)}}{2}}[A_{nm}^{g}P_{nm}(\sin\theta\prime)\cos m\lambda+B_{nm}^{g}P_{nm}(\sin\theta\prime)\sin m\lambda]\quad,\quad r>R$
Component ξ	$\xi(r,\theta\prime,\lambda)=-\frac{1}{r\gamma(r,\theta\prime,\lambda)}\frac{\partial T}{\partial\theta\prime}$
Component η	$\eta(r,\theta\prime,\lambda)=-\frac{1}{r\gamma(r,\theta\prime,\lambda)\sin\theta\prime}\frac{\partial T}{\partial\lambda}$
Normal gravity on the geoid	$\gamma_{geoid}(r,\theta\prime,\lambda)=\sum_{n=0}^{n_0}a_{2n}r^{-\frac{1+\sqrt{9+4n(n+1)}}{2}}\cdot\cdot\{P_{2n}(\sin\theta\prime)+e^2(2n+1)\sin\theta\prime[P_{2n}(\sin\theta\prime)\sin\theta\prime-P_{2n+1}(\sin\theta\prime)]\}$

Along with the Laplace equation, the G-modified Helmholtz equation can also be classified as a fundamental partial differential equation of Physical geodesy.

7. Conclusions

In this work, we presented new alternative formulae for various quantities related to the Earth's gravity field.

The G-modified Helmholtz equation is a partial differential equation that allows us to calculate the intensity of gravity as a series of spherical harmonics. The intensity of gravity plays a crucial role in a wide range of applications. The harmonic series of gravity intensity is a series in which the powers of the radial distance are irrational numbers. This series allows us to represent the vertical gradient of gravity as a series of spherical harmonics. The vertical gradient of gravity is very important for numerous geophysical applications.

Gravity disturbance and gravity anomaly are very significant. They are expressed as a series of spherical harmonics, as provided by the Laplace equation or the G-modified Helmholtz equation. In the first scenario, their determination is not straightforward, as we

have to solve either a Neumann (Hotine problem) or Robin (Stokes problem) boundary value problem. In the second case, their determinations occur as solutions to Dirichlet boundary value problems, which are simpler than the previous two cases.

The G-modified Helmholtz equation offers an additional advantage: it enables us to calculate the Earth's disturbing potential using a simple partial differential equation. The solution of this equation involves only the integration of the harmonic series of gravity anomaly with respect to radial distance, r. Moreover, the G-modified Helmholtz equation can also be used to find the components ξ and η of the deflection of the vertical and the geoid undulation.

In addition, on the surface of the geoid, the disturbing potential and the geoid undulation take the elegant form of a series of spherical harmonics, which shows the dependence on the difference between gravity disturbance and gravity anomaly on this surface. An interesting theoretical result that occurs is that both disturbing potential and geoid undulation can be determined if the value of normal gravity is known on the surface of the geoid. All the new formulae that occur from the G-modified Helmholtz equation are tabulated in Section 6.

The harmonic series of gravitational intensity gives us a significant advantage in studying the Earth's isogravitational surfaces. The Earth's gravitational potential's first- and second-order partial derivatives can be conveniently expressed in Cartesian systems. This simplifies the expressions of the partial derivatives and the required approximations. It is shown that isogravitational surfaces are closed surfaces with a non-negative Gauss curvature.

At very high altitudes, when the value of the radial distance "r" is large, isogravitational surfaces are spheres with a radius equal to the selected radial distance. As we move closer to Earth, their appearance becomes progressively more intricate. Each isogravitational surface near Earth has at least one set of points defined as a singular ring. This name is attributed to the fact that all points belonging to this set are singular points. The family of singular rings converges to a single point located on Earth's mean axis of rotation.

Future work could involve the development of new Global Gravity Models for gravity intensity g, gravity disturbance δg, and gravity anomaly Δg. A Global Gravity Model for gravity intensity would represent a significant advancement with a wide range of applications, such as in the fields of geophysics, archaeology, geodesy, orbit determination, and inertial navigation. Furthermore, it would be advantageous to develop new Global Gravity Models for gravity anomaly and gravity disturbance, separate from Global Geopotential Models. This evolution may present us with opportunities to improve present techniques for manipulating gravity data, leading to a deeper comprehension of the Earth's gravity field.

Funding: This research received no external funding.

Institutional Review Board Statement: The study did not require ethical approval.

Informed Consent Statement: Not applicable.

Data Availability Statement: No data was used.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Jarosz, W.; Carr, N.A.; Jensen, H.W. Importance Sampling Spherical Harmonics. *Comput. Graph. Forum* **2009**, *28*, 577–586. Available online: http://graphics.ucsd.edu/~henrik/papers/importance_sampling_spherical_harmonics.pdf (accessed on 2 February 2024). [CrossRef]
2. Silvennoinen, A.; Sloan, P.P. Moving Basis Decomposition for Precomputed Light Transport. *Comput. Graph. Forum* **2021**, *40*, 127–137. Available online: https://www.ppsloan.org/publications/mbd.pdf (accessed on 2 February 2024). [CrossRef]
3. Wang, J.; Xu, K.; Zhou, K.; Lin, S.; Hu, S.; Guo, B. Spherical Harmonics Scaling. *Vis. Comput.* **2006**, *22*, 713–720. Available online: https://link.springer.com/article/10.1007/s00371-006-0057-8 (accessed on 2 February 2024). [CrossRef]

4. Annen, T.; Kautz, J.; Durand, F.; Seidel, H.P. Spherical Harmonic Gradients for Mid-Range Illumination. In Proceedings of the 15th Eurographics Workshop in Rendering Techniques, Norrköping, Sweden, 21–24 June 2004; pp. 331–336. Available online: https://diglib.eg.org/bitstream/handle/10.2312/EGWR.EGSR04.331-336/331-336.pdf?sequence=1&isAllowed=y (accessed on 2 February 2024).
5. Sanso, F.; Sideris, M. *Geoid Determination Theory and Methods*; Springer Science & Business Media: Berlin, Germany, 2013; Chapter 6; Volume 262, p. 131. Available online: https://link.springer.com/book/10.1007/978-3-540-74700-0 (accessed on 2 February 2024).
6. Shouny, A.; Khalil, R.; Kamel, A.; Miky, Y. Assessments of recent Global Geopotential Models based on GPS levelling and gravity data along coastal zones of Egypt. *Open Gosciences* **2023**, *15*, 20220450. Available online: https://www.degruyter.com/document/doi/10.1515/geo-2022-0450/html (accessed on 2 February 2024). [CrossRef]
7. Wu, Y.; He, X.; Luo, Z.; Shi, H. An Assessment of Recently Released High Degree Global Geopotential Models Based on Heterogeneous Geodetic and Ocean Data. *Front. Earth Sci.* **2021**, *9*, 749611. Available online: https://www.frontiersin.org/articles/10.3389/feart.2021.749611/full (accessed on 2 February 2024). [CrossRef]
8. Isik, M.S.; Çevikalp, M.R.; Erol, B.; Erol, S. Improvement of GOCE-Based Geopotential Models for Gravimetric Geoid Modeling in Turkey. *Geosciences* **2022**, *12*, 432. Available online: https://www.mdpi.com/2076-3263/12/12/432 (accessed on 2 February 2024). [CrossRef]
9. Elizondo, N.A.; Lopez, R.V.G.; Carillo, X.G.T.; Becerra, G.E.V. Combining Global Geopotential Models, Digital Elevation Models, and GNSS/Leveling for Precise Local Geoid Determination in Some Mexico Urban Areas: Case Study. *ISPRS Int. J. Geo-Inf.* **2021**, *10*, 819. Available online: https://www.mdpi.com/2220-9964/10/12/819 (accessed on 2 February 2024). [CrossRef]
10. Roithmayr, C.M. *Contributions of Spherical Harmonics to Magnetic and Gravitational Fields*; NASA TM-2004-213007; National Aeronautics and Space Administration: Washington, DC, USA, 2004. Available online: https://ntrs.nasa.gov/api/citations/20040047194/downloads/20040047194.pdf (accessed on 2 February 2024).
11. Campbell, W.H. *Geomagnetism Applications*; U.S. Geological Survey Sircular 1109; United States Government Printing Office: Washington, DC, USA, 1995. Available online: https://pubs.usgs.gov/circ/1995/1109/report.pdf (accessed on 2 February 2024).
12. Prada, K.O.; Antelo, M.J.; Murillo, N.A.; Cózar, J.R.; Contreras–de–Villar, F.; Pérez, J.J.M. A New Method for the Collection of Marine Geomagnetic Information: Survey Application in the Colombian Caribbean. *J. Mar. Sci. Eng.* **2021**, *9*, 10. Available online: https://www.mdpi.com/2077-1312/9/1/10 (accessed on 2 February 2024). [CrossRef]
13. Zigan, L. Overview of Electric Field Applications in Energy and Process Engineering. *Energies* **2018**, *11*, 1361. Available online: https://www.mdpi.com/1996-1073/11/6/1361 (accessed on 15 February 2024). [CrossRef]
14. Tognolatti, L.; Ponti, C.; Santarsiero, M.; Schettini, G. An Efficient Computational Technique for the Electromagnetic Scattering by Prolate Spheroids. *Mathematics* **2022**, *10*, 1761. Available online: https://www.mdpi.com/2227-7390/10/10/1761 (accessed on 15 February 2024). [CrossRef]
15. Vafeas, P.; Perrusson, G.; Lesselier, D. Low-frequency scattering from perfectly conducting spheroidal bodies in a conductive medium with magnetic dipole excitation. *Int. J. Eng. Sci.* **2009**, *47*, 372–390. Available online: https://www.researchgate.net/publication/222923079_Lowfrequency_scattering_fro_perfectly_conducting_spheroidal_bodies_in_a_conductive_medium_with_magnetic_dipole_excitation (accessed on 15 February 2024). [CrossRef]
16. Perrusson, G.; Vafeas, P.; Chatjigeorgiou, I.K.; Lesselier, D. Low-frequency on-site identification of a highly conductive body buried in Earth from a model ellipsoid. *IMA J. Appl. Math.* **2015**, *80*, 963–980. Available online: https://www.researchgate.net/publication/274162167_Low-frequency_onsite_identification_of_a_highly_conductive_body_buried_in_Earth_from_a_model_ellipsoid (accessed on 15 February 2024). [CrossRef]
17. Ming, Y.; Niu, X.; Xie, X.; Han, Z.; Li, Q.; Sun, S. Application on Gravity Exploration in Urban Active Fault Detection. In The 9th International Conference on Environmental and Engineering Geophysics 2021, Proceedings of the IOP Conference Series: Earth and Environmental Sciences, Changchun, China, 11–14 October 2020; IOP Publishing: Bristol, UK, 2021; Volume 660, p. 012057. Available online: https://www.researchgate.net/publication/349532168_Application_of_gravity_exploration_in_urban_active_fault_detection (accessed on 21 February 2024).
18. Sjöberg, L.E.; Bagherbagdi, M. Gravity Inversion and Integration. In *Theory and Applications in Geodesy and Geophysics*; Springer: Basel, Switzerland, 2017; p. 251. Available online: https://www.geokniga.org/bookfiles/geokniga-gravity-inversion-and-integration.pdf (accessed on 21 February 2024).
19. Scheinert, M.; Ferranccioli, F.; Schwabe, J.; Studinger, M.; Damaske, D.; Jokat, W.; Aleshkova, N.; Jordan, T.; Leitchenkov, G.; Blankenship, D.D.; et al. New Antarctic gravity gnomaly grid for enhanced geodetic and geophysical studies in Antarctica. *Geophys. Res. Lett.* **2016**, *43*, 600–610. Available online: https://www.researchgate.net/publication/289983658_New_Antarctic_Gravity_Anomaly_Grid_for_Enhanced_Geodetic_and_Geophysical_Studies_in_Antarctica (accessed on 21 February 2024). [CrossRef] [PubMed]
20. Ismail, N.; Yanis, M.; Abdullah, F.; Irfansyam, A.; Atmojo, B.S.W. Mapping buried ancient structure using gravity method: A case study from Cot Sidi Abdullah North Aceh. *J. Phys.* **2018**, *1120*, 012035. Available online: https://iopscience.iop.org/article/10.1088/1742-6596/1120/1/012035/pdf (accessed on 21 February 2024). [CrossRef]
21. Manoussakis, G. The Gravity Force Generated by a Non-Rotating Level Ellipsoid of Revolution with Low Eccentricity as a Series of Spherical Harmonics. *Mathematics* **2023**, *11*, 1974. Available online: https://www.mdpi.com/journal/mathematics/special_issues/Applications_Partial_Differential_Equations_Mathematical_Physics (accessed on 21 February 2024). [CrossRef]

22. Heiskanen, W.; Moritz, H. Physical Geodesy. *W.H. Freeman Co.* **1967**, 82–112. Available online: https://archive.org/details/heiskanen_morits_1967_physical_geodesy/page/n89/mode/2up?view=theater (accessed on 3 March 2024).
23. Bagherbandi, M.; Jouybari, A.; Niflourousan, F.; Ägren, J. Deflection of vertical effect on direct georeferencing in aerial mobile mapping systems: A case study in Sweden. *Photogramm. Rec.* **2022**, *37*, 285–305. Available online: https://www.diva-portal.org/smash/get/diva2:1684428/FULLTEXT01.pdf (accessed on 3 March 2024). [CrossRef]
24. Barzaghi, R.; Carrion, D.; Pepe, M.; Prezioso, G. Computing the Deflection of the Vertical for Improving Aerial Surveys: A Comparison between EGM2008 and ITALGEO05 Estimates. *Sensors* **2016**, *16*, 1168. Available online: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5017334/pdf/sensors-16-01168.pdf (accessed on 3 March 2024). [CrossRef] [PubMed]
25. Šprlák, M.; Novak, P. Integral formulas for computing a third order gravitational tensor from volumetric mass density, disturbing gravitational potential, gravity anomaly and gravity disturbance. *J. Geod.* **2015**, *89*, 141–157. Available online: https://www.researchgate.net/publication/271211475_Integral_formulas_for_computing_a_thirdorder_gravitational_tensor_from_volumetric_mass_density_disturbing_gravitational_potential_gravity_anomaly_and_gravity_disturbance (accessed on 5 March 2024). [CrossRef]
26. Gorshkov, G.P. Isogravitational surfaces. *Geophys. Abstr.* **1934**, *82*. Available online: https://books.google.gr/books?id=1RxUCPOWOtEC&pg=PA1738&lpg=PA1738&dq=isogravitational+surfaces+of+the+Earth&source=bl&ots=9VnOhvB6Cs&sig=ACfU3U2hlSJxufGjWTktf6yfI13mwKuylA&hl=el&sa=X&ved=2ahUKEwjx8pD85tqEAxVh9LsIHbuUA_E4ChDoAXoECAYQAw#v=onepage&q=isogravitational%20surfaces%20of%20the%20Earth&f=false (accessed on 5 March 2024). [CrossRef]
27. Koop, R. *Global Gravity Field Modelling Using Satellite Gravity Gradiometry*; Nederlandse Commissie voor Geodesie: Delft, Netherlands, 1993; Volume 38, pp. 32–183. Available online: https://ncgeo.nl/index.php/en/publicatiesgb/publications-on-geodesy/item/2544-pog-38-r-koop-global-gravity-modelling-using-satellite-gravity-gradiometry (accessed on 5 March 2024).
28. Giblin, P. *Differential Geometry of Curves and Surfaces*; Department of Mathematical Sciences, University of Liverpool: Liverpool, UK, 1999; pp. 32–39. Available online: https://www.liverpool.ac.uk/~pjgiblin/papers/diff-geom-book.pdf (accessed on 5 March 2024).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

AppliedMath

MDPI

Article

Three-Body 3D-Kepler Electromagnetic Problem—Existence of Periodic Solutions

Vasil Georgiev Angelov

Department of Mathematics and Informatics, University of Mining and Geology, "St. I. Rilski", 1700 Sofia, Bulgaria; angelov@mgu.bg

Abstract: The main purpose of the present paper is to prove the existence of periodic solutions of the three-body problem in the 3D Kepler formulation. We have solved the same problem in the case when the three particles are considered in an external inertial system. We start with the three-body equations of motion, which are a subset of the equations of motion (previously derived by us) for any number of bodies. In the Minkowski space, there are 12 equations of motion. It is proved that three of them are consequences of the other nine, so their number becomes nine, as much as the unknown trajectories are. The Kepler formulation assumes that one particle (the nucleus) is placed at the coordinate origin. The motion of the other two particles is described by a neutral system with respect to the unknown velocities. The state-dependent delays arise as a consequence of the finite vacuum speed of light. We obtain the equations of motion in spherical coordinates and split them into two groups. In the first group all arguments of the unknown functions are delays. We take their solutions as initial functions. Then, the equations of motion for the remaining two particles must be solved to the right of the initial point. To prove the existence–uniqueness of a periodic solution, we choose a space consisting of periodic infinitely smooth functions satisfying some supplementary conditions. Then, we use a suitable operator which acts on these spaces and whose fixed points are periodic solutions. We apply the fixed point theorem for the operators acting on the spaces of periodic functions. In this manner, we show the stability of the He atom in the frame of classical electrodynamics. In a previous paper of ours, we proved the existence of spin functions for plane motion. Thus, we confirm the Bohr and Sommerfeld's hypothesis for the He atom.

Keywords: classical electrodynamics; three-body problem; periodic solutions; fixed point method; neutral system

MSC: 78A35; 34K40; 47H10

check for updates

Citation: Angelov, V.G. Three-Body 3D-Kepler Electromagnetic Problem—Existence of Periodic Solutions. *AppliedMath* **2024**, *4*, 612–640. https://doi.org/10.3390/appliedmath4020034

Academic Editor: Claude Chaudet

Received: 25 March 2024
Revised: 19 April 2024
Accepted: 30 April 2024
Published: 16 May 2024

Copyright: © 2024 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The primary goal of the present paper is to investigate the three-body problem of classical electrodynamics in the 3D Kepler formulation. In [1], we formulated the same three-body problem externally for the particle inertial system, and in [2], we proved the existence–uniqueness of a periodic solution. The three-body problem is a particular case of the general formulation of the N-body problem (cf. [3]), where the existence–uniqueness of escape trajectories of N charged particles has been proved. Here, we add radiation terms for each of the three particles. We use the relativistically invariant form of the Dirac radiation term introduced in [4]. We proceed from the system from [1] and passing to the spherical coordinates, we formulate the 3D Kepler problem for three charged particles. The results obtained allow us to extend qualitative investigations concerning He atom and He-like atoms such as Li^+, and others. The method for solving the equations of motion is based on the choice of suitable spaces of functions consisting of an infinitely smooth periodic satisfying some additional conditions and introducing operators acting on these spaces. Their fixed points are periodic solutions of the equations of motion. It is not superfluous

to note that, using the 3D Kepler formulation, we put the nuclei at the origin and look for trajectories of two moving particles circling around the nuclei. From a mathematical point of view, it is a formal operation, but from a physical point of view, this means that we pass from the relativistic case to the quasi-classical one, because the Kepler problem is in fact a classical (Newton) one. This is also confirmed by the type of radiation term, which turns out to be in the Lorentz form.

This paper consists of seven sections, five appendices and conclusions. Section 1 is the introduction. Section 2 contains the 3D Kepler formulation for the equations of motion in Cartesian coordinates. This leads to the separation of the system into two groups. One of them consists of three equations containing unknown functions on the initial interval. We call them initial equations. We follow R.D. Driver's approach (cf. [5]) to describe the retarded interaction between particles, which is based on the idea of W. Pauli [6] and J.L. Synge [7]. The state-dependent delays are obtained. In fact, with respect to the unknown velocities, the system obtained is a neutral type, and the delays depend on the unknown trajectories. In Section 3, we derive the basic equations of motion in spherical coordinates. The equations obtained are again of a neutral type, with delays depending on the unknown trajectories. In Section 4, we introduce the function spaces consisting of infinitely differentiable periodic functions which satisfy some additional conditions. Some auxiliary propositions are also proved. In Section 5, we introduce operators whose fixed points are periodic solutions to the equations of motion and prove the Main Lemma. Section 6 contains the main result, which guarantees the existence–uniqueness of the periodic solution to equations of motion for the three-body problem in the 3D Kepler form. The proof is based on a fixed point theorem [8]. A numerical example is given. In Section 7, the Bohr and Sommerfeld hypotheses are discussed.

We mention some results in which different approaches are applied [9–18].

We recall some denotations and results from [1,2] concerning the three-body problem in Cartesian coordinates. We consider the system of 12 equations of motion using radiation terms for the three-body system [1]:

$$\begin{aligned} m_1 \frac{d\lambda_r^{(1)}}{ds_1} &= \frac{e_1}{c^2}\left(F_{rl}^{(12)}\lambda_l^{(1)} + F_{rl}^{(13)}\lambda_l^{(1)} + F_{rl}^{(1)rad}\lambda_l^{(1)}\right), \\ m_2 \frac{d\lambda_r^{(2)}}{ds_2} &= \frac{e_2}{c^2}\left(F_{rl}^{(21)}\lambda_l^{(2)} + F_{rl}^{(23)}\lambda_l^{(2)} + F_{rl}^{(2)rad}\lambda_l^{(2)}\right), \\ m_3 \frac{d\lambda_r^{(3)}}{ds_3} &= \frac{e_3}{c^2}\left(F_{rl}^{(31)}\lambda_l^{(3)} + F_{rl}^{(32)}\lambda_l^{(3)} + F_{rl}^{(3)rad}\lambda_l^{(3)}\right), \end{aligned} \tag{1}$$

where $r = 1,2,3,4$; c is the vacuum speed of light; m_k are the masses; and $e_k (k = 1,2,3)$ are the charges of the particles. Recall that there is a summation in repeating l in the right-hand sides of the above system. The elements of the electromagnetic tensors $F_{rl}^{(kn)} = \frac{\partial \Phi_l^{(n)}}{\partial x_r^{(k)}} - \frac{\partial \Phi_r^{(n)}}{\partial x_l^{(k)}}$ can be calculated by the retarded Lienard–Wiechert potentials $\Phi_r^{(n)} = -\frac{e_n \lambda_r^{(n)}}{\left\langle \lambda^{(n)}, \xi^{(kn)} \right\rangle_4}$, while the radiation terms can be calculated as a half of the difference between the retarded and advance potentials, in accordance with the Dirac assumption:

$$F_{mn}^{(k)rad} = \frac{1}{2}\left[\left(\frac{\partial A_n^{(k)ret}}{\partial x_m^{(k)ret}} - \frac{\partial A_m^{(n)ret}}{\partial x_n^{(k)ret}}\right) - \left(\frac{\partial A_n^{(k)adv}}{\partial x_m^{(k)adv}} - \frac{\partial A_m^{(n)adv}}{\partial x_n^{(k)adv}}\right)\right],$$

where $A_n^{(k)ret} = -\frac{e_k \lambda_n^{(k)ret}}{\left\langle \lambda^{(k)ret}, \xi^{(k)ret} \right\rangle_4}$, $A_n^{(k)adv} = -\frac{e_k \lambda_n^{(k)adv}}{\left\langle \lambda^{(k)adv}, \xi^{(k)adv} \right\rangle_4}$.

Here, $\left(x_1^{(k)}(t), x_2^{(k)}(t), x_3^{(k)}(t), x_4^{(k)}(t) = ict\right)$, $(k = 1,2,3)$ are space–time coordinates of the charged particles. The dot product in the Minkowski space is $\langle a, b \rangle_4 = a_n b_n = \sum_{n=1}^{4} a_n b_n$,

while in the three-dimensional subspace, $\langle a,b\rangle = a_\alpha b_\alpha = \sum\limits_{\alpha=1}^{3} a_\alpha b_\alpha$. The elements of proper times are ds_k $(k=1,2,3)$ and the unit tangent vectors to the world lines are

$$\lambda^{(k)} = \left(\lambda_1^{(k)},\lambda_2^{(k)},\lambda_3^{(k)},\lambda_4^{(k)}\right) = \left(\frac{u_1^{(k)}(t)}{\Delta_k},\frac{u_2^{(k)}(t)}{\Delta_k},\frac{u_3^{(k)}(t)}{\Delta_k},\frac{ic}{\Delta_k}\right),\ \overrightarrow{u}^{(k)} = \left(u_1^{(k)}(t),u_2^{(k)}(t),u_3^{(k)}(t)\right),$$
$$\Delta_k = \sqrt{c^2 - \left\langle \overrightarrow{u}^{(k)},\overrightarrow{u}^{(k)}\right\rangle}.$$

We have proved that three equations from (1) are a consequence of the rest [1].

The equations of motion become nine in number, as do the unknown functions $\left(u_1{}^{(k)}(t),u_2{}^{(k)}(t),\ u_3{}^{(k)}(t)\right),\ (k=1,2,3)$.

The isotropic vectors $\xi^{(km)}(k=1,2,3\,;\ m\neq k)$ are obtained in the following way: fix any event on the world line of the k-th particle and draw the light cone into the past. This cone intersects the world line of the m-th particle in any other (past) event. Then,

$$\xi^{(km)} = \left(\xi_1^{(km)},\xi_2^{(km)},\xi_3^{(km)},\xi_4^{(km)}\right) = \left(\overrightarrow{\xi}^{(km)},\xi_4^{(km)}\right) =$$
$$= \left(x_1^{(k)}(t) - x_1^{(m)}(t-\tau_{km}(t)), x_2^{(k)}(t) - x_2^{(m)}(t-\tau_{km}(t)), x_3^{(k)}(t) - x_3^{(m)}(t-\tau_{km}(t)), ic\tau_{km}(t)\right).$$

$\xi^{(km)}$ are isotropic four-vectors $\left\langle \xi^{(km)},\xi^{(km)}\right\rangle_4 = 0$, which implies that the retarded functions $\tau_{km}(t)$ should satisfy the following functional equations:

$$\tau_{km}(t) = \frac{1}{c}\sqrt{\left\langle \overrightarrow{\xi}^{(km)},\overrightarrow{\xi}^{(km)}\right\rangle} \equiv \frac{1}{c}\sqrt{\sum_{\alpha=1}^{3}\left[x_\alpha^{(k)}(t) - x_\alpha^{(m)}(t-\tau_{km}(t))\right]^2}.$$

The above equations are six in number because $(km) = (12),(13),(23),(21),(31),(32)$ and $u^{(m)} \equiv u^{(m)}(t-\tau_{km}(t))$, for $m\neq k$, $\Delta_{km} = \sqrt{c^2 - \left\langle \overrightarrow{u}^{(m)}(t-\tau_{km}),\overrightarrow{u}^{(m)}(t-\tau_{km})\right\rangle}$,

$$\lambda^{(m)} = \left(\lambda_1^{(m)},\lambda_2^{(m)},\lambda_3^{(m)},\lambda_4^{(m)}\right) = \left(\frac{u_1^{(m)}(t-\tau_{km})}{\Delta_{km}},\frac{u_2^{(m)}(t-\tau_{km})}{\Delta_{km}},\frac{u_3^{(m)}(t-\tau_{km})}{\Delta_{km}},\frac{ic}{\Delta_{km}}\right) = \left(\frac{\overrightarrow{u}^{(m)}(t-\tau_{km})}{\Delta_{km}},\frac{ic}{\Delta_{km}}\right).$$

2. Separation of the Equations of Motion and Derivation of the Initial Equations

The assumption (C): $\left\|\overrightarrow{u}^{(k)}\right\| = \sqrt{\left\langle \overrightarrow{u}^{(k)}(t),\overrightarrow{u}^{(k)}(t)\right\rangle} \leq \bar{c} < c\ (k=1,2,3;\ \bar{c}=const)$ implies $c^2 - \left\langle \overrightarrow{u}^{(k)},\overrightarrow{u}^{(k)}\right\rangle \geq c^2 - \bar{c}^2 > 0$. This condition allows us to obtain from (1) a system containing three groups of equations (cf. [1]):

$$\begin{aligned}
\dot{u}_1^{(1)} &= \frac{c^2-\left(u_1^{(1)}\right)^2}{c^2}\left(G_1^{(12)}+G_1^{(13)}+G_1^{(1)rad}\right) - \frac{u_1^{(1)}u_2^{(1)}}{c^2}\left(G_2^{(12)}+G_2^{(13)}+G_2^{(1)rad}\right) - \frac{u_1^{(1)}u_3^{(1)}}{c^2}\left(G_3^{(12)}+G_3^{(13)}+G_3^{(1)rad}\right) \equiv U_1^1\\
\dot{u}_2^{(1)} &= -\frac{u_1^{(1)}u_2^{(1)}}{c^2}\left(G_1^{(12)}+G_1^{(13)}+G_1^{(1)rad}\right) + \frac{c^2-\left(u_2^{(1)}\right)^2}{c^2}\left(G_2^{(12)}+G_2^{(13)}+G_2^{(1)rad}\right) - \frac{u_2^{(1)}u_3^{(1)}}{c^2}\left(G_3^{(12)}+G_3^{(13)}+G_3^{(1)rad}\right) \equiv U_2^1\\
\dot{u}_3^{(1)} &= -\frac{u_1^{(1)}u_3^{(1)}}{c^2}\left(G_1^{(12)}+G_1^{(13)}+G_1^{(1)rad}\right) - \frac{u_2^{(1)}u_3^{(1)}}{c^2}\left(G_2^{(12)}+G_2^{(13)}+G_2^{(1)rad}\right) + \frac{c^2-\left(u_3^{(1)}\right)^2}{c^2}\left(G_3^{(12)}+G_3^{(13)}+G_3^{(1)rad}\right) \equiv U_3^1,
\end{aligned} \tag{2}$$

$$\begin{aligned}
\dot{u}_1^{(2)} &= \frac{c^2-\left(u_1^{(2)}\right)^2}{c^2}\left(G_1^{(21)}+G_1^{(23)}+G_1^{(2)rad}\right) - \frac{u_1^{(2)}u_2^{(2)}}{c^2}\left(G_2^{(21)}+G_2^{(23)}+G_2^{(2)rad}\right) - \frac{u_1^{(2)}u_3^{(2)}}{c^2}\left(G_3^{(21)}+G_3^{(23)}+G_3^{(2)rad}\right) \equiv U_1^2\\
\dot{u}_2^{(2)} &= -\frac{u_1^{(2)}u_2^{(2)}}{c^2}\left(G_1^{(21)}+G_1^{(23)}+G_1^{(2)rad}\right) + \frac{c^2-\left(u_2^{(2)}\right)^2}{c^2}\left(G_2^{(21)}+G_2^{(23)}+G_2^{(2)rad}\right) - \frac{u_2^{(2)}u_3^{(2)}}{c^2}\left(G_3^{(21)}+G_3^{(23)}+G_3^{(2)rad}\right) \equiv U_2^2\\
\dot{u}_3^{(2)} &= -\frac{u_1^{(2)}u_3^{(2)}}{c^2}\left(G_1^{(21)}+G_1^{(23)}+G_1^{(2)rad}\right) - \frac{u_2^{(2)}u_3^{(2)}}{c^2}\left(G_2^{(21)}+G_2^{(23)}+G_2^{(2)rad}\right) + \frac{c^2-\left(u_3^{(2)}\right)^2}{c^2}\left(G_3^{(21)}+G_3^{(23)}+G_3^{(2)rad}\right) \equiv U_3^2,
\end{aligned} \tag{3}$$

$$\begin{aligned}
\dot{u}_1^{(3)} &= \frac{c^2-\left(u_1^{(3)}\right)^2}{c^2}\left(G_1^{(31)}+G_1^{(32)}+G_1^{(3)rad}\right)-\frac{u_1^{(3)}u_2^{(3)}}{c^2}\left(G_2^{(31)}+G_2^{(32)}+G_2^{(3)rad}\right)-\frac{u_1^{(3)}u_3^{(3)}}{c^2}\left(G_3^{(31)}+G_3^{(32)}+G_3^{(3)rad}\right)\equiv U_1^3\\
\dot{u}_2^{(3)} &= -\frac{u_1^{(3)}u_2^{(3)}}{c^2}\left(G_1^{(31)}+G_1^{(32)}+G_1^{(3)rad}\right)+\frac{c^2-\left(u_2^{(3)}\right)^2}{c^2}\left(G_2^{(31)}+G_2^{(32)}+G_2^{(3)rad}\right)-\frac{u_2^{(3)}u_3^{(3)}}{c^2}\left(G_3^{(31)}+G_3^{(32)}+G_3^{(3)rad}\right)\equiv U_2^3\\
\dot{u}_3^{(3)} &= -\frac{u_1^{(3)}u_3^{(3)}}{c^2}\left(G_1^{(31)}+G_1^{(32)}+G_1^{(3)rad}\right)-\frac{u_2^{(3)}u_3^{(3)}}{c^2}\left(G_2^{(31)}+G_2^{(32)}+G_2^{(3)rad}\right)+\frac{c^2-\left(u_3^{(3)}\right)^2}{c^2}\left(G_3^{(31)}+G_3^{(32)}+G_3^{(3)rad}\right)\equiv U_3^3,\ t\geq 0,
\end{aligned} \tag{4}$$

where

$$G_\alpha^{(12)}=\frac{e_1e_2}{m_1}\frac{\Delta_1}{c^3}\left\{\frac{\left(c^2-\left\langle\vec{u}^{(1)},\vec{u}^{(2)}\right\rangle\right)\xi_\alpha^{(12)}-\left(c^2\tau_{12}-\left\langle\vec{\xi}^{(12)},\vec{u}^{(1)}\right\rangle\right)u_\alpha^{(2)}}{\left(c^2\tau_{12}-\left\langle\vec{\xi}^{(12)},\vec{u}^{(2)}\right\rangle\right)^3}\frac{\Delta_{12}^4+D_{12}\left[\Delta_{12}^2\left\langle\vec{\xi}^{(12)},\dot{\vec{u}}^{(2)}\right\rangle+\left(\left\langle\vec{\xi}^{(12)},\vec{u}^{(2)}\right\rangle-c^2\tau_{12}\right)\left\langle u^{(2)},\dot{u}^{(2)}\right\rangle\right]}{\Delta_{12}^2}-\right.$$
$$\left.-\frac{D_{12}}{\left(\left\langle\vec{\xi}^{(12)},\vec{u}^{(2)}\right\rangle-c^2\tau_{12}\right)^2}\frac{\left(\Delta_{12}^2\dot{u}_\alpha^{(2)}+u_\alpha^{(2)}\left\langle\vec{u}^{(2)},\dot{\vec{u}}^{(2)}\right\rangle\right)\left(c^2\tau_{12}-\left\langle\vec{\xi}^{(12)},\vec{u}^{(1)}\right\rangle\right)+\left(\Delta_{12}^2\left\langle\vec{u}^{(1)},\dot{\vec{u}}^{(2)}\right\rangle+\left(\left\langle\vec{u}^{(1)},\vec{u}^{(2)}\right\rangle-c^2\right)\left\langle\vec{u}^{(2)},\dot{\vec{u}}^{(2)}\right\rangle\right)\xi_\alpha^{(12)}}{\Delta_{12}^2}\right\};$$

$$G_\alpha^{(13)}=\frac{e_1e_3}{m_1}\frac{\Delta_1}{c^3}\left\{\frac{\left(c^2-\left\langle\vec{u}^{(1)},\vec{u}^{(3)}\right\rangle\right)\xi_\alpha^{(13)}-\left(c^2\tau_{13}-\left\langle\vec{\xi}^{(13)},\vec{u}^{(1)}\right\rangle\right)u_\alpha^{(3)}}{\left(c^2\tau_{13}-\left\langle\vec{\xi}^{(13)},\vec{u}^{(3)}\right\rangle\right)^3}\frac{\Delta_{13}^4+D_{13}\left[\Delta_{13}^2\left\langle\vec{\xi}^{(13)},\dot{\vec{u}}^{(3)}\right\rangle+\left(\left\langle\vec{\xi}^{(13)},\vec{u}^{(3)}\right\rangle-c^2\tau_{13}\right)\left\langle\vec{u}^{(3)},\dot{\vec{u}}^{(3)}\right\rangle\right]}{\Delta_{13}^2}-\right.$$
$$\left.-\frac{D_{13}}{\left(\left\langle\vec{\xi}^{(13)},\vec{u}^{(3)}\right\rangle-c^2\tau_{13}\right)^2}\frac{\left(\Delta_{13}^2\dot{u}_\alpha^{(3)}+u_\alpha^{(3)}\left\langle\vec{u}^{(3)},\dot{\vec{u}}^{(3)}\right\rangle\right)\left(c^2\tau_{13}-\left\langle\vec{\xi}^{(13)},\vec{u}^{(1)}\right\rangle\right)+\left(\Delta_{13}^2\left\langle\vec{u}^{(1)},\dot{\vec{u}}^{(3)}\right\rangle+\left(\left\langle\vec{u}^{(1)},\vec{u}^{(3)}\right\rangle-c^2\right)\left\langle\vec{u}^{(3)},\dot{\vec{u}}^{(3)}\right\rangle\right)\xi_\alpha^{(13)}}{\Delta_{13}^2}\right\};$$

$$G_\alpha^{(21)}=\frac{e_2e_1}{m_2}\frac{\Delta_2}{c^3}\left\{\frac{\left(c^2-\left\langle\vec{u}^{(2)},\vec{u}^{(1)}\right\rangle\right)\xi_\alpha^{(21)}-\left(c^2\tau_{21}-\left\langle\vec{\xi}^{(21)},\vec{u}^{(2)}\right\rangle\right)u_\alpha^{(1)}}{\left(c^2\tau_{21}-\left\langle\vec{\xi}^{(21)},\vec{u}^{(1)}\right\rangle\right)^3}\frac{\Delta_{21}^4+D_{21}\left[\Delta_{21}^2\left\langle\vec{\xi}^{(21)},\dot{\vec{u}}^{(1)}\right\rangle+\left(\left\langle\vec{\xi}^{(21)},\vec{u}^{(1)}\right\rangle-c^2\tau_{21}\right)\left\langle\vec{u}^{(1)},\dot{\vec{u}}^{(1)}\right\rangle\right]}{\Delta_{21}^2}-\right.$$
$$\left.-\frac{D_{23}}{\left(\left\langle\vec{\xi}^{(21)},\vec{u}^{(1)}\right\rangle-c^2\tau_{21}\right)^2}\frac{\left(\Delta_{21}^2\dot{u}_\alpha^{(1)}+u_\alpha^{(1)}\left\langle\vec{u}^{(1)},\dot{\vec{u}}^{(1)}\right\rangle\right)\left(c^2\tau_{21}-\left\langle\vec{\xi}^{(21)},\vec{u}^{(1)}\right\rangle\right)+\left(\Delta_{21}^2\left\langle\vec{u}^{(2)},\dot{\vec{u}}^{(1)}\right\rangle+\left(\left\langle\vec{u}^{(2)},\vec{u}^{(1)}\right\rangle-c^2\right)\left\langle\vec{u}^{(1)},\dot{\vec{u}}^{(1)}\right\rangle\right)\xi_\alpha^{(21)}}{\Delta_{21}^2}\right\};$$

$$G_\alpha^{(23)}=\frac{e_2e_3}{m_2}\frac{\Delta_2}{c^3}\left\{\frac{\left(c^2-\left\langle\vec{u}^{(2)},\vec{u}^{(3)}\right\rangle\right)\xi_\alpha^{(23)}-\left(c^2\tau_{23}-\left\langle\vec{\xi}^{(23)},\vec{u}^{(2)}\right\rangle\right)u_\alpha^{(3)}}{\left(c^2\tau_{23}-\left\langle\vec{\xi}^{(23)},\vec{u}^{(3)}\right\rangle\right)^3}\frac{\Delta_{23}^4+D_{23}\left[\Delta_{23}^2\left\langle\vec{\xi}^{(23)},\dot{\vec{u}}^{(3)}\right\rangle+\left(\left\langle\vec{\xi}^{(23)},\vec{u}^{(3)}\right\rangle-c^2\tau_{23}\right)\left\langle\vec{u}^{(3)},\dot{\vec{u}}^{(3)}\right\rangle\right]}{\Delta_{23}^2}-\right.$$
$$\left.-\frac{D_{23}}{\left(\left\langle\vec{\xi}^{(23)},\vec{u}^{(3)}\right\rangle,-,c^2,\tau_{23}\right)^2}\frac{\left(\Delta_{23}^2\dot{u}_\alpha^{(3)}+u_\alpha^{(3)}\left\langle\vec{u}^{(3)},\dot{\vec{u}}^{(3)}\right\rangle\right)\left(c^2\tau_{23}-\left\langle\vec{\xi}^{(23)},\vec{u}^{(2)}\right\rangle\right)+\left(\Delta_{23}^2\left\langle\vec{u}^{(2)},\dot{\vec{u}}^{(3)}\right\rangle+\left(\left\langle\vec{u}^{(2)},\vec{u}^{(3)}\right\rangle-c^2\right)\left\langle\vec{u}^{(3)},\dot{\vec{u}}^{(3)}\right\rangle\right)\xi_\alpha^{(23)}}{\Delta_{23}^2}\right\};$$

$$G_\alpha^{(31)}=\frac{e_3e_1}{m_3}\frac{\Delta_3}{c^3}\left\{\frac{\left(c^2-\left\langle\vec{u}^{(3)},\vec{u}^{(1)}\right\rangle\right)\xi_\alpha^{(31)}-\left(c^2\tau_{31}-\left\langle\vec{\xi}^{(31)},\vec{u}^{(3)}\right\rangle\right)u_\alpha^{(1)}}{\left(c^2\tau_{31}-\left\langle\vec{\xi}^{(31)},,,\vec{u}^{(1)}\right\rangle\right)^3}\frac{\Delta_{31}^4+D_{31}\left[\Delta_{31}^2\left\langle\vec{\xi}^{(31)},\dot{\vec{u}}^{(1)}\right\rangle+\left(\left\langle\vec{\xi}^{(31)},\vec{u}^{(1)}\right\rangle-c^2\tau_{31}\right)\left\langle\vec{u}^{(1)},\dot{\vec{u}}^{(1)}\right\rangle\right]}{\Delta_{31}^2}-\right.$$
$$\left.-\frac{D_{31}}{\left(\left\langle\vec{\xi}^{(31)},\vec{u}^{(1)}\right\rangle-c^2\tau_{31}\right)^2}\frac{\left(\Delta_{31}^2\dot{u}_\alpha^{(1)}+u_\alpha^{(1)}\left\langle\vec{u}^{(1)},\dot{\vec{u}}^{(1)}\right\rangle\right)\left(c^2\tau_{31}-\left\langle\vec{\xi}^{(31)},\vec{u}^{(3)}\right\rangle\right)+\left(\Delta_{31}^2\left\langle\vec{u}^{(3)},\dot{\vec{u}}^{(1)}\right\rangle+\left(\left\langle\vec{u}^{(3)},\vec{u}^{(1)}\right\rangle-c^2\right)\left\langle\vec{u}^{(1)},\dot{\vec{u}}^{(1)}\right\rangle\right)\xi_\alpha^{(31)}}{\Delta_{31}^2}\right\};$$

$$G_\alpha^{(32)}=\frac{e_3e_2}{m_3}\frac{\Delta_3}{c^3}\left\{\frac{\left(c^2-\left\langle\vec{u}^{(3)},\vec{u}^{(2)}\right\rangle\right)\xi_\alpha^{(32)}-\left(c^2\tau_{32}-\left\langle\vec{\xi}^{(32)},\vec{u}^{(3)}\right\rangle\right)u_\alpha^{(2)}}{\left(c^2\tau_{32}-\left\langle\vec{\xi}^{(32)},\vec{u}^{(2)}\right\rangle\right)^3}\frac{\Delta_{32}^4+D_{32}\left[\Delta_{32}^2\left\langle\vec{\xi}^{(32)},\dot{\vec{u}}^{(2)}\right\rangle+\left(\left\langle\vec{\xi}^{(32)},\vec{u}^{(2)}\right\rangle-c^2\tau_{32}\right)\left\langle\vec{u}^{(2)},\dot{\vec{u}}^{(2)}\right\rangle\right]}{\Delta_{32}^2}-\right.$$
$$\left.-\frac{D_{32}}{\left(\left\langle\vec{\xi}^{(32)},\vec{u}^{(2)}\right\rangle-c^2\tau_{32}\right)^2}\frac{\left(\Delta_{32}^2\dot{u}_\alpha^{(2)}+u_\alpha^{(2)}\left\langle\vec{u}^{(2)},\dot{\vec{u}}^{(2)}\right\rangle\right)\left(c^2\tau_{32}-\left\langle\vec{\xi}^{(32)},\vec{u}^{(3)}\right\rangle\right)+\left(\Delta_{32}^2\left\langle\vec{u}^{(3)},\dot{\vec{u}}^{(2)}\right\rangle+\left(\left\langle\vec{u}^{(3)},\vec{u}^{(2)}\right\rangle-c^2\right)\left\langle\vec{u}^{(2)},\dot{\vec{u}}^{(2)}\right\rangle\right)\xi_\alpha^{(32)}}{\Delta_{32}^2}\right\};$$

$$G_\alpha^{(k)rad}=-\frac{e_k^2}{m_kc^2}\left(\frac{u_\alpha^{(k)}(t)}{{\Delta_k}^3}\left\langle\vec{u}^{(k)}(t),\ddot{\vec{u}}^{(k)}(t)\right\rangle+\frac{\ddot{u}_\alpha^{(k)}(t)}{\Delta_k}\right),(k=1,2,3), \tag{5}$$

where $D_{km} = \frac{c\sqrt{\left\langle \vec{\xi}^{(km)}, \vec{\xi}^{(km)} \right\rangle} - \left\langle \vec{\xi}^{(km)}, \vec{u}^{(m)} \right\rangle}{c\sqrt{\left\langle \vec{\xi}^{(km)}, \vec{\xi}^{(km)} \right\rangle} - \left\langle \vec{\xi}^{(km)}, \vec{u}^{(k)} \right\rangle}$ for $m \neq k$.

In accordance with the Kepler formulation, the first particle P_1 is at the origin $O(0,0,0)$, and therefore,

$$x_\alpha^{(1)}(t) = 0,\ u_\alpha^{(1)}(t) = 0,\ \dot{u}_\alpha^{(1)}(t) = 0\ (\alpha = 1,2,3). \tag{6}$$

Then, we obtain $0 = G_1^{(12)} + G_1^{(13)}$, $0 = G_2^{(12)} + G_2^{(13)}$, $0 = G_3^{(12)} + G_3^{(13)}$.

It is easy to see that the equation $0 = G_1^{(12)} + G_1^{(13)}$ becomes $0 = 0$.

Remark 1. *We notice that in the equation* $0 = G_2^{(12)} + G_2^{(13)}$*, the unknown functions and their derivatives have retarded arguments* $t - \tau_{12}$*, while in the equation* $0 = G_3^{(12)} + G_3^{(13)}$*, the arguments are* $t - \tau_{13}$*. This implies that we must consider these equations on the initial interval* $[-T, 0]$*. This is why we call the first group of three equations (2) initial equations. We consider the second and third group of Equations (3) and (4), respectively (six in number), on the interval* $[0, \infty)$*.*

To simplify the equations of motion, we notice that for the first Bohr orbit, the Sommerfeld fine structure constant is $\beta = \bar{c}/c \approx 1/137 \Rightarrow \beta^2 \approx 0$. Consequently, $\Delta_k = \sqrt{c^2 - \left\langle \vec{u}^{(k)}, \vec{u}^{(k)} \right\rangle} = c\sqrt{1 - \left\langle \vec{u}^{(k)}, \vec{u}^{(k)} \right\rangle / c^2} \approx c$, $\Delta_{km} = \sqrt{c^2 - \left\langle \vec{u}^{(m)}(t - \tau_{km}), \vec{u}^{(m)}(t - \tau_{km}) \right\rangle} \approx c$, and $c^2\left(1 - \frac{\left\langle \vec{u}^{(2)}, \vec{u}^{(3)} \right\rangle}{c^2}\right) \approx c^2$.

If ρ_{km} is the distance between the k-th and m-th particle via $\xi^{(km)} \approx \rho_{km} \approx 10^{-11}$ m (cf. [14]), we obtain

$$c^2\tau_{23} - \left\langle \vec{\xi}^{(23)}, \vec{u}^{(2)} \right\rangle \approx c^2\tau_{23},\ D_{23} = \frac{c^2\tau_{23} - \left\langle \vec{\xi}^{(23)}, \vec{u}^{(3)} \right\rangle}{c^2\tau_{23} - \left\langle \vec{\xi}^{(23)}, \vec{u}^{(2)} \right\rangle} \approx 1 \text{ and}$$

$$H_{23} = \frac{\Delta_{23}^4 + D_{23}\left[\Delta_{23}^2\left\langle \vec{\xi}^{(23)}, \dot{\vec{u}}^{(3)} \right\rangle - \left(c^2\tau_{23} - \left\langle \vec{\xi}^{(23)}, \vec{u}^{(3)} \right\rangle\right)\left\langle \vec{u}^{(3)}, \dot{\vec{u}}^{(3)} \right\rangle\right]}{\Delta_{23}^2} \approx c^2 + \left\langle \vec{\xi}^{(23)} - \tau_{23}\vec{u}^{(3)}, \dot{\vec{u}}^{(3)} \right\rangle.$$

We can simplify the above equations in accordance to the last relations:

$$G_\alpha^{(12)} = \frac{e_1e_2}{m_1}\frac{1}{c^6}\frac{\left(c^2 + \left\langle \vec{\xi}^{(12)}, \dot{\vec{u}}^{(2)} \right\rangle\right)\left(\xi_\alpha^{(12)} - \tau_{12}u_\alpha^{(2)}\right) - c^2{\tau_{12}}^2\dot{u}_\alpha^{(2)}}{{\tau_{12}}^3};$$

$$G_\alpha^{(13)} = \frac{e_1e_3}{m_1}\frac{1}{c^6}\frac{\left(c^2 + \left\langle \vec{\xi}^{(13)}, \dot{\vec{u}}^{(3)} \right\rangle\right)\left(\xi_\alpha^{(13)} - \tau_{13}u_\alpha^{(3)}\right) - c^2{\tau_{13}}^2\dot{u}_\alpha^{(3)}}{{\tau_{13}}^3};$$

$$G_\alpha^{(21)} = \frac{e_2e_1}{m_2}\frac{\Delta_2}{c^3}\frac{c^2\xi_\alpha^{(21)}}{\left(c^2\tau_{21}\right)^3}\frac{\Delta_{21}^4}{\Delta_{21}^2} \approx \frac{e_2e_1}{m_2}\frac{\xi_\alpha^{(21)}}{c^4{\tau_{21}}^3};\ G_\alpha^{(23)} = \frac{e_2e_3}{m_2c^3}\left(\frac{\xi_\alpha^{(23)}}{{\tau_{23}}^3}\frac{\left\langle \vec{\xi}^{(23)}, \dot{\vec{u}}^{(3)} \right\rangle}{c^3} - \frac{\dot{u}_\alpha^{(3)}}{c\tau_{23}}\right);$$

$$G_\alpha^{(31)} = \frac{e_3e_1}{m_3}\frac{\Delta_3}{c^3}\frac{c^2\xi_\alpha^{(31)}}{\left(c^2\tau_{31}\right)^3}\frac{\Delta_{31}^4}{\Delta_{31}^2} \approx \frac{e_3e_1}{m_3}\frac{\xi_\alpha^{(31)}}{c^4{\tau_{31}}^3};\ G_\alpha^{(32)} = \frac{e_2e_3}{m_3c^3}\left(\frac{\xi_\alpha^{(32)}}{{\tau_{32}}^3}\frac{\left\langle \vec{\xi}^{(32)}, \dot{\vec{u}}^{(3)} \right\rangle}{c^3} - \frac{\dot{u}_\alpha^{(2)}}{c\tau_{32}}\right);$$

$$G_\alpha^{(2)rad} = -\frac{e_2^2}{m_2c^2}\left(\frac{u_\alpha^{(2)}\left\langle \vec{u}^{(2)}, \ddot{\vec{u}}^{(2)} \right\rangle}{c^3} + \frac{\ddot{u}_\alpha^{(2)}}{c}\right);\ G_\alpha^{(3)rad} = -\frac{e_3^2}{m_3c^2}\left(\frac{u_\alpha^{(3)}\left\langle \vec{u}^{(3)}, \ddot{\vec{u}}^{(3)} \right\rangle}{c^3} + \frac{\ddot{u}_\alpha^{(3)}}{c}\right).$$

For the stationary particle, we have $G_\alpha^{(1)rad} = 0$. Therefore, the initial (vector) equations, $G_1^{(12)} + G_1^{(13)} = 0, G_2^{(12)} + G_2^{(13)} = 0, G_3^{(12)} + G_3^{(13)} = 0$, become:

$$\begin{aligned}
&\tfrac{e_1e_2}{m_1}\left[\left(c^2+\left\langle \vec{\xi}^{(12)}, \dot{\vec{u}}^{(2)}\right\rangle\right)\left(\xi_1^{(12)}-\tau_{12}u_1^{(2)}\right)-c^2{\tau_{12}}^2\dot{u}_1^{(2)}\right]+\tfrac{e_3e_1}{m_1}\left[\left(c^2+\left\langle \vec{\xi}^{(13)}, \dot{\vec{u}}^{(3)}\right\rangle\right)\left(\xi_1^{(13)}-\tau_{13}u_1^{(3)}\right)-c^2{\tau_{13}}^2\dot{u}_1^{(3)}\right]=0\\
&\tfrac{e_1e_2}{m_2}\left[\left(c^2+\left\langle \vec{\xi}^{(12)}, \dot{\vec{u}}^{(2)}\right\rangle\right)\left(\xi_2^{(12)}-\tau_{12}u_2^{(2)}\right)-c^2{\tau_{12}}^2\dot{u}_2^{(2)}\right]+\tfrac{e_3e_1}{m_2}\left[\left(c^2+\left\langle \vec{\xi}^{(13)}, \dot{\vec{u}}^{(3)}\right\rangle\right)\left(\xi_2^{(13)}-\tau_{13}u_2^{(3)}\right)-c^2{\tau_{13}}^2\dot{u}_2^{(3)}\right]=0,\\
&\tfrac{e_1e_2}{m_3}\left[\left(c^2+\left\langle \vec{\xi}^{(12)}, \dot{\vec{u}}^{(2)}\right\rangle\right)\left(\xi_3^{(12)}-\tau_{12}u_3^{(2)}\right)-c^2{\tau_{12}}^2\dot{u}_3^{(2)}\right]+\tfrac{e_3e_1}{m_3}\left[\left(c^2+\left\langle \vec{\xi}^{(13)}, \dot{\vec{u}}^{(3)}\right\rangle\right)\left(\xi_3^{(13)}-\tau_{13}u_3^{(3)}\right)-c^2{\tau_{13}}^2\dot{u}_3^{(3)}\right]=0.
\end{aligned} \tag{7}$$

We assume that

$$\begin{aligned}
&\left(c^2+\left\langle \vec{\xi}^{(12)}, \dot{\vec{u}}^{(2)}\right\rangle\right)\left(\xi_1^{(12)}-\tau_{12}u_1^{(2)}\right)-c^2{\tau_{12}}^2\dot{u}_1^{(2)}=0,\\
&\left(c^2+\left\langle \vec{\xi}^{(12)}, \dot{\vec{u}}^{(2)}\right\rangle\right)\left(\xi_2^{(12)}-\tau_{12}u_2^{(2)}\right)-c^2{\tau_{12}}^2\dot{u}_2^{(2)}=0,\\
&\left(c^2+\left\langle \vec{\xi}^{(12)}, \dot{\vec{u}}^{(2)}\right\rangle\right)\left(\xi_3^{(12)}-\tau_{12}u_3^{(2)}\right)-c^2{\tau_{12}}^2\dot{u}_3^{(2)}=0.
\end{aligned} \tag{8}$$

and

$$\begin{aligned}
&\left(c^2+\left\langle \vec{\xi}^{(13)}, \dot{\vec{u}}^{(3)}\right\rangle\right)\left(\xi_1^{(13)}-\tau_{13}u_1^{(3)}\right)-c^2{\tau_{13}}^2\dot{u}_1^{(3)}=0\\
&\left(c^2+\left\langle \vec{\xi}^{(13)}, \dot{\vec{u}}^{(3)}\right\rangle\right)\left(\xi_2^{(13)}-\tau_{13}u_2^{(3)}\right)-c^2{\tau_{13}}^2\dot{u}_2^{(3)}=0,\\
&\left(c^2+\left\langle \vec{\xi}^{(13)}, \dot{\vec{u}}^{(3)}\right\rangle\right)\left(\xi_3^{(13)}-\tau_{13}u_3^{(3)}\right)-c^2{\tau_{13}}^2\dot{u}_3^{(3)}=0.
\end{aligned} \tag{9}$$

The system (8) contains functions with the argument $t-\tau_{12}(t)\in[-T,0]$ for $t\in[0,T]$. The system (9) contains functions with the argument $t-\tau_{13}(t)\in[-T,0]$ for $t\in[0,T]$. If (8) and (9) are satisfied, then (7) is also satisfied.

Remark 2. *The choice of* $t-\tau_{km}(t)\in[-T,0]$ *follows from the next reasoning: Let T be the period of the solution. We notice that all the arguments of the unknown functions are* $t-\tau_{km}$, *that is,* $\vec{u}=\vec{u}^{(m)}(t-\tau_{km})$. *Therefore, we must look for a solution on the initial set, that is, for* $t-\tau_{km}(t)\in[\tau_0;\ 0]$, *where* $\tau_0=\min\{t-\tau_{km}(t): t\in[0,T]\}$. *Since* $1-d\tau_{km}(t)/dt>0$, *then* $t-\tau_{km}(t)$ *is an increasing function. We have proved that if the trajectories are T-periodic, then* $\tau_{km}(t)$ *is T-periodic too.*

It follows that $-T-\tau_{km}(-T)\le 0-\tau_{km}(0)\Leftrightarrow -T-\tau_{km}(0)\le-\tau_{km}(0)\Leftrightarrow\tau_{km}(0)\le T-\tau_{km}(0)$ and then $0-\tau_{km}(0)\le t-\tau_{km}(t)\le T-\tau_{km}(T)=T-\tau_{km}(0)=0$. Consequently, $-T\le t-\tau_{km}(t)\le 0$. Therefore, the functions $t-\tau_{km}(t):[0,T]\to[-T,0]$, which means that $\tau_0=-T$, and we put $t-\tau_{km}(t)=\theta\in[-T,0]$. The arguments of the unknown functions $\vec{u}^{(2)}=\vec{u}^{(2)}(\theta)$, $\vec{u}^{(3)}=\vec{u}^{(3)}(\theta)$ belong to $[-T,0]$.

After the transformations given in Appendix A, we obtain from (8) and (9) the *initial equations*:

$$\begin{aligned}
\ddot{\rho}_2(t-\tau_{23})&=\rho_2(t-\tau_{23}){\dot{\varphi}_2}^2(t-\tau_{23})\cos^2\lambda_2(t-\tau_{23})+\rho_2(t-\tau_{23}){\dot{\lambda}_2}^2(t-\tau_{23})-\tfrac{c^2}{\rho_2(t-\tau_{23})},\\
\ddot{\varphi}_2(t-\tau_{23})&=2\dot{\varphi}_2(t-\tau_{23})\dot{\lambda}_2(t-\tau_{23})tg\lambda_2(t-\tau_{23})-\tfrac{2\dot{\rho}_2(t-\tau_{23})\dot{\varphi}_2(t-\tau_{23})}{\rho_n(t-\tau_{23})},\\
\ddot{\lambda}_2(t-\tau_{23})&=-\tfrac{2\dot{\rho}_2(t-\tau_{23})\dot{\lambda}_2(t-\tau_{23})+\rho_2(t-\tau_{23}){\dot{\varphi}_2}^2(t-\tau_{23})\sin\lambda_2(t-\tau_{23})\cos\lambda_2(t-\tau_{23})}{\rho_2(t-\tau_{23})}.
\end{aligned} \tag{10-1}$$

for the second particle and

$$\ddot{\rho}_3(t-\tau_{32}) = \rho_3(t-\tau_{32})\dot{\varphi}_3{}^2(t-\tau_{32})\cos^2\lambda_3(t-\tau_{32}) + \rho_3(t-\tau_{32})\dot{\lambda}_3{}^2(t-\tau_{32}) - \frac{c^2}{\rho_3(t-\tau_{32})},$$
$$\ddot{\varphi}_3(t-\tau_{32}) = 2\dot{\varphi}_3(t-\tau_{32})\dot{\lambda}_3(t-\tau_{32})tg\lambda_3(t-\tau_{32}) - \frac{2\dot{\rho}_3(t-\tau_{32})\dot{\varphi}_3(t-\tau_{32})}{\rho_3(t-\tau_{32})}, \quad (10\text{-}2)$$
$$\ddot{\lambda}_3(t-\tau_{32}) = -\frac{2\dot{\rho}_3(t-\tau_{32})\dot{\lambda}_3(t-\tau_{32})+\rho_3(t-\tau_{32})\dot{\varphi}_3{}^2(t-\tau_{32})\sin\lambda_3(t-\tau_{32})\cos\lambda_3(t-\tau_{32})}{\rho_3(t-\tau_{32})}.$$

for the third particle.

The reasoning from Remark 2 implies that we look for a solution (10-1) and (10-2) on the interval $[-T, 0]$.

The existence–uniqueness result for $(10-n)$ $(n = 1, 2)$ can be obtained as in [19] in the functional space $M_{r_2}{}^0 \times M_{\phi_2}{}^0 \times M_{\eta_2}{}^0 \times M_{r_3}{}^0 \times M_{\phi_3}{}^0 \times M_{\eta_3}{}^0$, where

$$M_{r_2}{}^0 = \left\{r_{20}(.) \in C_T{}^\infty[-T,0] : \left|r_{20}{}^{(m)}(t)\right| \le \omega_2{}^m R_2 e^{\mu(t-T_k)};\ r_{20}(-T) = r_{20}(0) = 0\right\},$$
$$M_{\phi_2}{}^0 = \left\{\phi_{20}(.) \in C_T{}^\infty[-T,0] : \left|\phi_{20}{}^{(m)}(t)\right| \le \omega_2{}^m \Phi_2 e^{\mu(t-T_k)},\ \phi_{20}(-T) = \phi_{20}(0) = 0\right\},$$
$$M_{\eta_2}{}^0 = \left\{\eta_{20}(.) \in C_T{}^\infty[-T,0] : \left|\eta_2{}^{(m)}(t)\right| \le \omega_2{}^m \Phi_2 e^{\mu(t-T_k)},\ \eta_{20}(-T) = \eta_{20}(0) = 0\right\},$$
$$M_{r_3}{}^0 = \left\{r_{30}(.) \in C_T{}^\infty[-T,0] : \left|r_{30}{}^{(m)}(t)\right| \le \omega_3{}^m R_3 e^{\mu(t-T_k)};\ r_{30}(-T) = r_{30}(0) = 0\right\},$$
$$M_{\phi_3}{}^0 = \left\{\phi_{30} \in C_T{}^\infty[-T,0] : \left|\phi_{30}{}^{(m)}(t)\right| \le \omega_3{}^m \Phi_3 e^{\mu(t-T_k)},\ \phi_{30}(-T) = \phi_{30}(0) = 0\right\},$$
$$M_{\eta_3}{}^0 = \left\{\eta_{30} \in C_T{}^\infty[-T,0] : \left|\eta_{30}{}^{(m)}(t)\right| \le \omega_3{}^m Y_3 e^{\mu(t-T_k)},\ \eta_{30}(-T) = \eta_{30}(0) = 0\right\}.$$

as fixed points of the operators

$$B_{r_2}(r_2,\phi_2,\eta_2,r_3,\phi_3,\eta_3)(t) := \int_{-T}^{t} G_{r_2}(s)ds - \frac{t+T}{T}\int_{-T}^{0} G_{r_2}(s)ds,\ t \in [-T,0],$$
$$B_{r_3}(r_2,\phi_2,\eta_2,r_3,\phi_3,\eta_3)(t) := \int_{-T}^{t} G_{r_3}(s)ds - \frac{t+T}{T}\int_{-T}^{0} G_{r_3}(s)ds,\ t \in [-T,0],$$
$$B_{\phi_2}(r_2,\phi_2,\eta_2,r_3,\phi_3,\eta_3)(t) := \int_{-T}^{t} G_{\phi_2}(s)ds - \frac{t+T}{T}\int_{-T}^{0} G_{\phi_2}(s)ds,\ t \in [-T,0],$$
$$B_{\phi_3}(r_2,\phi_2,\eta_2,r_3,\phi_3,\eta_3)(t) := \int_{-T}^{t} G_{r_3}(s)ds - \frac{t+T}{T}\int_{-T}^{0} G_{r_3}(s)ds,\ t \in [-T,0],$$
$$B_{\eta_2}(r_2,\phi_2,\eta_2,r_3,\phi_3,\eta_3)(t) := \int_{-T}^{t} G_{\eta_2}(s)ds - \frac{t+T}{T}\int_{-T}^{0} G_{\eta_2}(s)ds,\ t \in [-T,0],$$
$$B_{\eta_3}(r_2,\phi_2,\eta_2,r_3,\phi_3,\eta_3)(t) := \int_{-T}^{t} G_{\eta_3}(s)ds - \frac{t+T}{T}\int_{-T}^{0} G_{\eta_3}(s)ds,\ t \in [-T,0].$$

It is easy to verify that

$$B_{r_2}(.)(-T) = \int_{-T}^{-T} G_{r_2}(s)ds - \frac{-T+T}{T}\int_{-T}^{0} G_{r_2}(s)ds = 0 = \int_{-T}^{0} G_{r_2}(s)ds - \frac{T}{T}\int_{-T}^{0} G_{r_2}(s)ds = B_{r_2}(.)(0)$$

and so on. Then, we take the solutions as the initial functions $r_{20}(t), \phi_{20}(t), \eta_{20}(t), r_{30}(t), \phi_{30}(t), \eta_{30}(t)$ for the basic equations below.

3. Basic Equations in Spherical Coordinates

We note that $\varphi_{23} = \varphi_2(t) - \varphi_3(t-\tau_{23})$, $\varphi_{32} = \varphi_3(t) - \varphi_2(t-\tau_{32})$.

From (2)–(4), one obtains, in Appendix B, the final form of the equations of motion of the second and third particles:

$$\ddot{\rho}_2(t) = \frac{e_2e_1}{m_2}\frac{1}{c\rho_2{}^2(t)} - \frac{e_2e_3}{m_2}\frac{\rho_2(t)-\rho_3(t-\tau_{23})(\cos\varphi_{23}\cos\lambda_2(t)\cos\lambda_3(t-\tau_{23})+\sin\lambda_2(t)\sin\lambda_3(t-\tau_{23}))}{c^3}\frac{\left\langle \vec{\xi}^{(23)}, \dot{\vec{u}}^{(3)} \right\rangle}{c^3\tau_{23}{}^3} -$$

$$- \frac{e_2e_3}{m_2}\frac{\ddot{\rho}_3(\cos\varphi_{23}\cos\lambda_2(t)\cos\lambda_3(t-\tau_{23})+\sin\lambda_2(t)\sin\lambda_3(t-\tau_{23}))+\rho_3(t-\tau_{23})\ddot{\varphi}_3(t-\tau_{23})\sin\varphi_{23}\cos\lambda_2(t)\cos\lambda_3(t-\tau_{23})}{c^4\tau_{23}} -$$

$$- \frac{e_2e_3}{m_2}\frac{\rho_3\ddot{\lambda}_3(\sin\lambda_2(t)\cos\lambda_3(t-\tau_{23})-\cos\varphi_{23}\sin\lambda_3(t-\tau_{23})\cos\lambda_2(t))}{c^4\tau_{23}} - \frac{e_2^2}{m_2}\frac{\dddot{\rho}_2(t)}{c^3};$$

$$\ddot{\varphi}_2(t) = \frac{e_2e_3}{m_2}\left[\frac{\rho_3(t-\tau_{23})\sin\varphi_{23}\cos\lambda_3(t-\tau_{23})}{\tau_{23}{}^3\rho_2(t)\cos\lambda_2(t)}\frac{\left\langle \vec{\xi}^{(23)}, \dot{\vec{u}}^{(3)} \right\rangle}{c^6} + \right.$$

$$+ \frac{\ddot{\rho}_3(t-\tau_{23})\cos\lambda_3(t-\tau_{23})\sin\varphi_{23}+\rho_3(t-\tau_{23})\ddot{\varphi}_3(t-\tau_{23})\cos\lambda_3(t-\tau_{23})\cos\varphi_{23}}{c^4\tau_{23}\rho_2(t)\cos\lambda_2(t)} -$$

$$\left. - \frac{\rho_3(t-\tau_{23})\ddot{\lambda}_3(t-\tau_{23})\sin\lambda_3(t-\tau_{23})\sin\varphi_{23}}{c^4\tau_{23}\rho_2(t)\cos\lambda_2(t)} - \frac{\dddot{\varphi}_2(t)}{c^3}\right];$$

$$\ddot{\lambda}_2(t) = \frac{e_2e_3}{m_2}\frac{\rho_3(t-\tau_{23})\cos\varphi_{23}\sin\lambda_2(t)\cos\lambda_3(t-\tau_{23})}{c^3\tau_{23}{}^3}\frac{\left\langle \vec{\xi}^{(23)}, \dot{\vec{u}}^{(3)} \right\rangle}{c^3\rho_2} +$$

$$+ \frac{e_2e_3}{m_2}\frac{\ddot{\rho}_3(\cos\varphi_{23}\cos\lambda_3(t-\tau_{23})\sin\lambda_2(t)-\sin\lambda_3(t-\tau_{23})\cos\lambda_2(t))+\rho_3\ddot{\varphi}_3\sin\varphi_{23}\sin\lambda_2(t)\cos\lambda_3(t-\tau_{23})}{c^4\tau_{23}\rho_2(t)} +$$

$$+ \frac{e_2e_3}{m_2}\frac{-\rho_3\ddot{\lambda}_3(\cos\varphi_{23}\sin\lambda_2(t)\sin\lambda_3(t-\tau_{23})+\cos\lambda_2(t)\cos\lambda_3(t-\tau_{23})}{c^4\tau_{23}\rho_2(t)} - \frac{e_2^2}{m_2}\frac{\dddot{\lambda}_2(t)}{c^3};$$

where

$$\ddot{\rho}_3(t) = \frac{e_3e_1}{m_3}\frac{1}{c\rho_3{}^2(t)} - \frac{e_2e_3}{m_3}\frac{\rho_3(t)-\rho_2(t-\tau_{32})(\cos\varphi_{32}\cos\lambda_2(t-\tau_{32})\cos\lambda_3(t)+\sin\lambda_2(t-\tau_{32})\sin\lambda_3(t))}{c^3}\frac{\left\langle \vec{\xi}^{(32)}, \dot{\vec{u}}^{(2)} \right\rangle}{c^3\tau_{32}{}^3}$$

$$- \frac{e_3e_2}{m_3}\frac{\ddot{\rho}_2(\cos\varphi_{32}\cos\lambda_2(t-\tau_{32})\cos\lambda_3(t)+\sin\lambda_2(t-\tau_{32})\sin\lambda_3(t))+\rho_2(t-\tau_{32})\ddot{\varphi}_2(t-\tau_{32})\sin\varphi_{32}\cos\lambda_2(t-\tau_{32})\cos\lambda_3(t)}{c^4\tau_{23}}$$

$$- \frac{e_3e_2}{m_3}\frac{\rho_2(t-\tau_{32})\ddot{\lambda}_2(t-\tau_{32})(\sin\lambda_3(t)\cos\lambda_2(t-\tau_{32})-\cos\varphi_{32}\sin\lambda_2(t-\tau_{32})\cos\lambda_3(t))}{c^4\tau_{23}} - \frac{e_3^2}{m_3}\frac{\dddot{\rho}_3(t)}{c^3};$$

$$\ddot{\varphi}_3 = \frac{e_3e_2}{m_3}\left[\frac{\rho_2(t-\tau_{32})\sin\varphi_{32}\cos\lambda_2(t-\tau_{32})}{\tau_{32}{}^3\rho_3(t)\cos\lambda_3(t)}\frac{\left\langle \vec{\xi}^{(32)}, \dot{\vec{u}}^{(2)} \right\rangle}{c^6} + \right.$$

$$+ \frac{\ddot{\rho}_2(t-\tau_{32})\cos\lambda_2(t-\tau_{32})\sin\varphi_{32}+\rho_2(t-\tau_{32})\ddot{\varphi}_2(t-\tau_{32})\cos\lambda_2(t-\tau_{32})\cos\varphi_{32}}{c^4\tau_{32}\rho_3\cos\lambda_3(t)} -$$

$$\left. - \frac{\rho_2(t-\tau_{32})\ddot{\lambda}_2(t-\tau_{32})\sin\lambda_2(t-\tau_{32})\sin\varphi_{32}}{c^4\tau_{32}\rho_3\cos\lambda_3(t)} - \frac{\dddot{\varphi}_3(t)}{c^3}\right];$$

$$\ddot{\lambda}_3(t) = \frac{e_3e_2}{m_3}\frac{\rho_2(t-\tau_{32})\cos\varphi_{32}\sin\lambda_3(t)\cos\lambda_2(t-\tau_{32})}{c^3\tau_{32}{}^3}\frac{\left\langle \vec{\xi}^{(32)}, \dot{\vec{u}}^{(2)} \right\rangle}{c^3\rho_3(t)} +$$

$$+ \frac{e_3e_2}{m_3c^3}\frac{\ddot{\rho}_3(\cos\varphi_{32}\cos\lambda_2(t-\tau_{32})\sin\lambda_3(t)-\sin\lambda_2(t-\tau_{32})\cos\lambda_3(t))+\rho_2(t-\tau_{32})\ddot{\varphi}_2(t-\tau_{32})\sin\varphi_{32}\sin\lambda_3(t)\cos\lambda_2(t-\tau_{32})}{c\tau_{32}\rho_3(t)}$$

$$+ \frac{e_3e_2}{m_3c^3}\frac{-\rho_2(t-\tau_{32})\ddot{\lambda}_2(t-\tau_{32})(\cos\varphi_{23}\sin\lambda_2(t-\tau_{32})\sin\lambda_3(t)+\cos\lambda_2(t-\tau_{32})\cos\lambda_3(t)}{c\tau_{32}\rho_3(t)} - \frac{e_3^2}{m_3}\frac{\dddot{\lambda}_3(t)}{c^3}.$$

and

$$\left\langle \overset{\rightarrow}{\xi}^{(32)}, \overset{\cdot}{\vec{u}}^{(2)} \right\rangle =$$

$$= [\rho_3(t)\cos(\varphi_3(t)-\varphi_2(t-\tau_{32})\cos\lambda_3(t)\cos\lambda_2(t-\tau_{32})+\rho_3(t)\sin\lambda_3(t)\sin\lambda_2(t-\tau_{32})-\rho_2(t-\tau_{32})]\ddot{\rho}_2(t-\tau_{32})+$$

$$+\sin(\varphi_3(t)-\varphi_2(t-\tau_{32}))\cos\lambda_3(t)\cos\lambda_2(t-\tau_{32})\rho_2(t-\tau_{32})\rho_3(t)\ddot{\varphi}_2(t-\tau_{32})+$$

$$+[\sin\lambda_3(t)\cos\lambda_2(t-\tau_{32})-\cos(\varphi_3(t)-\varphi_2(t-\tau_{32}))\sin\lambda_2(t-\tau_{32})\cos\lambda_3(t)]\rho_2(t-\tau_{32})\rho_3(t)\ddot{\lambda}_2(t-\tau_{32}).$$

After the substitution,

$$\dot{\rho}_n = r_n,\ \dot{\varphi}_n = \phi_n, \dot{\lambda}_n = \eta_n\ (n=2,3) \Rightarrow \rho_n(t) = \rho_{n0} + \int_0^t r_n(s)ds,\ \varphi_n = \varphi_{n0} + \int_0^t \phi_n(s)ds,\ \lambda_n = \lambda_{n0} + \int_0^t \eta_n(s)ds$$

the system of equations of motion of the second and third particles can be rewritten as

$$\dot{r}_2 = \frac{e_2e_1}{m_2}\frac{1}{c\rho_2{}^2} - \frac{e_2e_3}{m_2}\frac{\rho_2-\rho_3(\cos\varphi_{23}\cos\lambda_2\cos\lambda_3+\sin\lambda_2\sin\lambda_3)}{c^3}\frac{\left\langle \overset{\rightarrow}{\xi}^{(23)}, \overset{\cdot}{\vec{u}}^{(3)} \right\rangle}{c^3\tau_{23}{}^3}$$

$$-\frac{e_2e_3}{m_2}\frac{\dot{r}_3(\cos\varphi_{23}\cos\lambda_2\cos\lambda_3+\sin\lambda_2\sin\lambda_3)+\rho_3\dot{\phi}_3\sin\varphi_{23}\cos\lambda_2\cos\lambda_3+\rho_3\dot{\eta}_3(\sin\lambda_2\cos\lambda_3-\cos\varphi_{23}\sin\lambda_3\cos\lambda_2)}{c^4\tau_{23}} - \frac{e_2^2}{m_2}\frac{\ddot{r}_2}{c^3} \equiv G_{r_2};$$

$$\dot{\phi}_2 = \frac{e_2e_3}{m_2}\left[\frac{\rho_3\sin\varphi_{23}\cos\lambda_3}{\tau_{23}{}^3\rho_2\cos\lambda_2}\frac{\left\langle \overset{\rightarrow}{\xi}^{(23)}, \overset{\cdot}{\vec{u}}^{(3)} \right\rangle}{c^6} + \frac{\dot{r}_3\cos\lambda_3\sin\varphi_{23}+\rho_3\dot{\phi}_3\cos_3\cos\varphi_{23}-\rho_3\dot{\eta}_3\sin\lambda_3\sin\varphi_{23}}{c^4\tau_{23}\rho_2\cos\lambda_2} - \frac{\ddot{\phi}_2}{c^3}\right] \equiv G_{\phi_2};$$

$$\dot{\eta}_2 = \frac{e_2e_3}{m_2}\frac{\rho_3\cos\varphi_{23}\sin\lambda_2\cos\lambda_3}{c^3\tau_{23}{}^3}\frac{\left\langle \overset{\rightarrow}{\xi}^{(23)}, \overset{\cdot}{\vec{u}}^{(3)} \right\rangle}{c^3\rho_2}+$$

$$+\frac{e_2e_3}{m_2}\frac{\dot{r}_3(\cos\varphi_{23}\cos\lambda_3\sin\lambda_2-\sin\lambda_3\cos\lambda_2)+\rho_3\dot{\phi}_3\sin\varphi_{23}\sin\lambda_2\cos\lambda_3-\rho_3\dot{\eta}_3(\cos\varphi_{23}\sin\lambda_2\sin\lambda_3+\cos\lambda_2\cos\lambda_3}{c^4\tau_{23}\rho_2} - \frac{e_2^2}{m_2}\frac{\ddot{\eta}_2}{c^3} \equiv G_{\eta_2};$$

$$\dot{r}_3 = \frac{e_3e_1}{m_3}\frac{1}{c\rho_3{}^2} - \frac{e_2e_3}{m_3}\frac{\rho_3-\rho_2(\cos\varphi_{32}\cos\lambda_2\cos\lambda_3+\sin\lambda_2\sin\lambda_3)}{c^3}\frac{\left\langle \overset{\rightarrow}{\xi}^{(32)}, \overset{\cdot}{\vec{u}}^{(2)} \right\rangle}{c^3\tau_{32}{}^3}$$

$$-\frac{e_3e_2}{m_3}\frac{\dot{r}_2(\cos\varphi_{32}\cos\lambda_2\cos\lambda_3+\sin\lambda_2\sin\lambda_3)+\rho_2\dot{\phi}_2\sin\varphi_{32}\cos\lambda_2\cos\lambda_3+\rho_2\dot{\eta}_2(\sin\lambda_3\cos\lambda_2-\cos\varphi_{32}\sin\lambda_2\cos\lambda_3)}{c^4\tau_{23}} - \frac{e_3^2}{m_3}\frac{\ddot{r}_3}{c^3} \equiv G_{r_3};$$

$$\dot{\phi}_3 = \frac{e_3e_2}{m_3}\left[\frac{\rho_2\sin\varphi_{32}\cos\lambda_2}{\tau_{32}{}^3\rho_3\cos\lambda_3}\frac{\left\langle \overset{\rightarrow}{\xi}^{(32)}, \overset{\cdot}{\vec{u}}^{(2)} \right\rangle}{c^6} + \frac{\dot{r}_2\cos\lambda_2\sin\varphi_{32}+\rho_2\dot{\phi}_2\cos_2\cos\varphi_{32}-\rho_2\dot{\eta}_2\sin\lambda_2\sin\varphi_{32}}{c^4\tau_{32}\rho_3\cos\lambda_3} - \frac{\ddot{\phi}_3}{c^3}\right] \equiv G_{\phi_3};$$

$$\dot{\eta}_3 = \frac{e_3e_2}{m_3}\frac{\rho_2\cos\varphi_{32}\sin\lambda_3\cos\lambda_2}{c^3\tau_{32}{}^3}\frac{\left\langle \overset{\rightarrow}{\xi}^{(32)}, \overset{\cdot}{\vec{u}}^{(2)} \right\rangle}{c^3\rho_3}+$$

$$+\frac{e_3e_2}{m_3c^3}\frac{\dot{r}_3(\cos\varphi_{32}\cos\lambda_2\sin\lambda_3-\sin\lambda_2\cos\lambda_3)+\rho_2\dot{\phi}_2\sin\varphi_{32}\sin\lambda_3\cos\lambda_2-\rho_2\dot{\eta}_2(\cos\varphi_{23}\sin\lambda_2\sin\lambda_3+\cos\lambda_2\cos\lambda_3}{c\tau_{32}\rho_3} - \frac{e_3^2}{m_3}\frac{\ddot{\eta}_3}{c^3} \equiv G_{\eta_3}.$$

Further on, we refer to the above system as a (BS)—basic system. It is a first-order neutral system with respect to spherical velocities.

4. Introducing Function Spaces and Some Auxiliary Propositions

It is proved that for Lipschitz continuous T-periodic trajectories, each of the functional equations,

$$\tau_{23}(t) = \frac{1}{c}\sqrt{\sum_{\alpha=1}^{3}\left[x_\alpha^{(2)}(t) - x_\alpha^{(3)}(t-\tau_{23}(t))\right]^2}, \tau_{32}(t) = \frac{1}{c}\sqrt{\sum_{\alpha=1}^{3}\left[x_\alpha^{(3)}(t) - x_\alpha^{(2)}(t-\tau_{32}(t))\right]^2}$$

has a unique T-periodic solution.

By $C_T{}^{\infty}[0,\infty)$, we denote the set of all infinitely differentiable functions, satisfying the condition $r(-T) = r(0) = 0$. We introduce the following sets, where $0 < \delta < \frac{\pi}{2}$, $\mu > \omega > 0$ are strictly positive constants, $r^{(m)}$ means the m-th derivative of r and $r^{(0)}(t) = r(t)$.

Then, we introduce the following function spaces:

$$M_{r_n} = \left\{ r_n \in C_T{}^{\infty}[0,\infty) : \left| r_n{}^{(m)}(t) \right| \le \omega_n{}^m R_n e^{\mu(t-T_k)};\ r_n{}^{(0)}(T_k) = 0;\ \int_{T_k}^{T_{k+1}} r_n(s)ds = 0 \right\},\ t \in [T_k, T_{k+1}];$$

$$d_{(k,m)}(r_n, \bar{r}_n) = \sup\left\{ \frac{\left| r_n{}^{(m)}(t) - \bar{r}_n{}^{(m)}(t) \right|}{\omega_n{}^m} e^{-\mu(t-T_k)} :\ t \in [T_k, T_{k+1}] \right\};$$

$$M_{\phi_n} = \left\{ \phi_n \in C_T{}^{\infty}[0,\infty) : \left| \phi_n{}^{(m)}(t) \right| \le \omega_n{}^m \Phi_n e^{\mu(t-T_k)},\ \int_{T_k}^{T_{k+1}} \phi_n(s)ds = T\phi_{n0} \right\},\ t \in [T_k, T_{k+1}];$$

$$d_{(k,m)}(\phi_n, \bar{\phi}_n) = \sup\left\{ \frac{\left| \phi_n{}^{(m)}(t) - \bar{\phi}_n{}^{(m)}(t) \right|}{\omega_n{}^m} e^{-\mu(t-T_k)} :\ t \in [T_k, T_{k+1}] \right\};$$

$$M_{\eta_n} = \left\{ \eta_n \in C_T{}^{\infty}[0,\infty) : \left| \eta_n{}^{(m)}(t) \right| \le \omega^m Y_n e^{\mu(t-T_k)};\ \eta_n{}^{(0)}(T_k) = 0;\ \int_{T_k}^{T_{k+1}} \eta_n(s)ds = 0;\ \left| \int_{T_k}^{t} \eta(s)ds \right| \le \frac{\pi}{2} - \delta \right\},\ t \in [T_k, T_{k+1}];$$

$$d_{(k,m)}(\eta_n, \bar{\eta}_n) = \sup\left\{ \frac{\left| \eta_n{}^{(m)}(t) - \bar{\eta}_n{}^{(m)}(t) \right|}{\omega_n{}^m} e^{-\mu(t-T_k)} :\ t \in [T_k, T_{k+1}] \right\}$$

$(k = 0, 1, 2, \dots),\ \ (m = 0, 1, 2, \dots), (n = 2, 3).$

Here, ω_n, R_n, Φ_n, Y_n $(n = 2, 3)$ and μ are strictly positive constants, such that $\mu > \omega_n$.

We consider the product of uniform spaces, $M_{r_2} \times M_{\phi_2} \times M_{\eta_2} \times M_{r_3} \times M_{\phi_3} \times M_{\eta_3}$, endowed with the countable family of pseudo-metrics:

$$d_{(k,m)}((r_2, \phi_2, \eta_2, r_3, \phi_3, \eta_3), (\bar{r}_2, \bar{\phi}_2, \bar{\eta}_2, \bar{r}_3, \bar{\phi}_3, \bar{\eta}_3)) =$$
$$= d_{(k,m)}(r_2, \bar{r}_2) + d_{(k,m)}(\phi_2, \bar{\phi}_2) + d_{(k,m)}(\eta_2, \bar{\eta}_2) + d_{(k,m)}(r_3, \bar{r}_3) + d_{(k,m)}(\phi_3, \bar{\phi}_3) + d_{(k,m)}(\eta_3, \bar{\eta}_3)$$

where the index set A consists of all ordered pairs (k, m).

Lemma 1 ([20]). *If* $r_n(.) \in C_T{}^{\infty}[0,\infty)$ and $\int_0^T r_n(t)dt = 0$, *then,* $\rho_n(t) = \rho_{n0} + \int_0^t r_n(s)ds$ $(n = 2, 3)$ *is a T-periodic function. In addition,*

$$\varphi_2(t+T) = \varphi_{20} + \int_0^{t+T} \phi_2(s)ds = \varphi_{20} + \int_0^t \phi_2(s)ds + \int_{T_k}^{T_{k+1}} \phi_2(s)ds = \varphi_2(t) + \phi_{20}T = \varphi_2(t) + 2\pi \Rightarrow$$
$$\Rightarrow \cos\varphi_2(t+T) = \cos(\varphi_2(t) + 2\pi) = \cos\varphi_2(t)$$

$$\varphi_3(t+T-\tau_{23}(t+T)) = \varphi_3(t+T-\tau_{23}(t)) = \varphi_{30} + \int_0^{t+T-\tau_{23}(t)} \phi_3(s)ds = \varphi_{30} + \int_0^{t-\tau_{23}(t)} \phi_3(s)ds + \int_{t-\tau_{23}(t)}^{t-\tau_{23}(t)+T} \phi_3(s)ds =$$
$$= \varphi_3(t-\tau_{23}) + \phi_{30}T = \varphi_3(t-\tau_{23}) + 2\pi \Rightarrow$$
$$\cos\varphi_3(t+T-\tau_{23}(t+T)) = \cos(\varphi_3(t-\tau_{23}) + 2\pi) = \cos\varphi_3(t-\tau_{23})$$

It follows that the right-hand sides of the equations of (BS) are T-periodic functions.

Lemma 2. *The distances* $\rho_n(t)$, $(n = 2, 3)$ *are bounded functions.*

Proof. For sufficiently large $\mu > 0$ and $\mu T = const. > 0$, we obtain

$$\rho_n(t) \ge \rho_{n0} - \int_0^t |r_\alpha(s)|ds \ge \rho_{n0} - R_n \int_0^t e^{\mu s}ds = \rho_{n0} - R_n \frac{e^{\mu t} - 1}{\mu} \ge \rho_{n0} - \frac{R_n(e^{\mu T} - 1)}{\mu} \ge \rho_{n0} - \frac{R_n e^{\mu T}}{\mu} \equiv \widehat{\rho}_n > 0, \quad (n = 2, 3).$$

In a similar way, we infer

$$\rho_n(t) \le \rho_{n0} + \int_0^t |r_n(s)|ds \le \rho_{n0} + R_n \int_0^t e^{\mu s} ds \le \rho_{n0} + R_n \frac{e^{\mu T} - 1}{\mu} \equiv \breve{\rho}_n > 0\ , \quad (n = 2,3).$$

Finally, $\hat{\rho}_n < \breve{\rho}_n$ since $e^{\mu T} - 1 > 0$.
Lemma 2 is thus proved. □

Lemma 3. *If* $\int_0^T \phi_n(t)dt = \phi_{n0} T\ (\phi_{n0} > 0), (n = 2,3)$, *then* $\lim_{t\to\infty} \varphi_n(t) = \infty$.

Proof. Let us set $s + T = p$. Then, we obtain $\int_0^T \phi_n(s)ds = \int_T^{2T} \phi_n(p)dp = \ldots = \int_{T_k}^{T_{k+1}} \phi_n(p)dp$.
Therefore, $\lim_{t\to\infty} \varphi_n(t) = \lim_{t\to\infty} \left(\varphi_{n0} + \int_0^{\infty} \phi_n(s)ds \right) = \varphi_{n0} + \sum_{k=0}^{\infty} \int_{T_k}^{T_{k+1}} \phi_n(s)ds = \varphi_{n0} + T\phi_{n0} \sum_{k=1}^{\infty} 1 = \infty$.
Lemma 3 is thus proved. □

Lemma 4. *The sets* M_{r_n}, M_{ϕ_n}, M_{η_n} $(n = 2,3)$ *are closed.*

For the proof, cf. [19].
The following assertions give a qualitative estimate of the electron orbits.

Lemma 5. *The following inequality is valid:*
$A = |\cos\varphi_2 \cos\varphi_3 \cos\lambda_2 \cos\lambda_3 + \sin\varphi_2 \sin\varphi_3 \cos\lambda_2 \cos\lambda_3 + \sin\lambda_2 \sin\lambda_3| \le 1$.

Proof.

$$\begin{aligned} A &= |\cos\lambda_2 \cos\lambda_3 \cos(\varphi_2 - \varphi_3) + \sin\lambda_2 \sin\lambda_3| = \\ &= \left| \tfrac{\cos(\lambda_2-\lambda_3)+\cos(\lambda_2+\lambda_3)}{2} \cos(\varphi_2 - \varphi_3) + \tfrac{\cos(\lambda_2-\lambda_3)-\cos(\lambda_2+\lambda_3)}{2} \right| = \\ &= \left| \tfrac{\cos(\lambda_2-\lambda_3)[1+\cos(\varphi_2-\varphi_3)]-\cos(\lambda_2+\lambda_3)[1-\cos(\varphi_2-\varphi_3)]}{2} \right| = \\ &= \tfrac{1}{2} \left| 2\cos^2 \tfrac{\varphi_2-\varphi_3}{2} \cos(\lambda_2 - \lambda_3) - 2\sin^2 \tfrac{\varphi_2-\varphi_3}{2} \cos(\lambda_2 + \lambda_3) \right| \le \cos^2 \tfrac{\varphi_2-\varphi_3}{2} + \sin^2 \tfrac{\varphi_2-\varphi_3}{2} = 1. \end{aligned}$$

□

Lemma 6. *The following inequalities are valid:*
$|\rho_{20} - \rho_{30}| - (R_2 + R_3)\frac{e^{\mu T}-1}{\mu} \le \rho_{23}(t), \rho_{32}(t) \le \rho_{20} + \rho_{30} + (R_2 + R_3)\frac{e^{\mu T}-1}{\mu}$.

Proof. Since $\rho_2(t) = \rho_{20} + \int_0^t r_2(s)ds$, $\rho_3(t) = \rho_{30} + \int_0^t r_3(s)ds$, then, in view of Lemma 2,

$$\rho_{23}(t) = \sqrt{(\rho_2 \cos\varphi_2 \cos\lambda_2 - \rho_3 \cos\varphi_3 \cos\lambda_3)^2 + (\rho_2 \sin\varphi_2 \cos\lambda_2 - \rho_3 \sin\varphi_3 \cos\lambda_3)^2 + (\rho_2 \sin\lambda_2 - \rho_3 \sin\lambda_3)^2} \le$$
$$\le \sqrt{\rho_2{}^2 + \rho_3{}^2 + 2\rho_2\rho_3 A} \le \sqrt{\rho_2{}^2 + \rho_3{}^2 + 2\rho_2\rho_3} = \rho_2(t) + \rho_3(t) \le \rho_{20} + \rho_{30} + (R_2 + R_3)\frac{e^{\mu T}-1}{\mu}$$

and

$$\rho_{23}(t) = \sqrt{(\rho_2 \cos\varphi_2 \cos\lambda_2 - \rho_3 \cos\varphi_3 \cos\lambda_3)^2 + (\rho_2 \sin\varphi_2 \cos\lambda_2 - \rho_3 \sin\varphi_3 \cos\lambda_3)^2 + (\rho_2 \sin\lambda_2 - \rho_3 \sin\lambda_3)^2} \ge$$
$$\ge \sqrt{\rho_2{}^2 + \rho_3{}^2 - 2\rho_2\rho_3 A} \ge \sqrt{\rho_2{}^2 + \rho_3{}^2 - 2\rho_2\rho_3} = |\rho_2 - \rho_3| \ge |\rho_{20} - \rho_{30}| - (R_2 + R_3)\frac{e^{\mu T}-1}{\mu}$$

which completes the proof. □

Lemma 7. *The following inequalities are valid:*

$$\tau_{23} \geq \frac{|\rho_2(t)-\rho_3(t-\tau_{23})|}{c} = \frac{\left|\rho_{02}-\rho_{03}+\int\limits_0^t r_2(s)ds-\int\limits_0^t r_3(s)ds-\int\limits_t^{t-\tau_{23}} r_3(s)ds\right|}{c} \geq \frac{|\rho_{02}-\rho_{03}|-(R_2+R_3)\int\limits_0^t e^{\mu s}ds-\left|\int\limits_t^{t-\tau_{23}} r_3(s)ds\right|}{c} \geq$$

$$\geq \frac{|\rho_{02}-\rho_{03}|-(R_2+R_3)\frac{e^{\mu T}-1}{\mu}-R_3\left|\frac{e^{\mu(t-\tau_{23})}-e^{\mu t}}{\mu}\right|}{c} \geq \frac{|\rho_{02}-\rho_{03}|-(R_2+R_3)\frac{e^{\mu T}-1}{\mu}-R_3e^{\mu t}\left|\frac{e^{-\mu\tau_{23}}-1}{\mu}\right|}{c} \geq$$

$$\geq \frac{|\rho_{02}-\rho_{03}|-R_2\frac{e^{\mu T}-1}{\mu}-R_3\frac{e^{\mu T}-1}{\mu}-R_3\frac{e^{\mu T}}{\mu}}{c} = \frac{|\rho_{02}-\rho_{03}|-R_2\frac{e^{\mu T}-1}{\mu}-R_3\frac{e^{\mu T}}{\mu}}{c} \geq \frac{|\rho_{02}-\rho_{03}|-(R_2+R_3)\frac{e^{\mu T}}{\mu}}{c} \equiv \frac{\Delta}{c} > 0$$

and

$$\tau_{32} \geq \frac{|\rho_3(t)-\rho_2(t-\tau_{32})|}{c} = \frac{\left|\rho_{03}-\rho_{02}+\int\limits_0^t r_3(s)ds-\int\limits_0^t r_2(s)ds-\int\limits_t^{t-\tau_{32}} r_2(s)ds\right|}{c} \geq \frac{|\rho_{03}-\rho_{02}|-(R_2+R_3)\int\limits_0^t e^{\mu s}ds-\left|\int\limits_t^{t-\tau_{32}} r_2(s)ds\right|}{c} \geq$$

$$\geq \frac{|\rho_{03}-\rho_{02}|-(R_3+R_2)\frac{e^{\mu T}-1}{\mu}-R_2\left|\frac{e^{\mu(t-\tau_{32})}-e^{\mu t}}{\mu}\right|}{c} \geq \frac{|\rho_{03}-\rho_{02}|-(R_3+R_2)\frac{e^{\mu T}-1}{\mu}-R_2e^{\mu t}\left|\frac{e^{-\mu\tau_{32}}-1}{\mu}\right|}{c} \geq$$

$$\geq \frac{|\rho_{03}-\rho_{02}|-R_3\frac{e^{\mu T}-1}{\mu}-R_2\frac{e^{\mu T}-1}{\mu}-R_2\frac{e^{\mu T}}{\mu}}{c} = \frac{|\rho_{03}-\rho_{02}|-R_3\frac{e^{\mu T}-1}{\mu}-R_2\frac{e^{\mu T}}{\mu}}{c} \geq \frac{|\rho_{02}-\rho_{03}|-(R_2+R_3)\frac{e^{\mu T}}{\mu}}{c} \equiv \frac{\Delta}{c} > 0$$

for sufficiently large $\mu > 0$ *and* $\mu T = const. > 0$. *The above inequalities imply that* $|\rho_2(t)-\rho_3(t-\tau_{23})|$, $|\rho_3(t)-\rho_2(t-\tau_{32})|$ *are strictly positive, which completes the proof.*

Remark 3. *Clearly,* $\frac{1}{\tau_{23}} \leq \frac{c}{\Delta}$, $\frac{1}{\tau_{32}} \leq \frac{c}{\Delta}$.

5. Introducing Operators and the Main Lemma

Prior to the formulation and proving of the Main Lemma, we give the compatibility condition (CC), which implies the continuity of solution for neutral equations: the initial functions satisfy the following relations:

$$\dot{r}_{20}(0) = \frac{e_2e_1}{m_2}\frac{1}{c\rho_{20}{}^2(0)} -$$

$$-\frac{e_2e_3}{m_2}\frac{\rho_{20}(0)-\rho_{30}(-\tau_{23}(0))(\cos\varphi_{23}\cos\lambda_{20}(0)\cos\lambda_{30}(-\tau_{23}(0))+\sin\lambda_{20}(0)\sin\lambda_{30}(-\tau_{23}(0)))}{c^3}\frac{\left\langle \vec{\xi}^{(23)}, \dot{\vec{u}}^{(3)} \right\rangle}{c^3\tau_{23}{}^3}$$

$$-\frac{e_2e_3}{m_2}\frac{\dot{r}_{30}(-\tau_{23}(0))(\cos\varphi_{23}\cos\lambda_{20}(0)\cos\lambda_{30}(-\tau_{23}(0))+\sin\lambda_{20}(0)\sin\lambda_{30}(-\tau_{23}(0)))}{c^4\tau_{23}}$$

$$-\frac{e_2e_3}{m_2}\frac{\rho_{30}(-\tau_{23}(0))\dot{\phi}_{30}(-\tau_{23}(0))\sin\varphi_{23}\cos\lambda_{20}(0)\cos\lambda_{30}(-\tau_{23}(0))}{c^4\tau_{23}}$$

$$-\frac{e_2e_3}{m_2}\frac{\rho_{30}(-\tau_{23}(0))\dot{\eta}_{30}(-\tau_{23}(0))(\sin\lambda_{20}(0)\cos\lambda_{30}(-\tau_{23}(0))-\cos\varphi_{23}\sin\lambda_{30}(-\tau_{23}(0))\cos\lambda_{20}(0))}{c^4\tau_{23}} - \frac{e_2^2}{m_2}\frac{\ddot{r}_{20}(0)}{c^3} \equiv G_{r_2}(0);$$

where $\varphi_{23} = \varphi_2(0) - \varphi_3(-\tau_{23}(0))$ and so on.

Define an operator B acting on $M_{r_2} \times M_{\phi_2} \times M_{\eta_2} \times M_{r_3} \times M_{\phi_3} \times M_{\eta_3}$ by the formulas

$$B_{r_2}^{(k)}(r_2,\phi_2,\eta_2,r_3,\phi_3,\eta_3)(t) := \begin{cases} \int_{T_k}^{t} G_{r_2}(s)ds - \left(\frac{t-T_k}{T}-\frac{1}{2}\right)\int_{T_k}^{T_{k+1}} G_{r_2}(s)ds - \frac{1}{T}\int_{T_k}^{T_{k+1}}\int_{T_k}^{p} G_{r_2}(s)dsdp,\ t\in[T_k,T_{k+1}] \\ r_3(t)=r_{30}(t),\phi_3(t)=\phi_{30}(t),\eta_3(t)=\eta_{30}(t),\ t\in[-T,0] \end{cases};$$

$$B_{\phi_2}^{(k)}(r_2,\phi_2,\eta_2,r_3,\phi_3,\eta_3)(t) := \begin{cases} \phi_{20}+\int_{T_k}^{t} G_{\phi_2}(s)ds - \left(\frac{t-T_k}{T}-\frac{1}{2}\right)\int_{T_k}^{T_{k+1}} G_{\phi_2}(s)ds - \frac{1}{T}\int_{T_k}^{T_{k+1}}\int_{T_k}^{p} G_{\phi_2}(s)dsdp,\ t\in[T_k,T_{k+1}] \\ r_3(t)=r_{30}(t),\phi_3(t)=\phi_{30}(t),\eta_3(t)=\eta_{30}(t),\ t\in[-T,0] \end{cases};$$

$$B_{\eta_2}^{(k)}(r_2,\phi_2,\eta_2,r_3,\phi_3,\eta_3)(t) := \begin{cases} \int_{T_k}^{t} G_{r_3}(s)ds - \left(\frac{t-T_k}{T}-\frac{1}{2}\right)\int_{T_k}^{T_{k+1}} G_{r_3}(s)ds - \frac{1}{T}\int_{T_k}^{T_{k+1}}\int_{T_k}^{p} G_{r_3}(s)dsdp\,,t\in[T_k,T_{k+1}] \\ r_3(t)=r_{30}(t),\phi_3(t)=\phi_{30}(t),\eta_3(t)=\eta_{30}(t),\ t\in[-T,0] \end{cases};$$

$$B_{r_3}^{(k)}(r_2,\phi_2,\eta_2,r_3,\phi_3,\eta_3)(t) := \begin{cases} \int_{T_k}^{t} G_{r_3}(s)ds - \left(\frac{t-T_k}{T}-\frac{1}{2}\right)\int_{T_k}^{T_{k+1}} G_{r_3}(s)ds - \frac{1}{T}\int_{T_k}^{T_{k+1}}\int_{T_k}^{p} G_{r_3}(s)dsdp,\ t\in[T_k,T_{k+1}] \\ r_2(t)=r_{20}(t),\phi_2(t)=\phi_{20}(t),\eta_2(t)=\eta_{20}(t),\ t\in[-T,0] \end{cases};$$

$$B_{\phi_3}^{(k)}(r_2,\phi_2,\eta_2,r_3,\phi_3,\eta_3)(t) := \begin{cases} \phi_{30}+\int_{T_k}^{t} G_{\phi_3}(s)ds - \left(\frac{t-T_k}{T}-\frac{1}{2}\right)\int_{T_k}^{T_{k+1}} G_{\phi_3}(s)ds - \frac{1}{T}\int_{T_k}^{T_{k+1}}\int_{T_k}^{p} G_{\phi_3}(s)dsdp, \\ r_2(t)=r_{20}(t),\phi_2(t)=\phi_{20}(t),\eta_2(t)=\eta_{20}(t),\ t\in[-T,0] \end{cases} t\in[T_k,T_{k+1}];$$

$$B_{\eta_3}^{(k)}(r_2,\phi_2,\eta_2,r_3,\phi_3,\eta_3)(t) := \begin{cases} \int_{T_k}^{t} G_{\eta_3}(s)ds - \left(\frac{t-T_k}{T}-\frac{1}{2}\right)\int_{T_k}^{T_{k+1}} G_{\eta_3}(s)ds - \frac{1}{T}\int_{T_k}^{T_{k+1}}\int_{T_k}^{p} G_{\eta_3}(s)dsdp,t\in[T_k,T_{k+1}] \\ r_2(t)=r_{20}(t),\phi_2(t)=\phi_{20}(t),\eta_2(t)=\eta_{20}(t),\ t\in[-T,0] \end{cases}$$

where $t\in[T_k,T_{k+1}]$, $T_k=kT$ $(k=0,1,2,\dots)$, $G_{\phi_2}=G_{\phi_2}(r_2,\phi_2,\eta_2,r_3,\phi_3,\eta_3)$, $G_{\eta_2}=G_{\eta_2}(r_2,\phi_2,\eta_2,r_3,\phi_3,\eta_3)$, $G_{r_3}=G_{r_3}(r_2,\phi_2,\eta_2,r_3,\phi_3,\eta_3)$, $G_{\phi_3}=G_{\phi_3}(r_2,\phi_2,\eta_2,r_3,\phi_3,\eta_3)$, $G_{\eta_3}=G_{\eta_3}(r_2,\phi_2,\eta_2,r_3,\phi_3,\eta_3)$ and the initial functions $r_{n0}(t),\phi_{n0}(t),\eta_{n0}(t),(n=2,3)$ are solutions of the system of initial equations.

Remark 4. *We notice that in view of the periodicity of $G_{r_2}(s)$, the operator $B_{r_2}(.)(t)$ is T-periodic on $[0,\infty)$ (cf. [20]). The conformity condition (CC) $\dot{r}_{20}(0)=G_{r_2}(0)$ implies*

$$\dot{r}_2(t)=G_{r_2}(t)-\frac{1}{T}\int_0^T G_{r_2}(s)ds \Leftrightarrow \dot{r}_{20}(0)=G_{r_2}(0)-\frac{1}{T}\int_0^T G_{r_2}(s)ds \Leftrightarrow \int_0^T G_{r_2}(s)ds=0$$

and

$$r_2(t)=B_{r_2}^{(0)}(.)(t)\Rightarrow r_2(0)=B_{r_2}^{(0)}(.)(0)\Leftrightarrow 0=\frac{1}{2}\int_0^T G_{r_2}(s)ds-\frac{1}{T}\int_0^T\int_0^p G_{r_2}(s)dsdp\Rightarrow \int_0^T\int_0^p G_{r_2}(s)dsdp=0$$

$$B_{r_2}^{(0)}(.)(0)=\frac{1}{2}\int_0^T G_{r_2}(s)ds-\frac{1}{T}\int_0^T\int_0^p G_{r_2}(s)dsdp=0;\ B_{r_2}^{(0)}(.)(T)=\frac{1}{2}\int_0^T G_{r_2}(s)ds-\frac{1}{T}\int_0^T\int_0^p G_{r_2}(s)dsdp=0.$$

But $G_{r_2}(s)$ is a T-periodic function. Therefore, one can extend the result for every $k=1,2,\dots$, that is, $\int_{T_k}^{T_{k+1}} G_{r_2}(s)ds=\int_{T_k}^{T_{k+1}}\int_{T_k}^{p} G_{r_2}(s)dsdp=0$.

We need the following inequalities:

$$\left\langle \overset{\rightarrow}{\xi}^{(23)}, \dot{\vec{u}}^{(3)} \right\rangle \le \sqrt{\left\langle \overset{\rightarrow}{\xi}^{(23)}, \overset{\rightarrow}{\xi}^{(23)} \right\rangle} \sqrt{\left\langle \dot{\vec{u}}^{(3)}, \dot{\vec{u}}^{(3)} \right\rangle} \le c\tau_{23}\sqrt{\ddot{\rho}_3{}^2 + \rho_3{}^2\ddot{\varphi}_3{}^2\cos^2\lambda_3 + \rho_3{}^2\ddot{\lambda}_3{}^2} =$$

$$= c\tau_{23}\sqrt{\dot{r}_3{}^2 + \rho_3{}^2\dot{\phi}_3{}^2\cos^2\lambda_3 + \rho_3{}^2\dot{\eta}_3{}^2} \le c\tau_{23}\omega_3 e^{\mu(t-T_k)}\sqrt{R_3{}^2 + \rho_3{}^2R_3{}^2 + \rho_3{}^2Y_3{}^2} \le \bar{c}c\tau_{23}\omega_3 e^{\mu(t-T_k)}$$

$$\left\langle \overset{\rightarrow}{\xi}^{(32)}, \dot{\vec{u}}^{(2)} \right\rangle \le \sqrt{\left\langle \overset{\rightarrow}{\xi}^{(32)}, \overset{\rightarrow}{\xi}^{(32)} \right\rangle} \sqrt{\left\langle \dot{\vec{u}}^{(2)}, \dot{\vec{u}}^{(2)} \right\rangle} \le c\tau_{32}\sqrt{\ddot{\rho}_2{}^2 + \rho_2{}^2\ddot{\varphi}_2{}^2\cos^2\lambda_2 + \rho_2{}^2\ddot{\lambda}_2{}^2} =$$

$$= c\tau_{23}\sqrt{\dot{r}_3{}^2 + \rho_3{}^2\dot{\phi}_3{}^2\cos^2\lambda_3 + \rho_3{}^2\dot{\eta}_3{}^2} \le c\tau_{32}\omega_2 e^{\mu(t-T_k)}\sqrt{R_2{}^2 + \rho_2{}^2R_2{}^2 + \rho_2{}^2Y_2{}^2} \le \bar{c}c\tau_{32}\omega_2 e^{\mu(t-T_k)}$$

In addition, the condition (C) implies $\left\langle \vec{u}^{(n)}, \vec{u}^{(n)} \right\rangle = r_n{}^2 + \rho_n{}^2\phi_n{}^2\cos^2\lambda_n + \rho_n{}^2\eta_n{}^2 \le R_n{}^2 + \rho_n{}^2\Phi_n{}^2 + \rho_n{}^2Y_n{}^2 \le \bar{c}^2 < c^2, (n = 2,3).$

Lemma 8. *The following inequalities are valid:*

$$\left|\int_{T_k}^{t} G_{r_2}(s)ds\right| \le \left|\left(\int_{T_k}^{t} \frac{|e_2e_1|}{m_2c}\frac{1}{\widehat{\rho}_2{}^2} + \frac{|e_2e_3|}{m_2}\frac{\bar{c}c\tau_{23}\omega_3 e^{\mu(t-T_k)}}{c^6}\frac{\rho_2+\rho_3}{\tau_{23}{}^3} + \frac{|e_2e_3|}{m_2}\frac{|\dot{r}_3|+|\dot{\phi}_3\rho_3|+|\dot{\eta}_3\rho_3|}{c^4\tau_{23}} + \frac{e_2^2}{m_2c^3}|\ddot{r}_2|\right)ds\right| \le$$

$$\le \frac{|e_2e_1|}{m_2}\left|\int_{T_k}^{t}\left(\frac{1}{c\widehat{\rho}_2{}^2} + \frac{\bar{c}\omega_3(\breve{\rho}_2+\breve{\rho}_3+3\Delta)}{c^3\Delta^2} + \frac{\omega_2{}^2R_2}{c^3}\right)e^{\mu(s-T_k)}ds\right| \le \frac{|e_2e_1|}{m_2}\left(\frac{1}{c\widehat{\rho}_2{}^2} + \frac{\bar{c}\omega_3(\breve{\rho}_2+\breve{\rho}_3+3\Delta)}{c^3\Delta^2} + \frac{\omega_2{}^2R_2}{c^3}\right)\frac{e^{\mu(t-T_k)}}{\mu};$$

$$\left|\int_{T_k}^{T_{k+1}} G_{r_2}(s)ds\right| \le \frac{|e_2e_1|}{m_2}\left(\frac{1}{c\widehat{\rho}_2{}^2} + \frac{\bar{c}\omega_3(\breve{\rho}_2+\breve{\rho}_3+3\Delta)}{c^2\Delta^2} + \frac{\omega_2{}^2R_2}{c^3}\right)\frac{e^{\mu T}-1}{\mu};$$

$$\left|\int_{T_k}^{t} G_{\phi_2}ds\right| \le \frac{|e_2e_3|}{m_2}\int_{T_k}^{t}\left(\frac{\rho_3\bar{c}\omega_3}{c^5\tau_{23}{}^2\rho_2\cos\lambda_2} + \frac{|\dot{r}_3|+|\rho_3\dot{\phi}_3|+|\rho_3\dot{\eta}_3|}{c^4\tau_{23}\rho_2\cos\lambda_2} + \frac{|\ddot{\phi}_2|}{c^3}\right)e^{\mu(s-T_k)}ds \le \frac{|e_2e_3|}{m_2}\left(\frac{\bar{c}\omega_3(\breve{\rho}_3+3\Delta)}{\widehat{\rho}_2c^3\Delta^2\cos(\pi/2-\delta)} + \frac{\omega_2{}^2\Phi_2}{c^3}\right)\frac{e^{\mu(t-T_k)}}{\mu};$$

$$\left|\int_{T_k}^{T_{k+1}} G_{\phi_2}ds\right| \le \frac{|e_2e_3|}{m_2}\left(\frac{\bar{c}\omega_3(\breve{\rho}_3+3\Delta)}{\widehat{\rho}_2c^3\Delta^2\cos(\pi/2-\delta)} + \frac{\omega_2{}^2\Phi_2}{c^3}\right)\frac{e^{\mu T}-1}{\mu};$$

$$\left|\int_{T_k}^{t} G_{\eta_2}ds\right| \le \left|\frac{e_2e_3}{m_2}\right|\int_{T_k}^{t}\left(\frac{\rho_3\bar{c}\omega_3 e^{\mu(s-T_k)}}{c^5\tau_{23}{}^2\rho_2} + \frac{|\dot{r}_3|+|\rho_3\dot{\phi}_3|+|\rho_3\dot{\eta}_3|}{c^4\tau_{23}\rho_2} + \frac{\omega_2{}^2Y_2}{c^3}\right)e^{\mu(s-T_k)}ds \le \frac{|e_2e_3|}{m_2}\left(\frac{\bar{c}\omega_3}{c^3}\frac{\breve{\rho}_3e^{\mu T}+3\Delta}{\widehat{\rho}_2\Delta^2} + \frac{\omega_2{}^2Y_2}{c^3}\right)\frac{e^{\mu(t-T_k)}}{\mu};$$

$$\left|\int_{T_k}^{T_{k+1}} G_{\eta_2}ds\right| \le \frac{|e_2e_3|}{m_2}\left(\frac{\bar{c}\omega_3}{c^3}\frac{\breve{\rho}_3e^{\mu T}+3\Delta}{\widehat{\rho}_2\Delta^2} + \frac{\omega_2{}^2Y_2}{c^3}\right)\frac{e^{\mu T}-1}{\mu};$$

$$\left|\int_{T_k}^{t} G_{r_3}(s)ds\right| \le \frac{|e_2e_1|}{m_3}\left(\frac{1}{c\widehat{\rho}_3{}^2} + \frac{\bar{c}\omega_2(\breve{\rho}_2+\breve{\rho}_3+3\Delta)}{c^2\Delta^2} + \frac{\omega_3{}^2R_3}{c^3}\right)\frac{e^{\mu(t-T_k)}}{\mu};$$

$$\left|\int_{T_k}^{T_{k+1}} G_{r_3}(s)ds\right| \le \frac{|e_2e_1|}{m_3}\left(\frac{1}{c\widehat{\rho}_3{}^2} + \frac{\bar{c}\omega_2(\breve{\rho}_2+\breve{\rho}_3+3\Delta)}{c^2\Delta^2} + \frac{\omega_3{}^2R_3}{c^3}\right)\frac{e^{\mu T}-1}{\mu};$$

$$\left|\int_{T_k}^{t} G_{\phi_3}ds\right| \le \frac{|e_2e_3|}{m_3}\left(\frac{\bar{c}\omega_2(\breve{\rho}_2+3\Delta)}{\widehat{\rho}_3c^3\Delta^2\cos(\pi/2-\delta)} + \frac{\omega_3{}^2\Phi_3}{c^3}\right)\frac{e^{\mu(t-T_k)}}{\mu};$$

$$\left|\int_{T_k}^{T_{k+1}} G_{\phi_3}ds\right| \le \frac{|e_2e_3|}{m_3}\left(\frac{\bar{c}\omega_2(\breve{\rho}_2+3\Delta)}{\widehat{\rho}_3c^3\Delta^2\cos(\pi/2-\delta)} + \frac{\omega_3{}^2\Phi_3}{c^3}\right)\frac{e^{\mu T}-1}{\mu};\left|\int_{T_k}^{t} G_{\eta_3}ds\right| \le \frac{|e_2e_3|}{m_3}\left(\frac{\bar{c}\omega_2}{c^3}\frac{\breve{\rho}_2e^{\mu T}+3\Delta}{\widehat{\rho}_3\Delta^2} + \frac{\omega_3{}^2Y_3}{c^3}\right)\frac{e^{\mu(t-T_k)}}{\mu};$$

$$\left|\int_{T_k}^{T_{k+1}} G_{\eta_3}ds\right| \le \frac{|e_2e_3|}{m_3}\left(\frac{\bar{c}\omega_2}{c^3}\frac{\breve{\rho}_2e^{\mu T}+3\Delta}{\widehat{\rho}_3\Delta^2} + \frac{\omega_3{}^2Y_3}{c^3}\right)\frac{e^{\mu T}-1}{\mu}.$$

Lemma 9 (Main Lemma). *Let the compatibility condition (CC) be satisfied. The periodic problem for the system (BS) has a solution* $(r_2, \phi_2, \eta_2, r_3, \phi_3, \eta_3) \in M_{r_2} \times M_{\phi_2} \times M_{\eta_2} \times M_{r_3} \times M_{\phi_3} \times M_{\eta_3}$ *if the operator B has a fixed point belonging to* $M_{r_2} \times M_{\phi_2} \times M_{\eta_2} \times M_{r_3} \times M_{\phi_3} \times M_{\eta_3}$.

Proof. Let the system (BS) have a solution belonging to $M_{r_2} \times M_{\phi_2} \times M_{\eta_2} \times M_{r_3} \times M_{\phi_3} \times M_{\eta_3}$. □

We show that $B = \left(B_{r_2}, B_{\phi_2}, B_{\eta_2}, B_{r_3}, B_{\phi_3}, B_{\eta_3}\right)$ possesses a fixed point belonging to $M_{r_2} \times M_{\phi_2} \times M_{\eta_2} \times M_{r_3} \times M_{\phi_3} \times M_{\eta_3}$. Indeed, integrating the first equation of (BS), we obtain $r_2(t) = \int\limits_{T_k}^{t} G_{r_2}(s)ds$, and therefore, in view of Remark 4, one obtains

$$r_2(t) = \int\limits_{T_k}^{t} G_{r_2}(s)ds - \left(\frac{t-T_k}{T} - \frac{1}{2}\right) \int\limits_{T_k}^{T_{k+1}} G_{r_2}(s)ds - \frac{1}{T} \int\limits_{T_k}^{T_{k+1}} \int\limits_{T_k}^{p} G_{r_2}(s)dsdp,$$

that is, $r_2 = B_{r_2}{}^{(k)}(r_2, \phi_2, \eta_2, r_3, \phi_3, \eta_3)$.

For the second component ϕ_2, we have $\int\limits_{T_k}^{T_{k+1}} G_{\phi_2}(s)ds = 0$. Therefore,

$$\phi_2(T_{k+1}) = \phi_2(T_k) + \int\limits_{T_k}^{T_{k+1}} G_{\phi_2}(s)ds \Leftrightarrow \phi_{02} = \phi_{02} + \int\limits_{T_k}^{T_{k+1}} G_{\phi_2}(s)ds \Rightarrow \int\limits_{T_k}^{T_{k+1}} G_{\phi_2}(s)ds = 0.$$

Therefore, $\phi_2(t) = \phi_{02} + \int\limits_{T_k}^{t} G_{\phi_2}(s)ds - \left(\frac{t-T_k}{T} - \frac{1}{2}\right) \int\limits_{T_k}^{T_{k+1}} G_{\phi_2}(s)ds$. In the same way,

$$\int\limits_{T_k}^{T_{k+1}} \int\limits_{T_k}^{p} G_{\phi_2}(s)dsdp = \int\limits_{T_k}^{T_{k+1}} \int\limits_{s}^{T_{k+1}} G_{\phi_2}(s)dpds = \int\limits_{T_k}^{T_{k+1}} (T_{k+1}-s)G_{\phi_2}(s)ds = T_{k+1} \int\limits_{T_k}^{T_{k+1}} G_{\phi_2}(s)ds - \int\limits_{T_k}^{T_{k+1}} sG_{\phi_2}(s)ds = -\int\limits_{T_k}^{T_{k+1}} sG_{\phi_2}(s)ds =$$
$$= -\int\limits_{T_k}^{T_{k+1}} s\frac{d\phi_2(s)}{ds}ds = -\int\limits_{T_k}^{T_{k+1}} sd\phi_2(s) = -T_{k+1}\phi_2(T_{k+1}) + T_k\,\phi_2(T_k) + \int\limits_{T_k}^{T_{k+1}} \phi_2(s)ds = -T\phi_{02} + T\,\phi_{02}(T_k) + \int\limits_{T_k}^{T_{k+1}} \phi_2(s)ds = -T\phi_{02} + T\phi_{02} = 0.$$

Therefore,

$$\phi_2(t) = \phi_{02} + \int\limits_{T_k}^{t} G_{\phi_2}(s)ds - \left(\frac{t-T_k}{T} - \frac{1}{2}\right) \int\limits_{T_k}^{T_{k+1}} G_{\phi_2}(s)ds - \frac{1}{T} \int\limits_{T_k}^{T_{k+1}} \int\limits_{T_k}^{p} G_{\phi_2}(s)dsdp$$

which means $\phi_2 = B_{\phi_2}(r_2, \phi_2, \eta_2, r_3, \phi_3, \eta_3)$.

Repeating the same reasoning, we obtain

$$\eta_2 = B_{\eta_2}(r_2, \phi_2, \eta_2, r_3, \phi_3, \eta_3),\ r_3 = B_{r_3}(r_2, \phi_2, \eta_2, r_3, \phi_3, \eta_3),$$
$$\phi_3 = B_{\phi_3}(r_2, \phi_2, \eta_2, r_3, \phi_3, \eta_3),\ \ \eta_3 = B_{\eta_3}(r_2, \phi_2, \eta_2, r_3, \phi_3, \eta_3)).$$

Conversely, let $(r_2, \phi_2, \eta_2, r_3, \phi_3, \eta_3) \in M_{r_2} \times M_{\phi_2} \times M_{\eta_2} \times M_{r_3} \times M_{\phi_3} \times M_{\eta_3}$ be a fixed point of B. Then,

$$r_2(t) = \int\limits_{T_k}^{t} G_{r_2}(s)ds - \left(\frac{t-T_k}{T} - \frac{1}{2}\right) \int\limits_{T_k}^{T_{k+1}} G_{r_2}(s)ds - \frac{1}{T} \int\limits_{T_k}^{T_{k+1}} \int\limits_{T_k}^{s} G_{r_2}(\theta)d\theta ds.$$

gives for $t = T_k$

$$r_2(T_k) = \int\limits_{T_k}^{T_k} G_{r_2}(s)ds - \left(\frac{T_k-T_k}{T} - \frac{1}{2}\right) \int\limits_{T_k}^{T_{k+1}} G_{r_2}(s)ds - \frac{1}{T} \int\limits_{T_k}^{T_{k+1}} \int\limits_{T_k}^{s} G_{r_2}(\theta)d\theta ds \Rightarrow 0 = \frac{1}{2} \int\limits_{T_k}^{T_{k+1}} G_{r_2}(s)ds - \frac{1}{T} \int\limits_{T_k}^{T_{k+1}} \int\limits_{T_k}^{s} G_{r_2}(\theta)d\theta ds.$$

But Remark 4 implies $\int\limits_{T_k}^{T_{k+1}} G_{r_2}(s)ds = 0 \Rightarrow \frac{1}{T} \int\limits_{T_k}^{T_{k+1}} \int\limits_{T_k}^{s} G_{r_2}(\theta)d\theta ds = 0$.

Consequently, $r_2(t) = \int\limits_{T_k}^{t} G_{r_2}(s)ds \Rightarrow \dot{r}_2(t) = G_{r_2}(t)$.

For the remaining components of B, we proceed in the same way.

Lemma 9 is thus proved.

6. The Main Result

The main result of the paper is as follows:

Theorem 1 (Main Theorem). *Let the compatibility condition (CC) be satisfied, and initial data be sufficiently small (Appendix C). If the following inequalities are satisfied*:

(1)
$$\frac{|e_2e_1|}{m_2}\left(\frac{1}{c\widehat{\rho}_2{}^2}+\frac{\bar{c}\omega_3(\breve{\rho}_2+\breve{\rho}_3+3|\Delta|)}{c^3\Delta^2}+\frac{\omega_2{}^2R_2}{c^3}\right)\frac{e^{\mu T}}{\mu}\le R_2\ (\mu>\omega_2,\omega_3);$$
$$\phi_{20}+\frac{|e_2e_1|}{m_2}\left(\frac{\bar{c}\omega_3(\breve{\rho}_3+3|\Delta|)}{\widehat{\rho}_2c^3\Delta^2\cos(\pi/2-\delta)}+\frac{\omega_2{}^2\Phi_2}{c^3}\right)\frac{e^{\mu T}}{\mu}\le\Phi_2\ (\phi_{20}>0);$$
$$\frac{|e_2e_1|}{m_2}\left(\frac{\bar{c}\omega_3}{c^3}\frac{\breve{\rho}_3e^{\mu T}+3|\Delta|}{\widehat{\rho}_2\Delta^2}+\frac{\omega_2{}^2Y_2}{c^3}\right)\frac{e^{\mu T}}{\mu}\le Y_2;$$
$$\frac{|e_3e_1|}{m_3}\left(\frac{1}{c\widehat{\rho}_3{}^2}+\frac{\bar{c}\omega_2(\breve{\rho}_2+\breve{\rho}_3+3)}{c^3\Delta^2}+\frac{\omega_3{}^2R_3}{c^3}\right)\frac{e^{\mu T}}{\mu}\le R_3;$$
$$\phi_{30}+\frac{|e_3e_1|}{m_3c^3}\left[\frac{\beta\omega_2}{\cos(\pi/2-\delta)}\frac{\breve{\rho}_2+3c\Delta}{\widehat{\rho}_3c\Delta^2}+\omega_3{}^2\Phi_3\right]\frac{e^{\mu T}}{\mu}\le\Phi_3\ (\phi_{30}>0);$$
$$\frac{|e_3e_1|}{m_3c^3}\left(\frac{\bar{c}\omega_2}{c^3}\frac{\breve{\rho}_2e^{\mu T}+3\Delta}{\widehat{\rho}_3\Delta^2}+\frac{\omega_3{}^2Y_3}{c^3}\right)\frac{e^{\mu T}}{\mu}\le Y_3;$$

(2)
$$\left(1+\frac{e^{\mu T}-1}{\mu T}\right)\frac{|e_2e_1|}{m_2}\left(\frac{1}{c\widehat{\rho}_2{}^2}+\frac{\bar{c}\omega_3(\breve{\rho}_2+\breve{\rho}_3+3|\Delta|)}{c^3\Delta^2}+\frac{\omega_2{}^2R_2}{c^3}\right)\frac{e^{\mu T}}{\mu}\le\omega_2R_2;$$
$$\left(1+\frac{e^{\mu T}-1}{\mu T}\right)\frac{|e_2e_1|}{m_2c^3}\left[\frac{\bar{c}\omega_3\left(\breve{\rho}_3+3|\Delta|\right)}{c^2\widehat{\rho}_2\cos(\pi/2-\delta)\Delta^2}+\omega_2{}^2\Phi_2\right]\le\omega_2\Phi_2;$$
$$\left(1+\frac{e^{\mu T}-1}{\mu T}\right)\frac{|e_2e_1|}{m_2c^3}\left[\frac{\bar{c}\omega_3\left(\breve{\rho}_3+3|\Delta|\right)}{c^2\widehat{\rho}_2\Delta^2}+\omega_2{}^2Y_2\right]\le\omega_2Y_2;$$
$$\left(1+\frac{e^{\mu T}-1}{\mu T}\right)\frac{|e_3e_1|}{m_3c}\left[\frac{1}{\widehat{\rho}_3{}^2}+\frac{\bar{c}\omega_2\left(\breve{\rho}_2+\breve{\rho}_3+3|\Delta|\right)}{c^3\Delta^2}+\frac{\omega_3{}^2R_3}{c^2}\right]\le\omega_3R_3;$$
$$\left(1+\frac{e^{\mu T}-1}{\mu T}\right)\frac{|e_3e_1|}{m_3c^3}\left[\frac{\bar{c}\omega_2\left(\breve{\rho}_2+3|\Delta|\right)}{c^2\widehat{\rho}_3\cos(\pi/2-\delta)\Delta^2}+\omega_3{}^2\Phi_3\right]\le\omega_3\Phi_3;$$
$$\left(1+\frac{e^{\mu T}-1}{\mu T}\right)\frac{|e_3e_1|}{m_3c^3}\left[\frac{\bar{c}\omega_2\left(\breve{\rho}_2+3|\Delta|\right)}{c^2\widehat{\rho}_3\Delta^2}+\omega_3{}^2Y_3\right]\le\omega_3Y_3$$

then, there exists a unique T-periodic solution
$(r_2,\phi_2,\eta_2,r_3,\phi_3,\eta_3)\in M_{r_2}\times M_{\phi_2}\times M_{\eta_2}\times M_{r_3}\times M_{\phi_3}\times M_{\eta_3}$ *of (BS).*

Principal remark. *Let us note that we have infinite number of inequalities in the theorem. The above inequalities imply that the operator functions and their first derivatives belong to* $M_{r_2}\times M_{\phi_2}\times M_{\eta_2}\times M_{r_3}\times M_{\phi_3}\times M_{\eta_3}$. *It is easy to check that for the third derivatives, we obtain respectively* $\omega_2{}^4$ *and* $\omega_2{}^3$ *and so on.*

Proof. First, we show that B maps the set $M_{r_2}\times M_{\phi_2}\times M_{\eta_2}\times M_{r_3}\times M_{\phi_3}\times M_{\eta_3}$ into itself. □

From Remark 4, we have $B^{(k)}_{r_2}(.)(T_k)=\frac{1}{2}\int\limits_{T_k}^{T_{k+1}}G_{r_2}(s)ds-\frac{1}{T}\int\limits_{T_k}^{T_{k+1}}\int\limits_{T_k}^{s}G_{r_2}(\theta)d\theta ds=0\,,(k=0,1,2,\dots)$ And since $\int\limits_{T_k}^{T_{k+1}}\left(\frac{t-T_k}{T}-\frac{1}{2}\right)dt=0$, we obtain

$$\int\limits_{T_k}^{T_{k+1}}B^{(k)}_{r_2}(.)(t)\,dt=\int\limits_{T_k}^{T_{k+1}}\int\limits_{T_k}^{t}G_{r_2}(s)ds\,dt-\int\limits_{T_k}^{T_{k+1}}\left(\frac{t-T_k}{T}-\frac{1}{2}\right)dt\int\limits_{T_k}^{T_{k+1}}G_{r_2}(s)ds-\int\limits_{T_k}^{T_{k+1}}\int\limits_{T_k}^{\theta}G_{r_2}(s)ds d\theta=0\,.$$

In addition, in view of $\rho\varphi\le\bar{c}$ and the inequalities from Remark 3, we have

$$\left|B_{r_2}^{(k)}(.)(t)\right| \le \int_{T_k}^{t} |G_{r_2}(s)|ds + \left|\frac{1}{2}\int_{T_k}^{T_{k+1}} G_{r_2}(s)ds + \frac{1}{T}\int_{T_k}^{T_{k+1}}\int_{T_k}^{t} G_{r_2}(s)dsdt\right| = \int_{T_k}^{t} |G_{r_2}(s)|ds + \left|\int_{T_k}^{T_{k+1}} G_{r_2}(s)ds\right| \le$$
$$\le e^{\mu(t-T_k)}\frac{|e_2e_1|}{m_2c}\left(\frac{1}{\widehat{\rho}_2{}^2} + \frac{\beta\omega_3(\breve{\rho}_2+\breve{\rho}_3+3\Delta)}{c^2\Delta^2} + \frac{\omega_2{}^2R_2}{c^2}\right)\left(\frac{1}{\mu} + \frac{e^{\mu T}-1}{\mu}\right) \le$$
$$\le e^{\mu(t-T_k)}\frac{|e_2e_1|}{m_2c}\left(\frac{1}{\widehat{\rho}_2{}^2} + \frac{\beta\omega_3(\breve{\rho}_2+\breve{\rho}_3+3\Delta)}{c^2\Delta^2} + \frac{\omega_2{}^2R_2}{c^2}\right)\frac{e^{\mu T}}{\mu} \le R_2e^{\mu(t-T_k)}.$$

For the second component of B, we have

$$B_{\phi_2}^{(k)}(.)(T_k) = \phi_{02} + \frac{1}{2}\int_{T_k}^{T_{k+1}} G_{\phi_2}(.)(s)ds - \frac{1}{T}\int_{T_k}^{T_{k+1}}\int_{T_k}^{s} G_{\phi_2}(.)(\theta)d\theta ds = \phi_{02},\ (k=0,1,2,\ldots);$$
$$\int_{T_k}^{T_{k+1}} B_{\phi_2}^{(k)}(.)(t)\,dt = T\phi_{02} + \int_{T_k}^{T_{k+1}}\int_{T_k}^{t} G_{\phi_2}(s)ds - \int_{T_k}^{T_{k+1}}\left(\frac{t-T_k}{T}-\frac{1}{2}\right)dt\int_{T_k}^{T_{k+1}} G_{\phi_2}(s)ds - \int_{T_k}^{T_{k+1}}\int_{T_k}^{t} G_{\phi_2}(s)dsdt = T\phi_{02}$$

and

$$\left|B_{\phi_2}^{(k)}(.)(t)\right| \le \phi_{02} + \int_{T_k}^{t}|G_{\phi_2}(s)|ds + \left|\int_{T_k}^{T_{k+1}} G_{\phi_2}(s)ds\right| \le e^{\mu(t-T_k)}\left(\phi_{02} + \frac{|e_2e_3|}{m_2c^3}\left(\frac{\beta\omega_3}{\cos(\pi/2-\delta)}\frac{\breve{\rho}_3+3c\Delta}{\widehat{\rho}_2c\Delta^2} + \omega_2{}^2\Phi_2\right)\frac{e^{\mu T}}{\mu}\right) \le \Phi_2e^{\mu(t-T_k)}.$$

For the third component of B, we obtain

$$\left|B_{\eta_2}^{(k)}(.)(t)\right| \le \int_{T_k}^{t}|G_{\eta_2}(s)|ds + \left|\int_{T_k}^{T_{k+1}} G_{\eta_2}(s)ds\right| \le e^{\mu(t-T_k)}\frac{|e_2e_1|}{m_2}\left(\frac{\bar{c}\omega_3}{c^3}\frac{\breve{\rho}_3e^{\mu T}+3|\Delta|}{\widehat{\rho}_2\Delta^2} + \frac{\omega_2{}^2Y_2}{c^3}\right)\frac{e^{\mu T}}{\mu} \le Y_2e^{\mu(t-T_k)}.$$

For the remaining three components of B, we have

$$\left|B_{r_3}^{(k)}(.)(t)\right| \le e^{\mu(t-T_k)}\frac{|e_3e_1|}{m_3c}\left(\frac{1}{\widehat{\rho}_3{}^2} + \frac{\beta\omega_2(\breve{\rho}_2+\breve{\rho}_3+3)}{c^2\Delta^2} + \frac{\omega_3{}^2R_3}{c^2}\right)\frac{e^{\mu T}}{\mu} \le R_3e^{\mu(t-T_k)};$$
$$\left|B_{\phi_3}^{(k)}(.)(t)\right| \le e^{\mu(t-T_k)}\phi_{03} + e^{\mu(t-T_k)}\frac{|e_3e_1|}{m_3c^3}\left[\frac{\beta\omega_2}{\cos(\pi/2-\delta)}\frac{\breve{\rho}_2+3c\Delta}{\widehat{\rho}_3c\Delta^2} + \omega_3{}^2\Phi_3\right]\frac{e^{\mu T}}{\mu} \le \Phi_3e^{\mu(t-T_k)};$$
$$\left|B_{\eta_3}^{(k)}(.)(t)\right| \le^{\mu(t-T_k)}\frac{|e_3e_1|}{m_3c^3}\left[\frac{\bar{c}\omega_2\left(\breve{\rho}_2+3|\Delta|\right)}{c^2\widehat{\rho}_3\Delta^2} + \omega_3{}^2Y_3\right]\frac{e^{\mu T}}{\mu} \le Y_3e^{\mu(t-T_k)}.$$

For the first-order derivatives of components of B, we obtain

$$\left|\frac{dB_{r_2}^{(k)}(.)(t)}{dt}\right| \le |G_{r_2}(.)(t)| + \frac{1}{T}\left|\int_{T_k}^{T_{k+1}} G_{r_2}(.)(s)ds\right| \le e^{\mu(t-T_k)}\left(1+\frac{e^{\mu T}-1}{\mu T}\right)\frac{|e_2e_1|}{m_2c}\left[\frac{1}{\widehat{\rho}_2{}^2} + \frac{\bar{c}\omega_3\left(\breve{\rho}_2+\breve{\rho}_3+3\Delta\right)}{c^3\Delta^2} + \frac{\omega_2{}^2R_2}{c^2}\right] \le \omega_2R_2e^{\mu(t-T_k)};$$
$$\left|\frac{dB_{\phi_2}^{(k)}(.)(t)}{dt}\right| \le |G_{\phi_2}(.)(t)| + \frac{1}{T}\left|\int_{T_k}^{T_{k+1}} G_{\phi_2}(.)(s)ds\right| \le e^{\mu(t-T_k)}\left(1+\frac{e^{\mu T}-1}{\mu T}\right)\frac{|e_2e_3|}{m_2c^3}\left[\frac{\bar{c}\omega_3\left(\breve{\rho}_3+3\Delta\right)}{c^2\widehat{\rho}_2\cos(\pi/2-\delta)\Delta^2} + \omega_2{}^2\Phi_2\right] \le \omega_2\Phi_2e^{\mu(t-T_k)};$$
$$\left|\frac{dB_{\eta_2}^{(k)}(.)(t)}{dt}\right| \le |G_{\eta_2}(.)(t)| + \frac{1}{T}\left|\int_{T_k}^{T_{k+1}} G_{\eta_2}(.)(s)ds\right| \le e^{\mu(t-T_k)}\left(1+\frac{e^{\mu T}-1}{\mu T}\right)\frac{|e_2e_3|}{m_2c^3}\left[\frac{\bar{c}\omega_3\left(\breve{\rho}_3+3\Delta e^{\mu T}\right)}{c^2\widehat{\rho}_2\Delta^2} + \omega_2{}^2Y_2\right] \le \omega_2Y_2e^{\mu(t-T_k)};$$
$$\left|\frac{dB_{r_3}^{(k)}(.)(t)}{dt}\right| \le e^{\mu(t-T_k)}\left(1+\frac{e^{\mu T}-1}{\mu T}\right)\frac{|e_3e_1|}{m_3c}\left[\frac{1}{\widehat{\rho}_3{}^2} + \frac{\bar{c}\omega_2\left(\breve{\rho}_2+\breve{\rho}_3+3\Delta\right)}{c^3\Delta^2} + \frac{\omega_3{}^2R_3}{c^2}\right] \le \omega_3R_3e^{\mu(t-T_k)};$$
$$\left|\frac{dB_{\phi_3}^{(k)}(.)(t)}{dt}\right| \le e^{\mu(t-kT)}\left(1+\frac{e^{\mu T}-1}{\mu T}\right)\frac{|e_2e_3|}{m_3c^3}\left[\frac{\bar{c}\omega_2\left(\breve{\rho}_2+3\Delta\right)}{c^2\widehat{\rho}_3\cos(\pi/2-\delta)\Delta^2} + \omega_3{}^2\Phi_3\right] \le \omega_3\Phi_3e^{\mu(t-T_k)};$$
$$\left|\frac{dB_{\eta_3}^{(k)}(.)(t)}{dt}\right| \le e^{\mu(t-T_k)}\frac{|e_2e_3|}{m_3c^3}\left[\frac{\bar{c}\omega_2\left(\breve{\rho}_2+3\Delta\right)}{c^2\widehat{\rho}_3\Delta^2} + \omega_3{}^2Y_3\right] \le \omega_3Y_3e^{\mu(t-T_k)}.$$

The estimates for the higher-order derivatives of the components of B are obtained in Appendix D.

In what follows, we show that operator B is a contractive one in the sense of [8].

In view of Remark 2, the retarded functions act in the following way:

$$t-\tau_{23}(t):[T_k,T_{k+1}]\to[T_{k-1},T_k],\ t-\tau_{32}(t):[T_k,T_{k+1}]\to[T_{k-1},T_k].$$

The index set of $M_{r_2}\times M_{\phi_2}\times M_{\eta_2}\times M_{r_3}\times M_{\phi_3}\times M_{\eta_3}$ is (k,m), where $k,m=0,1,2,\ldots$ We introduce the following mappings of the index set into itself: $j_1(k,m)=(k,m)$, $j_2(k,m)=(k-1,m+1)$, $j_3(k,m)=(k,m+2)$.

Denote by $\partial\overline{G}_{r_2}=\max\left\{\frac{\partial\overline{G}_{r_2}}{\partial\rho_2},\frac{\partial\overline{G}_{r_2}}{\partial\varphi_2},\frac{\partial\overline{G}_{r_2}}{\partial\lambda_2},\frac{\partial\overline{G}_{r_2}}{\partial\rho_3},\frac{\partial\overline{G}_{r_2}}{\partial\varphi_3},\frac{\partial\overline{G}_{r_2}}{\partial\lambda_3},\frac{\partial\overline{G}_{r_2}}{\partial\dot{r}_3},\frac{\partial\overline{G}_{r_2}}{\partial\dot{\phi}_3},\frac{\partial\overline{G}_{r_2}}{\partial\dot{\eta}_3},\frac{\partial\overline{G}_{r_2}}{\partial\ddot{r}_2}\right\}$ (cf. Appendix E). Then,

$$\left|B_{r_2}(r_2,\phi_2,\eta_2,r_3,\phi_3,\eta_3)-B_{r_2}(\bar{r}_2,\bar{\phi}_2,\bar{\eta}_2,\bar{r}_3,\bar{\phi}_3,\bar{\eta}_3)\right|\le$$

$$\le\tfrac{\partial\overline{G}_{r_2}}{\partial\rho_2}\int\limits_{T_k}^{t}|\rho_2(s)-\bar{\rho}_2(s)|ds+\tfrac{\partial\overline{G}_{r_2}}{\partial\varphi_2}\int\limits_{T_k}^{t}|\varphi_2(s)-\bar{\varphi}_2(s)|ds+\tfrac{\partial\overline{G}_{r_2}}{\partial\lambda_2}\int\limits_{T_k}^{t}|\lambda_2(s)-\bar{\lambda}_2(s)|ds+\tfrac{\partial\overline{G}_{r_2}}{\partial\rho_3}\int\limits_{T_k}^{t}|\rho_3(s-\tau_{23}(s))-\bar{\rho}_3(s-\tau_{23}(s))|ds+$$

$$+\tfrac{\partial\overline{G}_{r_2}}{\partial\varphi_3}\int\limits_{T_k}^{t}|\varphi_3(s-\tau_{23}(s))-\bar{\varphi}_3(s-\tau_{23}(s))|ds+\tfrac{\partial\overline{G}_{r_2}}{\partial\lambda_3}\int\limits_{T_k}^{t}|\lambda_3(s-\tau_{23}(s))-\bar{\lambda}_3(s-\tau_{23}(s))|ds+\tfrac{\partial\overline{G}_{r_2}}{\partial\dot{r}_3}\int\limits_{T_k}^{t}\left|\dot{r}_3(s-\tau_{23}(s))-\dot{\bar{r}}_3(s-\tau_{23}(s))\right|ds+$$

$$+\tfrac{\partial\overline{G}_{r_2}}{\partial\phi_3}\int\limits_{T_k}^{t}\left|\dot{\underset{\frown 3}{\phi}}(s-\tau_{23}(s))-\dot{\bar{\phi}}_3(s-\tau_{23}(s))\right|ds+\tfrac{\partial\overline{G}_{r_2}}{\partial\dot{\eta}_3}\int\limits_{T_k}^{t}\left|\dot{\eta}_3(s-\tau_{23}(s))-\dot{\bar{\eta}}_3(s-\tau_{23}(s))\right|ds+\tfrac{\partial\overline{G}_{r_2}}{\partial\ddot{r}_2}\int\limits_{T_k}^{t}\left|\ddot{r}_2(s)-\ddot{\bar{r}}_2(s)\right|ds\le$$

$$\le\tfrac{\partial\overline{G}_{r_2}}{\partial\rho_2}\int\limits_{T_k}^{t}\int\limits_{T_k}^{s}|r_2(p)-\bar{r}_2(p)|e^{-\mu(p-T_k)}e^{\mu(p-T_k)}dpds+\tfrac{\partial\overline{G}_{r_2}}{\partial\varphi_2}\int\limits_{T_k}^{t}\int\limits_{T_k}^{s}|\phi_2(p)-\bar{\phi}_2(p)|e^{-\mu(p-T_k)}e^{\mu(p-T_k)}dpds+$$

$$+\tfrac{\partial\overline{G}_{r_2}}{\partial\lambda_2}\int\limits_{T_k}^{t}\int\limits_{T_k}^{s}|\eta_2(p)-\bar{\eta}_2(p)|e^{-\mu(p-T_k)}e^{\mu(p-T_k)}dpds+\tfrac{\partial\overline{G}_{r_2}}{\partial\rho_3}\int\limits_{T_k}^{t}\int\limits_{T_k}^{s-\tau_{23}(s)}|r_3(p)-\bar{r}_3(p)|e^{-\mu(p-T_k)}e^{\mu(p-T_k)}dpds+$$

$$+\tfrac{\partial\overline{G}_{r_2}}{\partial\varphi_3}\int\limits_{T_k}^{t}\int\limits_{T_k}^{s-\tau_{23}(s)}|\phi_3(p)-\bar{\phi}_3(p)|e^{-\mu(p-T_k)}e^{\mu(p-T_k)}dpds+\tfrac{\partial\overline{G}_{r_2}}{\partial\lambda_3}\int\limits_{T_k}^{t}\int\limits_{T_k}^{s-\tau_{23}(s)}|\eta_2(p)-\bar{\eta}_2(p)|e^{-\mu(p-T_k)}e^{\mu(p-T_k)}dpds+$$

$$+\tfrac{\partial\overline{G}_{r_2}}{\partial\dot{r}_3}\int\limits_{T_k}^{t}\left|\dot{r}_3(s-\tau_{23}(s))-\dot{\bar{r}}_3(s-\tau_{23}(s))\right|e^{-\mu(s-\tau_{23}(s)-T_k)}e^{\mu(s-\tau_{23}(s)-T_k)}ds+\tfrac{\partial\overline{G}_{r_2}}{\partial\phi_3}\int\limits_{T_k}^{t}\left|\dot{\underset{\frown 3}{\phi}}(s-\tau_{23}(s))-\dot{\bar{\phi}}_3(s-\tau_{23}(s))\right|e^{-\mu(s-\tau_{23}(s)-T_k)}e^{\mu(s-\tau_{23}(s)-T_k)}ds+$$

$$+\tfrac{\partial\overline{G}_{r_2}}{\partial\dot{\eta}_3}\int\limits_{T_k}^{t}\left|\dot{\eta}_3(s-\tau_{23}(s))-\dot{\bar{\eta}}_3(s-\tau_{23}(s))\right|e^{-\mu(s-\tau_{23}(s)-T_k)}e^{\mu(s-\tau_{23}(s)-T_k)}ds+\tfrac{\partial\overline{G}_{r_2}}{\partial\ddot{r}_2}\int\limits_{T_k}^{t}\left|\ddot{r}_2(s)-\ddot{\bar{r}}_2(s)\right|e^{-\mu(s-T_k)}e^{\mu(s-T_k)}ds\le$$

$$\le\partial\overline{G}_{r_2}e^{\mu(t-T_k)}\left[\tfrac{d_{(k,0)}(r_2,\bar{r}_2)}{\mu^2}+\tfrac{d_{(k,0)}(\phi_2,\bar{\phi}_2)}{\mu^2}+\tfrac{d_{(k,0)}(\eta_2,\bar{\eta}_2)}{\mu^2}+\left(\tfrac{d_{(k,0)}(r_3,\bar{r}_3)}{\mu}+\tfrac{d_{(k,0)}(\phi_3,\bar{\phi}_3)}{\mu}+\tfrac{d_{(k,0)}(\eta_3,\bar{\eta}_3)}{\mu}\right)\left|\int\limits_{T_k}^{t}\left(e^{\mu(s-\tau_{23}-T_k)}-1\right)ds\right|+\right.$$

$$\left.+\left(\tfrac{\omega d_{(k-1,1)}(r_3,\bar{r}_3)}{\mu}+\tfrac{\omega d_{(k-1,1)}(\phi_3,\bar{\phi}_3)}{\mu}+\tfrac{\omega d_{(k-1,1)}(\eta_3,\bar{\eta}_3)}{\mu}\right)\int\limits_{T_k}^{t}e^{\mu(s-\tau_{23}(s)-T_k)}ds+\tfrac{\omega^2 d_{(k,2)}(r_2,\bar{r}_2)}{\mu}\int\limits_{T_k}^{t}e^{\mu(s-T_k)}ds\right]\le$$

$$\le\partial\overline{G}_{r_2}\left[e^{\mu(t-T_k)}\left(\tfrac{d_{(k,0)}(r_2,\bar{r}_2)}{\mu^2}+\tfrac{d_{(k,0)}(\phi_2,\bar{\phi}_2)}{\mu^2}+\tfrac{d_{(k,0)}(\eta_2,\bar{\eta}_2)}{\mu^2}\right)+\left(\tfrac{d_{(k,0)}(r_3,\bar{r}_3)}{\mu}+\tfrac{d_{(k,0)}(\phi_3,\bar{\phi}_3)}{\mu}+\tfrac{d_{(k,0)}(\eta_3,\bar{\eta}_3)}{\mu}\right)\left|\int\limits_{T_k}^{t}e^{\mu(s-T_k)}(1-e^{-\mu\tau_{23}})ds\right|+\right.$$

$$\left.+e^{\mu(t-T_k)}\tfrac{\omega}{\mu^2}\left(d_{(k-1,1)}(r_3,\bar{r}_3)+d_{(k-1,1)}(\phi_3,\bar{\phi}_3)+d_{(k-1,1)}(\eta_3,\bar{\eta}_3)\right)+e^{\mu(t-T_k)}\tfrac{\omega^2 d_{(k,2)}(r_2,\bar{r}_2)}{\mu^2}\right]\le$$

$$\le\partial\overline{G}_{r_2}e^{\mu(t-T_k)}\left[2\tfrac{d_{(k,0)}(r_2,\bar{r}_2)+d_{(k,0)}(\phi_2,\bar{\phi}_2)+d_{(k,0)}(\eta_2,\bar{\eta}_2)}{\mu^2}+\tfrac{\omega\left(d_{(k-1,1)}(r_3,\bar{r}_3)+d_{(k-1,1)}(\phi_3,\bar{\phi}_3)+d_{(k-1,1)}(\eta_3,\bar{\eta}_3)\right)}{\mu^2}+\tfrac{\omega^2 d_{(k,2)}(r_2,\bar{r}_2)}{\mu^2}\right].$$

It follows that

$$d_{(k,0)}\left(B^{(k)}_{r_2}(.),\overline{B}^{(k)}_{r_2}(.)\right)\le$$
$$\le\partial\overline{G}_{r_2}\left[\tfrac{2}{\mu^2}d_{(k,0)}\left((r_2,\phi_2,\eta_2,r_3,\phi_3,\eta_3),(\bar{r}_2,\bar{\phi}_2,\bar{\eta}_2,\bar{r}_3,\bar{\phi}_3,\bar{\eta}_3)\right)+\tfrac{\omega}{\mu^2}d_{(k-1,1)}\left((r_2,\phi_2,\eta_2,r_3,\phi_3,\eta_3),(\bar{r}_2,\bar{\phi}_2,\bar{\eta}_2,\bar{r}_3,\bar{\phi}_3,\bar{\eta}_3)\right)+\right.$$
$$\left.+\tfrac{\omega^2}{\mu^2}d_{(k,2)}\left((r_2,\phi_2,\eta_2,r_3,\phi_3,\eta_3),(\bar{r}_2,\bar{\phi}_2,\bar{\eta}_2,\bar{r}_3,\bar{\phi}_3,\bar{\eta}_3)\right)\right].$$

In a similar way, one obtains

$$d_{(k,0)}\left(B_{r_3}^{(k)}(.),\overline{B}_{r_3}^{(k)}(.)\right)\leq$$
$$\leq\partial\overline{G}_{r_3}\left[\frac{2}{\mu^2}d_{(k,0)}\left((r_2,\phi_2,\eta_2,r_3,\phi_3,\eta_3),(\overline{r}_2,\overline{\phi}_2,\overline{\eta}_2,\overline{r}_3,\overline{\phi}_3,\overline{\eta}_3)\right)+\frac{\omega}{\mu^2}d_{(k-1,1)}\left((r_2,\phi_2,\eta_2,r_3,\phi_3,\eta_3),(\overline{r}_2,\overline{\phi}_2,\overline{\eta}_2,\overline{r}_3,\overline{\phi}_3,\overline{\eta}_3)\right)+\right.$$
$$\left.+\frac{\omega^2}{\mu^2}d_{(k,2)}\left((r_2,\phi_2,\eta_2,r_3,\phi_3,\eta_3),(\overline{r}_2,\overline{\phi}_2,\overline{\eta}_2,\overline{r}_3,\overline{\phi}_3,\overline{\eta}_3)\right)\right].$$

Further on, we have

$$\left|B_{\phi_2}(r_2,\phi_2,\eta_2,r_3,\phi_3,\eta_3)-B_{\phi_2}(\overline{r}_2,\overline{\phi}_2,\overline{\eta}_2,\overline{r}_3,\overline{\phi}_3,\overline{\eta}_3)\right|\leq$$
$$\leq\frac{\partial\overline{G}_{\phi_2}}{\partial\rho_2}\int\limits_{T_k}^{t}|\rho_2(s)-\overline{\rho}_2(s)|ds+\frac{\partial\overline{G}_{\phi_2}}{\partial\varphi_2}\int\limits_{T_k}^{t}|\varphi_2(s)-\overline{\varphi}_2(s)|ds+\frac{\partial\overline{G}_{\phi_2}}{\partial\lambda_2}\int\limits_{T_k}^{t}\left|\lambda_2(s)-\overline{\lambda}_2(s)\right|ds+\frac{\partial\overline{G}_{\phi_2}}{\partial\rho_3}\int\limits_{T_k}^{t}\left|\rho_3(s-\tau_{23}(s))-\overline{\rho}_3(s-\tau_{23}(s))\right|ds+$$
$$\frac{\partial\overline{G}_{\phi_2}}{\partial\varphi_3}\int\limits_{T_k}^{t}\left|\varphi_3(s-\tau_{23}(s))-\overline{\varphi}_3(s-\tau_{23}(s))\right|ds+\frac{\partial\overline{G}_{\phi_2}}{\partial\lambda_3}\int\limits_{T_k}^{t}\left|\lambda_3(s-\tau_{23}(s))-\overline{\lambda}_3(s-\tau_{23}(s))\right|ds+\frac{\partial\overline{G}_{\phi_2}}{\partial\dot{r}_3}\int\limits_{T_k}^{t}\left|\dot{r}_3(s-\tau_{23}(s))-\dot{\overline{r}}_3(s-\tau_{23}(s))\right|ds+$$
$$\frac{\partial\overline{G}_{\phi_2}}{\partial\phi_3}\int\limits_{T_k}^{t}\left|\dot{\phi}_{\frown 3}(s-\tau_{23}(s))-\dot{\overline{\phi}}_3(s-\tau_{23}(s))\right|ds+\frac{\partial\overline{G}_{\phi_2}}{\partial\eta_3}\int\limits_{T_k}^{t}\left|\dot{\eta}_3(s-\tau_{23}(s))-\dot{\overline{\eta}}_3(s-\tau_{23}(s))\right|ds+\frac{\partial\overline{G}_{\phi_2}}{\partial\dot{r}_2}\int\limits_{T_k}^{t}\left|\ddot{\phi}_2(s)-\ddot{\overline{\phi}}_2(s)\right|ds\leq$$
$$\left|\frac{\partial G_{\phi_2}}{\partial\rho_2}\right|\leq\frac{|e_2e_3|\overline{c}}{m_2\cos(\pi/2-\delta)}\left(\frac{c^3\breve{\rho}_3}{(\Delta-(R_3/\mu)e^{\mu T})^3}\frac{2\omega}{\widehat{\rho}_2{}^2}+\frac{6c}{c^4\Delta-(R_3/\mu)e^{\mu T})\widehat{\rho}_2{}^2}+\frac{c\omega}{c^4\Delta-(R_3/\mu)e^{\mu T})\widehat{\rho}_2}\right)=\frac{\partial\overline{G}_{\phi_2}}{\partial\rho_2}$$

and so on.

Denoting by $\overline{B}_{r_2}=B_{r_2}(\overline{r}_2,\overline{\phi}_2,\overline{\eta}_2,\overline{r}_3,\overline{\phi}_3,\overline{\eta}_3),\ldots,\overline{B}_{\eta_3}=B_{\eta_3}(\overline{r}_2,\overline{\phi}_2,\overline{\eta}_2,\overline{r}_3,\overline{\phi}_3,\overline{\eta}_3)$, we obtain

$$d_{(k,0)}\left((B_{r_2},B_{\phi_2},B_{\eta_3},B_{r_3},B_{\phi_3},B_{\eta_3}),(\overline{B}_{r_2},\overline{B}_{\phi_2},\overline{B}_{\eta_3},\overline{B}_{r_3},\overline{B}_{\phi_3},\overline{B}_{\eta_3})\right)\leq$$
$$\leq\frac{2(\partial\overline{G}_{r_3}+\ldots+\partial\overline{G}_{\eta_3})}{\mu^2}d_{(k,0)}\left((r_2,\phi_2,\eta_2,r_3,\phi_3,\eta_3),(\overline{r}_2,\overline{\phi}_2,\overline{\eta}_2,\overline{r}_3,\overline{\phi}_3,\overline{\eta}_3)\right)+\frac{\omega(\partial\overline{G}_{r_3}+\ldots+\partial\overline{G}_{\eta_3})}{\mu^2}d_{(k-1,1)}\left((r_2,\phi_2,\eta_2,r_3,\phi_3,\eta_3),(\overline{r}_2,\overline{\phi}_2,\overline{\eta}_2,\overline{r}_3,\overline{\phi}_3,\overline{\eta}_3)\right)+$$
$$+\frac{\omega^2(\partial\overline{G}_{r_3}+\ldots+\partial\overline{G}_{\eta_3})}{\mu^2}d_{(k,2)}\left((r_2,\phi_2,\eta_2,r_3,\phi_3,\eta_3),(\overline{r}_2,\overline{\phi}_2,\overline{\eta}_2,\overline{r}_3,\overline{\phi}_3,\overline{\eta}_3)\right).$$

If $\frac{2(\partial\overline{G}_{r_3}+\ldots+\partial\overline{G}_{\eta_3})}{\mu^2}<\frac{1}{3};\ \frac{\omega(\partial\overline{G}_{r_3}+\ldots+\partial\overline{G}_{\eta_3})}{\mu^2}<\frac{1}{3};\ \frac{\omega^2(\partial\overline{G}_{r_3}+\ldots+\partial\overline{G}_{\eta_3})}{\mu^2}<\frac{1}{3}$ are not satisfied, then we apply the inequalities from Appendix A and obtain the desired inequalities for a sufficiently large $m\in N$:

$$\frac{\omega^m}{\mu^m}\frac{2(\partial\overline{G}_{r_3}+\ldots+\partial\overline{G}_{\eta_3})}{\mu^2}<\frac{1}{3};\ \frac{\omega^m}{\mu^m}\frac{\omega(\partial\overline{G}_{r_3}+\ldots+\partial\overline{G}_{\eta_3})}{\mu^2}<\frac{1}{3};\ \frac{\omega^m}{\mu^m}\frac{\omega^2(\partial\overline{G}_{r_3}+\ldots+\partial\overline{G}_{\eta_3})}{\mu^2}<\frac{1}{3}.$$

It is easy to see that the space $M_{r_2}\times M_{\phi_2}\times M_{\eta_2}\times M_{r_3}\times M_{\phi_3}\times M_{\eta_3}$ is $(j_1,j_2,j_3)-$bounded and the maps $j_s:A\to A\ (s=1,2,3)$ are commutative. Then, the operator $(B_{2r},B_{2\phi},B_{3r},B_{3\phi})$, in view of the fixed point theorem from [8], has a unique fixed point, which, in view of the Main Lemma, is a T-periodic solution of (BS).

Theorem 1 is thus proved.

Numerical results

We check the inequalities from the Main Theorem:

$$\frac{|e_2e_1|}{m_2}\left(\frac{1}{c\widehat{\rho}_2{}^2}+\frac{\overline{c}\omega_3(\breve{\rho}_2+\breve{\rho}_3+3\Delta)}{c^3\Delta^2}+\frac{\omega_2{}^2R_2}{c^3}\right)\frac{e^{\mu T}}{\mu}\leq R_2;\phi_{02}+\frac{|e_2e_3|}{m_2}\left(\frac{\overline{c}\omega_3(\breve{\rho}_3+3\Delta)}{\widehat{\rho}_2c^3\Delta^2\cos(\pi/2-\delta)}+\frac{\omega_2{}^2\Phi_2}{c^3}\right)\frac{e^{\mu T}}{\mu}\leq\Phi_2;$$
$$\frac{|e_2e_3|}{m_2}\left(\frac{\overline{c}\omega_3}{c^3}\frac{\breve{\rho}_3e^{\mu T}+3\Delta}{\widehat{\rho}_2\Delta^2}+\frac{\omega_2{}^2\Upsilon_2}{c^3}\right)\frac{e^{\mu T}}{\mu}\leq\Upsilon_2.$$

It is known (cf. [21]) that

$e_n=1.6\times10^{-19}C$; $m_n=9.11\times10^{-19}kg\ (n=2,3)$; $c=3\times10^8m/\sec$; $\breve{\rho}_n\approx\widehat{\rho}_n=5.3\times10^{-11}m$; $\omega_2=\omega_3=4\times10^{16}$, $T=\frac{2\pi}{\omega}=\frac{6.28}{4\times10^{16}}\approx1.57\times10^{-16}$, $\Delta\approx c$, $\delta=\frac{\pi}{6}$, $\phi_{20}<\Phi_0$.

Choosing $\mu=6\times10^{16}>\omega=4\times10^{16}$ and $\mu T=6\times10^6\times1,.57\times10^{-16}=9.42>\omega T=2\pi$, one obtains $e^{9.42}\approx1.2\times10^4$, $\frac{e^{\mu T}}{\mu}=\frac{1.2\times10^4}{6\times10^{16}}\approx2\times10^{-13}$.

$$\text{We have } (R_2+R_3)\tfrac{e^{\mu T}}{\mu} \le \tfrac{2\bar{c}}{\mu} = \tfrac{2.2\times10^6}{6\times10^{16}} \approx 0 \Rightarrow \Delta \approx |\rho_{02}-\rho_{03}|;$$

$$2.81\times10^{-8}\left(\tfrac{1}{3\times10^8\times5.3\times10^{-11}} + \tfrac{4\times10^{16}(2\times5.3\times10^{-11}+3|\rho_{02}-\rho_{03}|)}{137\times9\times10^{16}|\rho_{02}-\rho_{03}|^2} + \tfrac{16\times10^{32}}{27\times10^{24}}R_2\right)2\times10^{-13} \le R_2;$$

$$5.62\times10^{-21}\left(\tfrac{4\times10^{16}(5.3\times10^{-11}+3|\rho_{02}-\rho_{03}|)}{137\times5.3\times10^{-11}\times9\times10^{16}|\rho_{02}-\rho_{03}|^2\times0.5} + \tfrac{4^2\times10^{32}}{27\times10^{24}}\Phi_2\right) \le \Phi_2-\phi_{02};$$

$$2.81\times10^{-8}\left(\tfrac{4\times10^{16}}{137\times9\times10^{16}}\tfrac{5.3\times10^{-11}\times4.8+3|\rho_{02}-\rho_{03}|}{5.3\times10^{-11}|\rho_{02}-\rho_{03}|^2} + \tfrac{16\times10^{32}}{27\times10^{24}}Y_2\right)2\times10^{-13} \le Y_2.$$

Since $|\rho_{02}-\rho_{03}| \approx 10^{-11}$, then

$$5.62\left(\frac{10^3}{15.9} + \frac{1}{4.11\times10^{12}\times10^{-22}} + \frac{1}{4.11\times10^{18}\times10^{-11}} + 6\times10^7 R_2\right)10^{-21} \le R_2.$$

For the second inequality, we obtain

$$5.62\times10^{-21}\left(6.5\times10^{19}+3.7\times10^{19}+0.6\times10^8\Phi_2\right) \le \Phi_2-\phi_{02}.$$

Finally, $5.62\left(1.6\times10^{20}+1.8\times10^{19}+6.10^7Y_2\right)10^{-21} \le Y_2$.

7. Discussion on the Bohr and Sommerfeld Hypotheses

In accordance with Bohr's first hypothesis (cf. [22,23]) both electrons of the He atom move in circle orbits in a plane on opposite sides. This means that we have the Kepler problem with the conditions $\lambda_2=0,\ \lambda_3=0,\ \varphi_3(t)=\varphi_2(t)+\pi$, which yield

$$\varphi_2(t)-\varphi_2(t-\tau_{23}) \approx \phi_2\tau_{23} \approx \omega\tfrac{\rho_{23}}{2c} \approx 10^{16}\tfrac{5.3\times10^{-11}}{6\times10^8} \approx 10^{-3};$$

$$\varphi_3(t-\tau_{32})-\varphi_3(t) \approx \phi_3\tau_{32} \approx \omega\tfrac{\rho_{23}}{2c} \approx 10^{16}\tfrac{5.3\times10^{-11}}{6\times10^8} \approx 10^{-3}; \text{and then,}$$

$$\cos\varphi_{23} = \cos(\varphi_2(t)-\varphi_3(t-\tau_{23})) = \cos(\varphi_2(t)-\varphi_2(t-\tau_{23})-\pi) = -\cos(\varphi_2(t)-\varphi_2(t-\tau_{23})) \simeq -1;$$

$$\sin\varphi_{23} = -\sin(\varphi_2-\varphi_3(t-\tau_{23})) = -\sin(\varphi_2(t)-\varphi_2(t-\tau_{23})-\pi) = \sin(\varphi_2(t)-\varphi_2(t-\tau_{23})) \simeq 0.$$

Since $\left\langle \vec{\xi}^{(23)}, \dot{\vec{u}}^{(3)} \right\rangle = \rho_2[\cos(\varphi_2-\varphi_3)-1]\ddot{\rho}_3 = -2\rho_2\ddot{\rho}_3$, $\left\langle \vec{\xi}^{(32)}, \dot{\vec{u}}^{(2)} \right\rangle = -2\rho_3\ddot{\rho}_2$, the equations of motion become

$$\dot{r}_2 = \tfrac{e_2e_1}{m_2}\tfrac{1}{c\rho_2{}^2} + \tfrac{e_2e_3}{m_2}\left(\tfrac{4\rho_2{}^2\dot{r}_3}{c^6\tau_{23}{}^3} + \tfrac{\dot{r}_3}{c^4\tau_{23}}\right) - \tfrac{e_2^2}{m_2}\tfrac{\ddot{r}_2}{c^3};\ \dot{\phi}_2 = -\tfrac{e_2e_3}{m_2}\left(\tfrac{\rho_3\dot{\phi}_3}{c^4\tau_{23}\rho_2} + \tfrac{\ddot{\phi}_2}{c^3}\right);$$

$$\dot{r}_3 = \tfrac{e_3e_1}{m_3}\tfrac{1}{c\rho_3{}^2} - \tfrac{e_2e_3}{m_3}\left(\tfrac{4\rho_3{}^2\ddot{\rho}_2}{c^6\tau_{32}{}^3} + \tfrac{\dot{r}_2}{c^4\tau_{23}}\right) - \tfrac{e_3^2}{m_3}\tfrac{\ddot{r}_3}{c^3};\ \dot{\phi}_3 = \tfrac{e_3e_2}{m_3}\left(\tfrac{\rho_2\dot{\phi}_2\cos\varphi_{32}}{c^4\tau_{32}\rho_3} - \tfrac{\ddot{\phi}_3}{c^3}\right).$$

This system is a particular case of (BS) and has a unique periodic solution.

The second Bohr hypothesis (cf. [22,23]) is as follows: both electrons move in two different planes at an angle of 60 degrees. This means that one can choose $\lambda_2=0,\ \lambda_3=\pi/3 \Rightarrow \eta_3=0,\ \varphi_3(t)=\varphi_2(t)+\alpha>0$, where α will be defined below.

From (BS), the system of equations of motion is

$$\begin{aligned}
\dot{r}_2 &= \tfrac{e_2e_1}{m_2}\tfrac{1}{c\rho_2{}^2} - \tfrac{e_2e_3}{m_2}\tfrac{\rho_2-\rho_3\cos\varphi_{23}}{c^3}\tfrac{\left\langle \vec{\xi}^{(23)}, \dot{\vec{u}}^{(3)} \right\rangle}{c^3\tau_{23}{}^3} - \tfrac{e_2e_3}{m_2}\tfrac{\dot{r}_3\cos\varphi_{23}+\rho_3\dot{\phi}_3\sin\varphi_{23}}{c^4\tau_{23}} - \tfrac{e_2^2}{m_2}\tfrac{\ddot{r}_2}{c^3};\\
\dot{\phi}_2 &= \tfrac{e_2e_3}{m_2}\left[\tfrac{\rho_3\sin\varphi_{23}}{\tau_{23}{}^3\rho_2}\tfrac{\left\langle \vec{\xi}^{(23)}, \dot{\vec{u}}^{(3)} \right\rangle}{c^6} + \tfrac{\dot{r}_3\sin\varphi_{23}+\rho_3\dot{\phi}_3\cos\varphi_{23}}{c^4\tau_{23}\rho_2} - \tfrac{\ddot{\phi}_2}{c^3}\right];\\
\dot{r}_3 &= \tfrac{e_3e_1}{m_3}\tfrac{1}{c\rho_3{}^2} - \tfrac{e_2e_3}{m_3}\tfrac{\rho_3-\rho_2\cos\varphi_{32}}{c^3}\tfrac{\left\langle \vec{\xi}^{(32)}, \dot{\vec{u}}^{(2)} \right\rangle}{c^3\tau_{32}{}^3} - \tfrac{e_3e_2}{m_3}\tfrac{\dot{r}_2\cos\varphi_{32}+\rho_2\dot{\phi}_2\sin\varphi_{32}}{c^4\tau_{23}} - \tfrac{e_3^2}{m_3}\tfrac{\ddot{r}_3}{c^3};\\
\dot{\phi}_3 &= \tfrac{e_3e_2}{m_3}\left[\tfrac{\rho_2\sin\varphi_{32}}{\tau_{32}{}^3\rho_3}\tfrac{\left\langle \vec{\xi}^{(32)}, \dot{\vec{u}}^{(2)} \right\rangle}{c^6} + \tfrac{\dot{r}_2\sin\varphi_{32}+\rho_2\dot{\phi}_2\cos\varphi_{32}}{c^4\tau_{32}\rho_3} - \tfrac{\ddot{\phi}_3}{c^3}\right];\\
0 &= \tfrac{\rho_2\cos\varphi_{32}}{\tau_{32}{}^2}\tfrac{(\rho_3\cos\varphi_{32}-2\rho_2)\dot{r}_2+\sin\varphi_{32}\rho_2\rho_3\dot{\phi}_2}{2c^6} + \tfrac{\dot{r}_3\cos\varphi_{32}+\rho_2\dot{\phi}_2\sin\varphi_{32}}{c^4}.
\end{aligned} \tag{11}$$

Equation (11) can be considered as an additional condition for the existence of a periodic solution. Indeed,

$$\varphi_{23} = \varphi_2(t) - \varphi_3(t-\tau_{23}),\ \varphi_3(t) = \varphi_2(t) + \alpha, \varphi_{23} = \varphi_2(t) - \varphi_2(t-\tau_{23}) - \alpha \approx -\alpha;$$
$$\varphi_{32} = \varphi_2(t-\tau_{32}) - \varphi_3(t),\ \varphi_3(t) = \varphi_2(t) + \alpha, \varphi_{32} = \varphi_2(t-\tau_{32}) - \varphi_2(t) - \alpha \approx -\alpha;$$
$$\cos\varphi_{32} \approx \cos\alpha, \sin\varphi_{32} \approx -\sin\alpha.$$

Then, Equation (11) becomes

$$0 = \frac{\rho_2\cos\alpha}{{\tau_{32}}^2}\frac{(\rho_3\cos\alpha - 2\rho_2)\dot{r}_2 + \rho_2\rho_3\dot{\phi}_2\sin\alpha}{2c^6} + \frac{\dot{r}_3\cos\alpha + \rho_2\dot{\phi}_2\sin\alpha}{c^4}.$$

For $\rho_2 = \rho_3 = \rho = const.$, one obtains

$$\left(\frac{\cos\alpha}{{\tau_{32}}^2}\frac{\rho^2}{2c^2} + 1\right)\frac{\dot{\phi}_2\sin\alpha}{c^4} = 0 \Rightarrow \sin\alpha = 0 \Rightarrow \alpha = 0 \vee \alpha = \pi,$$

which confirms the Bohr hypothesis for $\alpha = \pi$. Then, the equations of motion become

$$\dot{r}_2 = \frac{e_2e_1}{m_2}\frac{1}{c{\rho_2}^2} - \frac{e_2e_3}{m_2}\frac{\rho_2}{c^3}\frac{(\rho_2+2\rho_3)\dot{r}_3}{2c^3{\tau_{23}}^3} - \frac{e_2e_3}{m_2}\frac{\dot{r}_3}{2c^4\tau_{23}} - \frac{e_2^2}{m_2}\frac{\ddot{r}_2}{c^3};\ \dot{\phi}_2 = -\frac{e_2e_3}{m_2}\left(\frac{2\rho_3\dot{\phi}_3}{c^4\tau_{23}\rho_2} + \frac{\ddot{\phi}_2}{c^3}\right);$$
$$\dot{r}_3 = \frac{e_3e_1}{m_3}\frac{1}{c{\rho_3}^2} + \frac{e_2e_3}{m_3}\frac{(\rho_3+2\rho_2)\dot{r}_2}{2c^3{\tau_{32}}^3} + \frac{e_3e_2}{m_3}\frac{\dot{r}_2}{2c^4\tau_{23}} - \frac{e_3^2}{m_3}\frac{\ddot{r}_3}{c^3};\ \dot{\phi}_3 = -\frac{e_3e_2}{m_3}\left(\frac{2\rho_2\dot{\phi}_2}{c^4\tau_{32}\rho_3} + \frac{\ddot{\phi}_3}{c^3}\right).$$

Later, additional assumptions are made by A. Sommerfeld [23], but they turn out, again, to be a particular case of our equations. We notice that the spin equation for the plane Kepler problem (investigated in [24]) has solutions with opposite signs.

8. Conclusions

We have obtained conditions for the existence–uniqueness of a periodic solution for the three-body problem of classical electrodynamics. This includes the He atom and the He-like atoms—for instance, the Li atom without one electron. We give a new formulation of the retarded action between moving particles, splitting the equations of motion. As in previous papers, we use an operator formulation of a periodic problem, and by the fixed point method, we prove the existence–uniqueness of a periodic solution of the 3D Kepler problem. In this manner, we show the stability of the He atom in the frame of classical electrodynamics. The same approach can be used for the investigation of a similar problem for atoms with more complicated structures. As a particular case, we obtain a confirmation of the hypotheses stated by Bohr and Sommerfeld (cf. [22,23]).

Funding: This research received no external funding.

Informed Consent Statement: Not applicable.

Acknowledgments: The author thanks Takayuki Hibi for his invitation to submit a paper and the referees, whose very useful suggestions improved the paper.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A. Derivation of Initial Equations in Spherical Coordinates

We pass to the spherical coordinates, that is, the particle $P_n(n = 2, 3)$ is located at the point

$$x_1{}^{(n)}(\theta) = \rho_n(\theta)\cos\varphi_n(\theta)\cos\lambda_n(\theta);$$
$$x_2{}^{(n)}(\theta) = \rho_n(\theta)\sin\varphi_n(\theta)\cos\lambda_n(\theta);$$
$$x_3{}^{(n)}(\theta) = \rho_n(\theta)\sin\ \lambda_n(\theta)$$

where $\rho_n(\theta) \geq 0$; $\varphi_n(\theta) \geq 0$; $\lambda_n(\theta) \in \left[-\frac{\pi}{2} + \delta, \frac{\pi}{2} - \delta\right]$, $0 < \delta < \frac{\pi}{2}$, and then the velocities are

$$u_1{}^{(n)} = \dot{\rho}_n\cos\varphi_n\cos\lambda_n - \rho_n\dot{\varphi}_n\sin\varphi_n\cos\lambda_n\ - \rho_n\dot{\lambda}_n\cos\varphi_n\sin\lambda_n$$
$$u_2{}^{(n)} = \dot{\rho}_n\sin\varphi_n\cos\lambda_n + \rho_n\dot{\varphi}_n\cos\varphi_n\ \cos\lambda_n - \rho_n\dot{\lambda}_n\sin\varphi_n\sin\lambda_n$$
$$u_3{}^{(n)} = \dot{\rho}_n\sin\lambda_n + \rho_n\ \dot{\lambda}_n\cos\lambda_n$$

where $\dot{\rho}^{(2)} = d\rho^{(2)}/d\theta$. For the accelerations, we have

$$\begin{aligned}
\dot{u}_1{}^{(n)} &\approx \ddot{\rho}_n \cos\varphi_n \cos\lambda_n - \rho_n \ddot{\varphi}_n \sin\varphi_n \cos\lambda_n - \rho_n \ddot{\lambda}_n \cos\varphi_n \sin\lambda_n \\
\dot{u}_2{}^{(n)} &\approx \ddot{\rho}_n \sin\varphi_n \cos\lambda_n + \rho_n \ddot{\varphi}_n \cos\varphi_n \cos\lambda_n - \rho_n \ddot{\lambda}_n \sin\varphi_n \sin\lambda_n \\
\dot{u}_3{}^{(n)} &\approx \ddot{\rho}_n \sin\lambda_n + \rho_n \ddot{\lambda}_n \cos\lambda_n.
\end{aligned}$$

In the same manner, we obtain the expressions for the second derivatives ($n = 2, 3$):

$$\begin{aligned}
\ddot{u}_1^{(n)} &\approx \dddot{\rho}_n \cos\varphi_n \cos\lambda_n - \dddot{\varphi}_n \rho_n \sin\varphi_n \cos\lambda_n - \dddot{\lambda}_n \rho_n \cos\varphi_n \sin\lambda_n \\
\ddot{u}_2^{(n)} &\approx \dddot{\rho}_n \sin\varphi_n \cos\lambda_n + \dddot{\varphi}_n \rho_n \cos\varphi_n \ \cos\lambda_n - \dddot{\lambda}_n \rho_n \sin\varphi_n \sin\lambda_n \\
\ddot{u}_3^{(n)} &\approx \dddot{\rho}_n \sin\lambda_n + \dddot{\lambda}_n \rho_n \cos\lambda_n
\end{aligned}$$

The condition (C) implies

$$\begin{aligned}
&\left\langle \vec{u}^{(2)}, \vec{u}^{(2)} \right\rangle = \dot{\rho}_2{}^2 + \rho_2{}^2 \dot{\varphi}_2{}^2 \cos^2\lambda_2 + \rho_2{}^2 \dot{\lambda}_2{}^2 \le \bar{c}^2 < c^2 \ \Rightarrow \ \tfrac{|\dot{\rho}_2|}{c} \approx 0, \ \tfrac{|\rho_2 \dot{\varphi}_2|}{c} \approx 0, \tfrac{\left|\rho_2 \dot{\lambda}_2\right|}{c} \approx 0; \\
&\left\langle \vec{u}^{(2)}, \vec{u}^{(2)} \right\rangle / c^2 \approx 0; \\
&\left\langle \dot{\vec{u}}^{(2)}, \dot{\vec{u}}^{(2)} \right\rangle = \ddot{\rho}_2{}^2 + \rho_2{}^2 \ddot{\varphi}_2{}^2 \cos^2\lambda_2 + \rho_2{}^2 \ddot{\lambda}_2{}^2; \ \left\langle \ddot{\vec{u}}^{(2)}, \ddot{\vec{u}}^{(2)} \right\rangle = \dddot{\rho}_2{}^2 + \rho_2{}^2 \dddot{\varphi}_2{}^2 \cos^2\lambda_2 + \rho_2{}^2 \dddot{\lambda}_2{}^2.
\end{aligned}$$

Further on, we have

$$\begin{aligned}
&\vec{\xi}^{(12)} = (\xi_1^{(12)}, \xi_1^{(12)}, \xi_1^{(12)}) = \left(x_1^{(1)}(t) - x_1^{(2)}(t - \tau_{12}(t)), x_2^{(1)}(t) - x_2^{(2)}(t - \tau_{12}(t)), x_3^{(1)}(t) - x_3^{(2)}(t - \tau_{12}(t))\right) = \\
&= \left(-x_1^{(2)}(\theta), -x_2^{(2)}(\theta), -x_3^{(2)}(\theta)\right) = (-\rho_2 \cos\varphi_2 \cos\lambda_2, -\rho_2 \sin\varphi_2 \cos\lambda_2, -\rho_2 \sin\ \lambda_2); \\
&\vec{\xi}^{(13)} = (\xi_1^{(13)}, \xi_1^{(13)}, \xi_1^{(13)}) = \left(x_1^{(1)}(t) - x_1^{(3)}(t - \tau_{13}(t)), x_2^{(1)}(t) - x_2^{(3)}(t - \tau_{13}(t)), x_3^{(1)}(t) - x_3^{(3)}(t - \tau_{13}(t))\right) = \\
&= \left(-x_1^{(3)}(\vartheta), -x_2^{(3)}(\vartheta), -x_3^{(3)}(\vartheta)\right) = (-\rho_3 \cos\varphi_3 \cos\lambda_3, -\rho_3 \sin\varphi_3 \cos\lambda_3, -\rho_3 \sin\ \lambda_3);
\end{aligned}$$

$$\begin{aligned}
\tau_{12}(t) &= \tfrac{1}{c} \sqrt{\sum_{\alpha=1}^{3} \left[x_\alpha^{(1)}(t) - x_\alpha^{(2)}(t - \tau_{12}(t)\right]^2} = \tfrac{1}{c} \sqrt{\sum_{\alpha=1}^{3} \left[-x_\alpha^{(2)}(t - \tau_{12}(t)\right]^2} = \tfrac{\rho_2(t - \tau_{12}(t))}{c} = \tfrac{\rho_2(\theta)}{c}; \\
\tau_{21} &= \tfrac{1}{c} \sqrt{\left\langle \vec{\xi}^{(21)}, \vec{\xi}^{(21)} \right\rangle} = \tfrac{1}{c} \sqrt{\left\langle \vec{x}^{(2)}, \vec{x}^{(2)} \right\rangle} = \tfrac{\rho_2}{c}; \\
\tau_{13}(t) &= \tfrac{1}{c} \sqrt{\sum_{\alpha=1}^{3} \left[x_\alpha^{(1)}(t) - x_\alpha^{(3)}(t - \tau_{13}(t)\right]^2} = \tfrac{1}{c} \sqrt{\sum_{\alpha=1}^{3} \left[-x_\alpha^{(2)}(t - \tau_{13}(t)\right]^2} = \tfrac{\rho_3(t - \tau_{13}(t))}{c} = \tfrac{\rho_3(\vartheta)}{c};
\end{aligned}$$

$$\begin{aligned}
&\xi^{(21)} = \left(\xi_1^{(21)}, \xi_2^{(21)}, \xi_3^{(21)}, \xi_4^{(21)}\right) = \left(\vec{\xi}^{(21)}, \xi_4^{(21)}\right) = \left(x_1^{(2)}(t) - x_1^{(1)}(t - \tau_{21}(t)), x_2^{(2)}(t) - x_2^{(1)}(t - \tau_{21}(t)), x_3^{(2)}(t) - x_3^{(1)}(t - \tau_{21}(t)), ic\tau_{21}\right) = \\
&= \left(x_1^{(2)}(t), x_2^{(2)}(t), x_3^{(2)}(t), ic\tau_{21}\right); \\
&\xi^{(23)} = \left(\xi_1^{(23)}, \xi_2^{(23)}, \xi_3^{(23)}, \xi_4^{(23)}\right) = \left(\vec{\xi}^{(23)}, \xi_4^{(23)}\right) = \left(x_1^{(2)}(t) - x_1^{(3)}(t - \tau_{23}), x_2^{(2)}(t) - x_2^{(3)}(t - \tau_{23}), x_3^{(2)}(t) - x_3^{(3)}(t - \tau_{23}), ic\tau_{23}\right) = \\
&= (\rho_2 \cos\varphi_2 \cos\lambda_2 - \rho_3 \cos\varphi_3 \cos\lambda_3,\ \rho_2 \sin\varphi_2 \cos\lambda_2 - \rho_3 \sin\varphi_3 \cos\lambda_3,\ \rho_2 \sin\lambda_2 - \rho_3 \sin\lambda_3); \\
&c^2\tau_{23}{}^2 = (\rho_2 \cos\varphi_2 \cos\lambda_2 - \rho_3 \cos\varphi_3 \cos\lambda_3)^2 + (\rho_2 \sin\varphi_2 \cos\lambda_2 - \rho_3 \sin\varphi_3 \cos\lambda_3)^2 + (\rho_2 \sin\lambda_2 - \rho_3 \sin\lambda_3)^2 = \\
&= \rho_2{}^2 + \rho_3{}^2 - 2\rho_2\rho_3[\cos\lambda_2 \cos\lambda_3 \cos(\varphi_2 - \varphi_3) + \sin\lambda_2 \sin\lambda_3]; \\
&\xi^{(31)} = \left(\xi_1^{(31)}, \xi_2^{(31)}, \xi_3^{(31)}, \xi_4^{(31)}\right) = \left(\vec{\xi}^{(31)}, \xi_4^{(31)}\right) = \left(x_1^{(3)}(t) - x_1^{(1)}(t - \tau_{31}(t)), x_2^{(3)}(t) - x_2^{(1)}(t - \tau_{31}(t)), x_3^{(3)}(t) - x_3^{(1)}(t - \tau_{31}(t)), ic\tau_{31}\right) = \\
&= \left(x_1^{(3)}(t), x_2^{(3)}(t), x_3^{(3)}(t), ic\tau_{31}\right); \\
&\xi^{(32)} = \left(\xi_1^{(32)}, \xi_2^{(32)}, \xi_3^{(32)}, \xi_4^{(32)}\right) = \left(\vec{\xi}^{(32)}, \xi_4^{(32)}\right) = \left(x_1^{(3)}(t) - x_1^{(2)}(t - \tau_{32}), x_2^{(3)}(t) - x_2^{(2)}(t - \tau_{32}), x_3^{(3)}(t) - x_3^{(2)}(t - \tau_{32}), ic\tau_{32}\right) = \\
&= (\rho_3 \cos\varphi_3 \cos\lambda_3 - \rho_2 \cos\varphi_2 \cos\lambda_2,\ \rho_3 \sin\varphi_3 \cos\lambda_3 - \rho_2 \sin\varphi_2 \cos\lambda_2,\ \rho_3 \sin\lambda_3 - \rho_2 \sin\lambda_2); \\
&c^2\tau_{32}{}^2 = (\rho_3 \cos\varphi_3 \cos\lambda_3 - \rho_2 \cos\varphi_2 \cos\lambda_2)^2 + (\rho_3 \sin\varphi_3 \cos\lambda_3 - \rho_2 \sin\varphi_2 \cos\lambda_2)^2 + (\rho_3 \sin\lambda_3 - \rho_2 \sin\lambda_2)^2 = \\
&= \rho_2{}^2 + \rho_3{}^2 - 2\rho_2\rho_3[\cos\lambda_2 \cos\lambda_3 \cos(\varphi_2 - \varphi_3) + \sin\lambda_2 \sin\lambda_3].
\end{aligned}$$

But

$$\begin{aligned}&\cos\lambda_2\cos\lambda_3\cos(\varphi_2-\varphi_3)+\sin\lambda_2\sin\lambda_3=\\&=\tfrac{\cos(\lambda_2-\lambda_3)+\cos(\lambda_2+\lambda_3)}{2}\cos(\varphi_2-\varphi_3)+\tfrac{\cos(\lambda_2-\lambda_3)-\cos(\lambda_2+\lambda_3)}{2}=\\&=\tfrac{\cos(\lambda_2-\lambda_3)[1+\cos(\varphi_2-\varphi_3)]-[\cos(\lambda_2+\lambda_3)-\cos(\lambda_2+\lambda_3)\cos(\varphi_2-\varphi_3)]}{2}=\\&=\tfrac{\cos(\lambda_2-\lambda_3)2\cos^2\frac{\varphi_2-\varphi_3}{2}-\cos(\lambda_2+\lambda_3)2\sin^2\frac{\varphi_2-\varphi_3}{2}}{2}\le\cos^2\tfrac{\varphi_2-\varphi_3}{2}+\sin^2\tfrac{\varphi_2-\varphi_3}{2}=1.\end{aligned}$$

Therefore,

$$c^2\tau_{23}{}^2=\rho_2{}^2+\rho_3{}^2-2\rho_2\rho_3[\cos\lambda_2\cos\lambda_3\cos(\varphi_2-\varphi_3)+\sin\lambda_2\sin\lambda_3]\ge(\rho_2-\rho_3)^2\Leftrightarrow\tau_{23}\ge\frac{|\rho_2(t)-\rho_3(t-\tau_{23})|}{c}$$

and $\tau_{32}\ge\frac{|\rho_3(t)-\rho_2(t-\tau_{32})|}{c}$.

For the accelerations, we obtain $(n=2,3)$

$$\begin{aligned}&\dot u_1{}^{(n)}=c\tfrac{\dot\rho_n\cos\varphi_n\cos\lambda_n-\rho_n\dot\varphi_n\sin\varphi_n\cos\lambda_n\ -\rho_n\dot\lambda_n\cos\varphi_n\sin\lambda_n-c\cos\varphi_n\cos\lambda_n}{\rho_n}\approx-\tfrac{c^2\cos\varphi_n\cos\lambda_n}{\rho_n}\equiv Q_1{}^{(n)},\\&\dot u_2{}^{(n)}=c\tfrac{\dot\rho_n\sin\varphi_n\cos\lambda_n+\rho_n\dot\varphi_n\cos\varphi_n\cos\lambda_n-\rho_n\dot\lambda_n\sin\varphi_n\sin\lambda_n-c\sin\varphi_n\cos\lambda_n}{\rho_n}\approx-\tfrac{c^2\sin\varphi_n\cos\lambda_n}{\rho_n}\equiv Q_2{}^{(n)},\\&\dot u_3{}^{(n)}=c\tfrac{\dot\rho_n\sin\lambda_n+\rho_n\ \dot\lambda_n\cos\lambda_n-c\sin\ \lambda_n}{\rho_n}\approx-\tfrac{c^2\sin\lambda_n}{\rho_n}\equiv Q_3{}^{(n)}\ ,\ \ (n=2,3).\end{aligned}$$

Passing to the spherical coordinates, we obtain

$$\begin{aligned}&\ddot\rho_n\cos\varphi_n\cos\lambda_n-\rho_n\ddot\varphi_n\sin\varphi_n\cos\lambda_n-\rho_n\ddot\lambda_n\cos\varphi_n\sin\lambda_n=\\&=2\dot\rho_n\dot\varphi_n\sin\varphi_n\cos\lambda_n+2\dot\rho_n\dot\lambda_n\cos\varphi_n\sin\lambda_n-2\rho_n\dot\varphi_n\lambda_n\sin\varphi_n\sin\lambda_n+\rho_n\dot\varphi_n{}^2\cos\varphi_n\cos\lambda_n+\rho_n\dot\lambda_n{}^2\cos\varphi_n\cos\lambda_n-\tfrac{c^2\cos\varphi_n\cos\lambda_n}{\rho_n};\\&\ddot\rho_n\sin\varphi_n\cos\lambda_n+\rho_n\ddot\varphi_n\cos\varphi_n\cos\lambda_n-\rho_n\ddot\lambda_n\sin\varphi_n\sin\lambda_n=\\&=-2\dot\rho_n\dot\varphi_n\cos\varphi_n\cos\lambda_n+2\dot\rho_n\dot\lambda_n\sin\varphi_n\sin\lambda_n+2\rho_n\dot\varphi_n\dot\lambda_n\cos\varphi_n\ \sin\lambda_n+\rho_n\dot\varphi_n{}^2\sin\varphi_n\ \cos\lambda_n+\rho_n\dot\lambda_n{}^2\sin\varphi_n\cos\lambda_n-\tfrac{c^2\sin\varphi_n\cos\lambda_n}{\rho_n};\\&\ddot\rho_n\sin\lambda_n+\rho_n\ddot\lambda_n\cos\lambda_n=-2\dot\rho_n\dot\lambda_n\cos\lambda_n+\rho_n\ \dot\lambda_n{}^2\sin\lambda_n-\tfrac{c^2\sin\ \lambda_n}{\rho_n}.\end{aligned}\tag{A1}$$

We introduce denotations:

$$\begin{aligned}P_1{}^{(n)}&=2\dot\rho_n\dot\varphi_n\sin\varphi_n\cos\lambda_n+2\dot\rho_n\dot\lambda_n\cos\varphi_n\sin\lambda_n-2\rho_n\dot\varphi_n\dot\lambda_n\sin\varphi_n\sin\lambda_n+\rho_n\dot\varphi_n{}^2\cos\varphi_n\cos\lambda_n+\rho_n\dot\lambda_n{}^2\cos\varphi_n\cos\lambda_n;\\P_2{}^{(n)}&=-2\dot\rho_n\dot\varphi_n\cos\varphi_n\cos\lambda_n+2\dot\rho_n\dot\lambda_n\sin\varphi_n\sin\lambda_n+2\rho_n\dot\varphi_n\dot\lambda_n\cos\varphi_n\ \sin\lambda_n+\rho_n\dot\varphi_n{}^2\sin\varphi_n\ \cos\lambda_n+\rho_n\dot\lambda_n{}^2\sin\varphi_n\cos\lambda_n;\\P_3{}^{(n)}&=-2\dot\rho_n\dot\lambda_n\cos\lambda_n+\rho_n\ \dot\lambda_n{}^2\sin\lambda_n.\end{aligned}$$

Then, (A1) takes the following form:

$$\begin{aligned}&\ddot\rho_n\cos\varphi_n\cos\lambda_n-\ddot\varphi_n\rho_n\sin\varphi_n\cos\lambda_n-\ddot\lambda_n\rho_n\cos\varphi_n\sin\lambda_n=P_1{}^{(n)}+Q_1{}^{(n)}\\&\ddot\rho_n\sin\varphi_n\cos\lambda_n+\ddot\varphi_n\rho_n\cos\varphi_n\ \cos\lambda_n-\ddot\lambda_n\rho_n\sin\varphi_n\sin\lambda_n=P_2{}^{(n)}+Q_2{}^{(n)};\\&\ddot\rho_n\sin\lambda_n+\ \ddot\lambda_n\rho_n\cos\lambda_n=P_3{}^{(n)}+Q_3{}^{(n)}.\end{aligned}$$

To solve the last system with respect to $\ddot\rho_n$, $\ddot\varphi_n$, $\ddot\lambda_n$, we notice that

$$\begin{vmatrix}\cos\varphi_n\cos\lambda_n & -\rho_n\sin\varphi_n\cos\lambda_n & -\rho_n\cos\varphi_n\sin\lambda_n\\ \sin\varphi_n\cos\lambda_n & \rho_n\cos\varphi_n\ \cos\lambda_n & -\rho_n\sin\varphi_n\sin\lambda_n\\ \sin\lambda_n & 0 & \rho_n\cos\lambda_n\end{vmatrix}=\rho_n{}^2\cos\lambda_n>0$$

and then, we obtain the initial equations $(n=2,3)$:

$$\begin{aligned}&\ddot\rho_n=\rho_n\dot\varphi_n{}^2\cos^2\lambda_n+\rho_n\dot\lambda_n{}^2-\tfrac{c^2}{\rho_n},\\&\ddot\varphi_n=2\dot\varphi_n\dot\lambda_n tg\lambda_n-\tfrac{2\dot\rho_n\dot\varphi_n}{\rho_n},\\&\ddot\lambda_n=-\tfrac{2\dot\rho_n\dot\lambda_n+\rho_n\dot\varphi_n{}^2\sin\lambda_n\ \cos\lambda_n}{\rho_n}.\end{aligned}$$

Appendix B. Final Form of Basic Equations

First, we calculate $\left\langle \dot{\vec{u}}^{(n)}, \dot{\vec{u}}^{(n)} \right\rangle = \ddot{\rho}_n{}^2 + \rho_n{}^2 \ddot{\varphi}_n{}^2 \cos^2\lambda_n + \rho_n{}^2 \ddot{\lambda}_n{}^2$;

$$\left\langle \vec{\xi}^{(23)}, \dot{\vec{u}}^{(3)} \right\rangle = [\rho_2 \cos(\varphi_2 - \varphi_3)\cos\lambda_2 \cos\lambda_3 + \rho_2 \sin\lambda_2 \sin\lambda_3 - \rho_3]\ddot{\rho}_3 +$$
$$+\rho_2\rho_3 \sin(\varphi_2 - \varphi_3)\cos\lambda_2 \cos\lambda_3 \ddot{\varphi}_3 + \rho_2\rho_3[\sin\lambda_2 \cos\lambda_3 - \cos(\varphi_2 - \varphi_3)\sin\lambda_3 \cos\lambda_2]\ddot{\lambda}_3;$$
$$\left\langle \vec{\xi}^{(32)}, \dot{\vec{u}}^{(2)} \right\rangle = [\rho_3 \cos(\varphi_3 - \varphi_2)\cos\lambda_3 \cos\lambda_2 + \rho_3 \sin\lambda_3 \sin\lambda_2 - \rho_2]\ddot{\rho}_2 +$$
$$+ \sin(\varphi_3 - \varphi_2)\cos\lambda_3 \cos\lambda_2 \rho_2\rho_3 \ddot{\varphi}_2 + [\sin\lambda_3 \cos\lambda_2) - \cos(\varphi_3 - \varphi_2)\sin\lambda_2 \cos\lambda_3]\rho_2\rho_3\ddot{\lambda}_2.$$

Introduce denotations

$$P_\alpha^{(2)} = G_\alpha^{(21)} + G_\alpha^{(23)} + G_\alpha^{(2)rad}; \quad P_\alpha^{(3)} = G_\alpha^{(31)} + G_\alpha^{(32)} + G_\alpha^{(3)rad}; \ (\alpha = 1,2,3).$$

Then, the system of equations of motion becomes

$$\begin{aligned}
\dot{u}_1^{(n)} &= \tfrac{c^2 - \left(u_1^{(n)}\right)^2}{c^2} P_1^{(n)} - \tfrac{u_1^{(n)} u_2^{(n)}}{c^2} P_2^{(n)} - \tfrac{u_1^{(n)} u_3^{(n)}}{c^2} P_3^{(n)} \equiv U_1^{(n)};\\
\dot{u}_2^{(n)} &= -\tfrac{u_1^{(n)} u_2^{(n)}}{c^2} P_1^{(n)} + \tfrac{c^2 - \left(u_2^{(n)}\right)^2}{c^2} P_2^{(n)} - \tfrac{u_2^{(n)} u_3^{(n)}}{c^2} P_3^{(n)} \equiv U_2^{(n)};\\
\dot{u}_3^{(n)} &= -\tfrac{u_1^{(n)} u_3^{(n)}}{c^2} P_1^{(n)} - \tfrac{u_2^{(n)} u_3^{(n)}}{c^2} P_2^{(n)} + \tfrac{c^2 - \left(u_3^{(n)}\right)^2}{c^2} P_3^{(n)} \equiv U_3^{(n)}, (n = 2,3).
\end{aligned}$$

But

$$\begin{aligned}
\dot{u}_1{}^{(n)} &\approx \ddot{\rho}_n \cos\varphi_n \cos\lambda_n - \rho_n \ddot{\varphi}_n \sin\varphi_n \cos\lambda_n - \rho_n \ddot{\lambda}_n \cos\varphi_n \sin\lambda_n\\
\dot{u}_2{}^{(n)} &\approx \ddot{\rho}_n \sin\varphi_n \cos\lambda_n + \rho_n \ddot{\varphi}_n \cos\varphi_n \cos\lambda_n - \rho_n \ddot{\lambda}_n \sin\varphi_n \sin\lambda_n\\
\dot{u}_3{}^{(n)} &\approx \ddot{\rho}_n \sin\lambda_n + \rho_n \ddot{\lambda}_n \cos\lambda_n
\end{aligned}$$

and since $\left\langle \vec{u}^{(n)}, \vec{u}^{(n)} \right\rangle / c^2 \approx 0$, we infer

$$\begin{aligned}
U_1^{(n)} &= \tfrac{c^2 - \left(u_1^{(n)}\right)^2}{c^2} P_1^{(n)} - \tfrac{u_1^{(n)} u_2^{(n)}}{c^2} P_2^{(n)} - \tfrac{u_1^{(n)} u_3^{(n)}}{c^2} P_3^{(n)} \approx P_1^{(n)};\\
U_2^{(n)} &= -\tfrac{u_1^{(n)} u_2^{(n)}}{c^2} P_1^{(n)} + \tfrac{c^2 - \left(u_2^{(n)}\right)^2}{c^2} P_2^{(n)} - \tfrac{u_2^{(n)} u_3^{(n)}}{c^2} P_3^{(n)} \approx P_2^{(n)};\\
U_3^{(n)} &= -\tfrac{u_1^{(n)} u_3^{(n)}}{c^2} P_1^{(n)} - \tfrac{u_2^{(n)} u_3^{(n)}}{c^2} P_2^{(n)} + \tfrac{c^2 - \left(u_3^{(n)}\right)^2}{c^2} P_3^{(n)} \approx P_3^{(n)}, (n = 2,3)
\end{aligned}$$

and therefore,

$$\begin{aligned}
&\ddot{\rho}_n \cos\varphi_n \cos\lambda_n - \rho_n \ddot{\varphi}_n \sin\varphi_n \cos\lambda_n - \rho_n \ddot{\lambda}_n \cos\varphi_n \sin\lambda_n = P_1^{(n)}\\
&\ddot{\rho}_n \sin\varphi_n \cos\lambda_n + \rho_n \ddot{\varphi}_n \cos\varphi_n \cos\lambda_n - \rho_n \ddot{\lambda}_n \sin\varphi_n \sin\lambda_n = P_2^{(n)}\\
&\ddot{\rho}_n \sin\lambda_n + \rho_n \ddot{\lambda}_n \cos\lambda_n = P_3^{(n)}.
\end{aligned}$$

Solving the above system with respect to $\ddot{\rho}_n$, $\ddot{\varphi}_n$, $\ddot{\lambda}_n$, we obtain

$$\begin{aligned}
&\ddot{\rho}_n = P_1^{(n)} \cos\varphi_n \cos\lambda_n + P_2^{(n)} \sin\varphi_n \cos\lambda_n + P_3^{(n)} \sin\lambda_n;\\
&\ddot{\varphi}_n = \tfrac{-P_1^{(n)} \sin\varphi_n + P_2^{(n)} \cos\varphi_n}{\rho_n \cos\lambda_n};\\
&\ddot{\lambda}_n = \tfrac{-P_1^{(n)} \cos\varphi_n \sin\lambda_n - P_2^{(n)} \sin\varphi_n \sin\lambda_n + P_3^{(n)} \cos\lambda_n}{\rho_n} (n = 2,3)
\end{aligned} \qquad . \qquad \text{(A2)}$$

Recall that $\varphi_{23} = \varphi_2(t) - \varphi_3(t-\tau_{23})$ and $\varphi_{32} = \varphi_2(t-\tau_{32}) - \varphi_3(t)$, and, in view of

$$G_\alpha^{(23)} \approx \frac{e_2 e_3}{m_2 c^3}\left(\frac{\xi_\alpha^{(23)}}{{\tau_{23}}^3}\frac{\left\langle \vec{\xi}^{(23)}, \dot{\vec{u}}^{(3)}\right\rangle}{c^3} - \frac{\dot{u}_\alpha^{(3)}}{c\tau_{23}}\right),\ G_\alpha^{(32)} \approx \frac{e_2 e_3}{m_3 c^3}\left(\frac{\xi_\alpha^{(32)}}{{\tau_{32}}^3}\frac{\left\langle \vec{\xi}^{(32)}, \dot{\vec{u}}^{(3)}\right\rangle}{c^3} - \frac{\dot{u}_\alpha^{(2)}}{c\tau_{32}}\right);$$

$$P_\alpha^{(2)} = G_\alpha^{(21)} + G_\alpha^{(23)} + G_\alpha^{(2)rad} \approx \frac{e_2 e_1}{m_2 c^4}\frac{\xi_\alpha^{(21)}}{{\tau_{21}}^3} + \frac{e_2 e_3}{m_2 c^3}\left(\frac{\xi_\alpha^{(23)}}{{\tau_{23}}^3}\frac{\left\langle \vec{\xi}^{(23)}, \dot{\vec{u}}^{(3)}\right\rangle}{c^3} - \frac{\dot{u}_\alpha^{(3)}}{c\tau_{23}}\right) - \frac{e_2^2}{m_2 c^3}\ddot{u}_\alpha^{(2)};$$

$$P_\alpha^{(3)} = G_\alpha^{(31)} + G_\alpha^{(32)} + G_\alpha^{(3)rad} \approx \frac{e_3 e_1}{m_3}\frac{\xi_\alpha^{(31)}}{c^4{\tau_{31}}^3} + \frac{e_2 e_3}{m_3 c^3}\left(\frac{\xi_\alpha^{(32)}}{{\tau_{32}}^3}\frac{\left\langle \vec{\xi}^{(32)}, \dot{\vec{u}}^{(2)}\right\rangle}{c^3} - \frac{\dot{u}_\alpha^{(2)}}{c\tau_{32}}\right) - \frac{e_3^2}{m_3 c^3}\ddot{u}_\alpha^{(3)}$$

$(\alpha = 1,2,3)$

we transform the right-hand sides of the equations of (A2).

Firstly,

$$\ddot{\rho}_2 = P_1^{(2)}\cos\varphi_2\cos\lambda_2 + P_2^{(2)}\sin\varphi_2\cos\lambda_2 + P_3^{(2)}\sin\lambda_2 =$$

$$= \cos\varphi_2\cos\lambda_2\left[\frac{e_2e_1}{m_2c^4}\frac{\xi_1^{(21)}}{{\tau_{21}}^3} + \frac{e_2e_3}{m_2c^3}\left(\frac{\xi_1^{(23)}}{{\tau_{23}}^3}\frac{\left\langle \vec{\xi}^{(23)}, \dot{\vec{u}}^{(3)}\right\rangle}{c^3} - \frac{\dot{u}_1^{(3)}}{c\tau_{23}}\right) - \frac{e_2^2}{m_2c^3}\ddot{u}_1^{(2)}\right] +$$

$$+ \sin\varphi_2\cos\lambda_2\left[\frac{e_2e_1}{m_2c^4}\frac{\xi_2^{(21)}}{{\tau_{21}}^3} + \frac{e_2e_3}{m_2c^3}\left(\frac{\xi_2^{(23)}}{{\tau_{23}}^3}\frac{\left\langle \vec{\xi}^{(23)}, \dot{\vec{u}}^{(3)}\right\rangle}{c^3} - \frac{\dot{u}_2^{(3)}}{c\tau_{23}}\right) - \frac{e_2^2}{m_2c^3}\ddot{u}_2^{(2)}\right] +$$

$$+ \sin\lambda_2\left[\frac{e_2e_1}{m_2c^4}\frac{\xi_3^{(21)}}{{\tau_{21}}^3} + \frac{e_2e_3}{m_2c^3}\left(\frac{\xi_3^{(23)}}{{\tau_{23}}^3}\frac{\left\langle \vec{\xi}^{(23)}, \dot{\vec{u}}^{(3)}\right\rangle}{c^3} - \frac{\dot{u}_3^{(3)}}{c\tau_{23}}\right) - \frac{e_2^2}{m_2c^3}\ddot{u}_3^{(2)}\right] =$$

$$= \frac{e_2e_1}{m_2c^4}\frac{\xi_1^{(21)}\cos\varphi_2\cos\lambda_2 + \xi_2^{(21)}\sin\varphi_2\cos\lambda_2 + \xi_3^{(21)}\sin\lambda_2}{{\tau_{21}}^3} + \frac{e_2e_3}{m_2c^4}\left(\xi_1^{(23)}\cos\varphi_2\cos\lambda_2 + \xi_2^{(23)}\sin\varphi_2\cos\lambda_2 + \xi_3^{(23)}\sin\lambda_2\right)\frac{\left\langle \vec{\xi}^{(23)}, \dot{\vec{u}}^{(3)}\right\rangle}{c^2{\tau_{23}}^3} -$$

$$- \frac{e_2e_3}{m_2c^3\tau_{23}}\frac{\dot{u}_1^{(3)}\cos\varphi_2\cos\lambda_2 + \dot{u}_2^{(3)}\sin\varphi_2\cos\lambda_2 + \dot{u}_3^{(3)}\sin\lambda_2}{c} - \frac{e_2^2}{m_2c^3}\left(\ddot{u}_1^{(2)}\cos\varphi_2\cos\lambda_2 + \ddot{u}_2^{(2)}\sin\varphi_2\cos\lambda_2 + \ddot{u}_3^{(2)}\sin\lambda_2\right) \approx$$

$$\approx \frac{e_2e_1}{m_2c}\frac{1}{{\rho_2}^2} - \frac{e_2e_3}{m_2}\frac{\rho_2 - \rho_3(\cos\varphi_{23}\cos\lambda_2\cos\lambda_3 + \sin\lambda_2\sin\lambda_3)}{c^3}\frac{\left\langle \vec{\xi}^{(23)}, \dot{\vec{u}}^{(3)}\right\rangle}{c^3{\tau_{23}}^3} -$$

$$- \frac{e_2e_3}{m_2}\frac{\ddot{\rho}_3(\cos\varphi_{23}\cos\lambda_2\cos\lambda_3 + \sin\lambda_2\sin\lambda_3) + \rho_3\ddot{\varphi}_3\sin\varphi_{23}\cos\lambda_2\cos\lambda_3 + \rho_3\dot{\lambda}_3(\sin\lambda_2\cos\lambda_3 - \cos\varphi_{23}\sin\lambda_3\cos\lambda_2)}{c^4\tau_{23}} - \frac{e_2^2}{m_2c^3}\dddot{\rho}_2.$$

For the second equation of (A2), we have

$$P_\alpha^{(2)} = G_\alpha^{(21)} + G_\alpha^{(23)} + G_\alpha^{(2)rad} \approx \frac{e_2e_1}{m_2c^4}\frac{x_\alpha^{(2)}}{{\tau_{21}}^3} + \frac{e_2e_3}{m_2c^3}\left(\frac{\xi_\alpha^{(23)}}{{\tau_{23}}^3}\frac{\left\langle \vec{\xi}^{(23)}, \dot{\vec{u}}^{(3)}\right\rangle}{c^3} - \frac{\dot{u}_\alpha^{(3)}}{c\tau_{23}}\right) - \frac{e_2^2}{m_2c^3}\ddot{u}_\alpha^{(2)}\ ;$$

$$\ddot{\varphi}_2 = \frac{-\sin\varphi_2 P_1^{(2)} + \cos\varphi_2 P_2^{(2)}}{\rho_2\cos\lambda_2} = \frac{1}{\rho_2\cos\lambda_2}\left[\frac{e_2e_1}{m_2c^4}\frac{-x_1^{(2)}\sin\varphi_2 + x_2^{(2)}\cos\varphi_2}{{\tau_{21}}^3} + \right.$$

$$\left. + \frac{e_2e_3}{m_2c^4}\left(\left(-\xi_1^{(23)}\sin\varphi_2 + \xi_2^{(23)}\cos\varphi_2\right)\frac{\left\langle \vec{\xi}^{(23)}, \dot{\vec{u}}^{(3)}\right\rangle}{c^2{\tau_{23}}^3} + \frac{\dot{u}_1^{(3)}\sin\varphi_2 - \dot{u}_2^{(3)}\cos\varphi_2}{\tau_{23}}\right) - \frac{e_2^2}{m_2c^3}\left(\ddot{u}_\alpha^{(2)}\sin\varphi_2 - \ddot{u}_2^{(2)}\cos\varphi_2\right)\right] = \tag{A3}$$

$$= \frac{1}{\rho_2\cos\lambda_2}\frac{e_2e_3}{m_2c^3}\left[\frac{\rho_3\sin\varphi_{23}\cos\lambda_3}{{\tau_{23}}^3}\frac{\left\langle \vec{\xi}^{(23)}, \dot{\vec{u}}^{(3)}\right\rangle}{c^3} + \frac{\ddot{\rho}_3\cos\lambda_3\sin\varphi_{23} + \rho_3\ddot{\varphi}_3\cos_3\cos\varphi_{23} - \rho_3\dot{\lambda}_3\sin\lambda_3\sin\varphi_{23}}{c\tau_{23}} - \dddot{\varphi}_2\rho_2\cos\lambda_2\right].$$

For the third equation of (A2), we have

$$\ddot{\lambda}_2 = \frac{-\cos\varphi_2 \sin\lambda_2}{\rho_2} P_1^{(2)} + \frac{-\sin\varphi_2 \sin\lambda_2}{\rho_2} P_2^{(2)} + \frac{\cos\lambda_2}{\rho_2} P_3^{(2)} =$$

$$= \frac{-\cos\varphi_2 \sin\lambda_2}{\rho_2} \left(\frac{e_2 e_1}{m_2 c^4} \frac{x_1^{(2)}}{\tau_{21}^{\ 3}} + \frac{e_2 e_3}{m_2 c^3} \left(\frac{\xi_1^{(23)}}{\tau_{23}^{\ 3}} \frac{\left\langle \vec{\xi}^{(23)}, \dot{\vec{u}}^{(3)} \right\rangle}{c^3} - \frac{\dot{u}_1^{(3)}}{c\tau_{23}} \right) - \frac{e_2^2}{m_2 c^3} \ddot{u}_1^{(2)} \right) +$$

$$+ \frac{-\sin\varphi_2 \sin\lambda_2}{\rho_2} \left(\frac{e_2 e_1}{m_2 c^4} \frac{x_2^{(2)}}{\tau_{21}^{\ 3}} + \frac{e_2 e_3}{m_2 c^3} \left(\frac{\xi_2^{(23)}}{\tau_{23}^{\ 3}} \frac{\left\langle \vec{\xi}^{(23)}, \dot{\vec{u}}^{(3)} \right\rangle}{c^3} - \frac{\dot{u}_2^{(3)}}{c\tau_{23}} \right) - \frac{e_2^2}{m_2 c^3} \ddot{u}_2^{(2)} \right) +$$

$$+ \frac{\cos\lambda_2}{\rho_2} \left(\frac{e_2 e_1}{m_2 c^4} \frac{x_3^{(2)}}{\tau_{21}^{\ 3}} + \frac{e_2 e_3}{m_2 c^3} \left(\frac{\xi_3^{(23)}}{\tau_{23}^{\ 3}} \frac{\left\langle \vec{\xi}^{(23)}, \dot{\vec{u}}^{(3)} \right\rangle}{c^3} - \frac{\dot{u}_3^{(3)}}{c\tau_{23}} \right) - \frac{e_2^2}{m_2 c^3} \ddot{u}_3^{(2)} \right) =$$

$$= \frac{e_2 e_3}{m_2 c^3} \frac{\rho_3 \cos\varphi_{23} \sin\lambda_2 \cos\lambda_3}{\tau_{23}^{\ 3}} \frac{\left\langle \vec{\xi}^{(23)}, \dot{\vec{u}}^{(3)} \right\rangle}{c^3 \rho_2} +$$

$$+ \frac{e_2 e_3}{m_2 c^3} \frac{\ddot{\rho}_3 (\cos\varphi_{23} \cos\lambda_3 \sin\lambda_2 - \sin\lambda_3 \cos\lambda_2) + \rho_3 \ddot{\varphi}_3 \sin\varphi_{23} \sin\lambda_2 \cos\lambda_3 - \rho_3 \dot{\lambda}_3 (\cos\varphi_{23} \sin\lambda_2 \sin\lambda_3 + \cos\lambda_2 \cos\lambda_3}{c\tau_{23}\rho_2} - \frac{e_2^2}{m_2 c^3} \dddot{\lambda}_2.$$

Appendix C

Since, in the Main Theorem, $\left| r_0^{\ (m)} \right| (m = 1, 2, \ldots)$ should be sufficiently small, we take $\left| r_0^{\ (m)} \right| = 1/\mu (m = 1, 2, \ldots)$, then,

$$|r_n(t)| = \left| r_{n0} + \int_{T_k}^{t} r_n^{\ (1)}(t_1) dt_1 \right| = \left| r_{n0} + \int_{T_k}^{t} \left(r_{n0}^{\ (1)} + \int_{T_k}^{t_1} r_n^{\ (2)}(t_2) dt_2 \right) dt_1 \right| = \ldots =$$

$$= \left| \int_{T_k}^{t} r_{n0}^{\ (1)} dt_1 + \int_{T_k}^{t} \int_{T_k}^{t_1} r_{n0}^{\ (2)} dt_2 dt_1 + \ldots + \int_{T_k}^{t} \int_{T_k}^{t_1} \ldots \int_{T_k}^{t_{m-1}} r_n^{(m)}(t_m) dt_m \ldots dt_1 \right| \leq$$

$$\frac{\left| r_{n0}^{\ (1)} \right|}{\mu} e^{\mu(t-T_k)} + \frac{\left| r_{n0}^{\ (2)} \right|}{\mu^2} e^{\mu(t-T_k)} + \ldots + \frac{\omega^m}{\mu^m} R_n e^{\mu(t-T_k)} \leq$$

$$\leq \frac{1}{\mu^2} e^{\mu T} + \frac{1}{\mu^3} e^{\mu T} + \ldots + \frac{\omega^m}{\mu^m} R_n e^{\mu(t-T_k)} \approx \frac{\omega^m}{\mu^m} R_n e^{\mu(t-T_k)},$$

That means $|r_n(t)| \leq \frac{\omega^m}{\mu^m} R_n e^{\mu(t-T_k)} (n = 2, 3)$, because $\frac{1}{\mu} \approx 0$.

In the same way, $\left| r_n^{\ (1)}(t) \right| \leq \frac{\omega^m}{\mu^m} \omega R_n e^{\mu(t-T_k)}$, $\left| r_n^{\ (2)}(t) \right| \leq \frac{\omega^m}{\mu^m} \omega^2 R_n e^{\mu(t-T_k)}$, and so on.

Appendix D. Estimates for the Higher Derivatives

Assuming $\beta \approx 0$, we obtain

$$\frac{d\tau_{pq}(t)}{dt} = \frac{\left\langle \vec{\xi}^{(pq)}, \vec{u}^{(p)} \right\rangle - \left\langle \vec{\xi}^{(pq)}, \vec{u}^{(q)} \right\rangle}{c\sqrt{\left\langle \vec{\xi}^{(pq)}, \vec{\xi}^{(pq)} \right\rangle} - \left\langle \vec{\xi}^{(pq)}, \vec{u}^{(q)} \right\rangle} = \frac{\left\langle \vec{\xi}^{(pq)}, \vec{u}^{(p)} - \vec{u}^{(q)} \right\rangle}{c^2 \tau_{pq} - \left\langle \vec{\xi}^{(pq)}, \vec{u}^{(q)} \right\rangle} \approx \frac{2\beta}{1-\beta} < 1 \Leftrightarrow \beta = \frac{1}{137} < \frac{1}{3} < 1$$

$$\left| \frac{d\xi_2^{(23)}}{dt} \right| = \left| \frac{d\left[x_2^{(2)}(t) - x_2^{(3)}(t - \tau_{23}) \right]}{dt} \right| = \left| u_2^{(2)}(t) - u_2^{(3)}(t - \tau_{23})(1 - \dot{\tau}_{23}) \right| \leq 2\bar{c}.$$

Then,

$$\left\langle \vec{u}^{(n)}, \vec{u}^{(n)} \right\rangle = \dot{\rho}_n^{\ 2} + \rho_n^{\ 2} \dot{\varphi}_n^{\ 2} \cos^2\lambda_n + \rho_n^{\ 2} \dot{\lambda}_n^{\ 2} \leq \bar{c}^2 < c^2 ; \left\langle \vec{u}^{(2)}, \vec{u}^{(2)} \right\rangle / c^2 \approx 0;$$

$$\sqrt{\left\langle \dot{\vec{u}}^{(n)}, \dot{\vec{u}}^{(n)} \right\rangle} = \sqrt{\dot{r}_n^{\ 2} + \rho_n^{\ 2} \dot{\phi}_n^{\ 2} \cos^2\lambda_n + \rho_n^{\ 2} \dot{\eta}_n^{\ 2}} \leq \sqrt{\omega_n^{\ 2} (R_n^{\ 2} + \rho_n^{\ 2} \Phi_n^{\ 2} + \rho_n^{\ 2} Y_n^{\ 2}) e^{2\mu(t-T_k)}} \leq \omega_n \bar{c} e^{\mu(t-T_k)};$$

$$\sqrt{\left\langle \ddot{\vec{u}}^{(n)}, \ddot{\vec{u}}^{(n)} \right\rangle} = \sqrt{\ddot{r}_n^{\ 2} + \rho_n^{\ 2} \ddot{\underline{\phi}}_n^{\ 2} \cos^2\lambda_n + \rho_n^{\ 2} \ddot{\eta}_n^{\ 2}} \leq \omega_n^{\ 2} \bar{c} e^{\mu(t-T_k)};$$

$$\sqrt{\left\langle \dddot{\vec{u}}^{(n)}, \dddot{\vec{u}}^{(n)} \right\rangle} \leq \sqrt{\dddot{r}_n^{\ 2} + \rho_n^{\ 2} \dddot{\phi}_n^{\ 2} \cos^2\lambda_n + \rho_n^{\ 2} \dddot{\eta}_n^{\ 2}} \leq \omega_n^{\ 3} \bar{c} e^{\mu(t-T_k)}, \frac{1}{\tau_{23}} \leq \frac{c}{\Delta}; \frac{1}{\tau_{32}} \leq \frac{c}{\Delta}.$$

Furthermore,

$$\frac{d^2B_{r_2}^{(k)}}{dt^2}=\frac{dG_{r_2}}{dt}=\frac{d(P_1^{(2)}\cos\varphi_2\cos\lambda_2+P_2^{(2)}\sin\varphi_2\cos\lambda_2+P_3^{(2)}\sin\lambda_2)}{dt}=$$
$$=\dot{P}_1^{(2)}\cos\varphi_2\cos\lambda_2-P_1^{(2)}\phi_2\sin\varphi_2\cos\lambda_2-P_1^{(2)}\eta_2\cos\varphi_2\sin\lambda_2+$$
$$+\dot{P}_2^{(2)}\sin\varphi_2\cos\lambda_2+P_2^{(2)}\phi_2\cos\varphi_2\cos\lambda_2-P_2^{(2)}\eta_2\sin\varphi_2\sin\lambda_2+\dot{P}_3^{(2)}\sin\lambda_2+P_3^{(2)}\eta_2\cos\lambda_2=$$
$$=\frac{d}{dt}\left[\frac{e_2e_1}{m_2c^4}\frac{x_1^{(2)}}{\tau_{21}^{\ 3}}+\frac{e_2e_3}{m_2c^3}\left(\frac{\xi_1^{(23)}}{\tau_{23}^{\ 3}}\frac{\left\langle\vec{\xi}^{(23)},\vec{\dot{u}}^{(3)}\right\rangle}{c^3}-\frac{\dot{u}_1^{(3)}}{c\tau_{23}}\right)-\frac{e_2^2}{m_2c^3}\ddot{u}_1^{(2)}\right]\cos\varphi_2\cos\lambda_2-$$
$$-\left[\frac{e_2e_1}{m_2c^4}\frac{x_1^{(2)}}{\tau_{21}^{\ 3}}+\frac{e_2e_3}{m_2c^3}\left(\frac{\xi_1^{(23)}}{\tau_{23}^{\ 3}}\frac{\left\langle\vec{\xi}^{(23)},\vec{\dot{u}}^{(3)}\right\rangle}{c^3}-\frac{\dot{u}_1^{(3)}}{c\tau_{23}}\right)-\frac{e_2^2}{m_2c^3}\ddot{u}_1^{(2)}\right](\phi_2\sin\varphi_2\cos\lambda_2+\eta_2\cos\varphi_2\sin\lambda_2)+$$
$$+\frac{d}{dt}\left[\frac{e_2e_1}{m_2c^4}\frac{x_2^{(2)}}{\tau_{21}^{\ 3}}+\frac{e_2e_3}{m_2c^3}\left(\frac{\xi_2^{(23)}}{\tau_{23}^{\ 3}}\frac{\left\langle\vec{\xi}^{(23)},\vec{\dot{u}}^{(3)}\right\rangle}{c^3}-\frac{\dot{u}_2^{(3)}}{c\tau_{23}}\right)-\frac{e_2^2}{m_2c^3}\ddot{u}_2^{(2)}\right]\sin\varphi_2\cos\lambda_2+$$
$$+\left[\frac{e_2e_1}{m_2c^4}\frac{x_2^{(2)}}{\tau_{21}^{\ 3}}+\frac{e_2e_3}{m_2c^3}\left(\frac{\xi_2^{(23)}}{\tau_{23}^{\ 3}}\frac{\left\langle\vec{\xi}^{(23)},\vec{\dot{u}}^{(3)}\right\rangle}{c^3}-\frac{\dot{u}_2^{(3)}}{c\tau_{23}}\right)-\frac{e_2^2}{m_2c^3}\ddot{u}_2^{(2)}\right](\phi_2\cos\varphi_2\cos\lambda_2-\eta_2\sin\varphi_2\sin\lambda_2)+$$
$$+\frac{d}{dt}\left[\frac{e_2e_1}{m_2c^4}\frac{x_3^{(2)}}{\tau_{21}^{\ 3}}+\frac{e_2e_3}{m_2c^3}\left(\frac{\xi_3^{(23)}}{\tau_{23}^{\ 3}}\frac{\left\langle\vec{\xi}^{(23)},\vec{\dot{u}}^{(3)}\right\rangle}{c^3}-\frac{\dot{u}_3^{(3)}}{c\tau_{23}}\right)-\frac{e_2^2}{m_2c^3}\ddot{u}_3^{(2)}\right]\sin\lambda_2+$$
$$+\left[\frac{e_2e_1}{m_2c^4}\frac{x_3^{(2)}}{\tau_{21}^{\ 3}}+\frac{e_2e_3}{m_2c^3}\left(\frac{\xi_3^{(23)}}{\tau_{23}^{\ 3}}\frac{\left\langle\vec{\xi}^{(23)},\vec{\dot{u}}^{(3)}\right\rangle}{c^3}-\frac{\dot{u}_3^{(3)}}{c\tau_{23}}\right)-\frac{e_2^2}{m_2c^3}\ddot{u}_3^{(2)}\right]\eta_2\sin\lambda_2\le$$
$$\le\frac{e_2e_1}{m_2c^4}\frac{u_1^{(2)}}{\tau_{21}^{\ 3}}-3\frac{e_2e_1}{m_2c^4}\frac{x_1^{(2)}\dot{\tau}_{21}}{\tau_{21}^{\ 4}}+$$
$$+\frac{e_2e_3}{m_2c^3}\left(\frac{\dot{\xi}_{\cdot1}^{(23)}}{\tau_{23}^{\ 3}}\frac{\left\langle\vec{\xi}^{(23)},\vec{\dot{u}}^{(3)}\right\rangle}{c^3}-\frac{3\xi_1^{(23)}\dot{\tau}_{23}}{\tau_{23}^{\ 4}}\frac{\left\langle\vec{\xi}^{(23)},\vec{\dot{u}}^{(3)}\right\rangle}{c^3}+\frac{\xi_1^{(23)}}{\tau_{23}^{\ 3}}\frac{\left\langle\vec{\dot{\xi}}^{(23)},\vec{\dot{u}}^{(3)}\right\rangle+\left\langle\vec{\xi}^{(23)},\vec{\ddot{u}}^{(3)}\right\rangle}{c^3}-\frac{\ddot{u}_1^{(3)}}{c\tau_{23}}+\frac{\dot{u}_1^{(3)}\dot{\tau}_{23}}{c\tau_{23}^{\ 2}}\right)-\frac{e_2^2}{m_2c^3}\dddot{u}_1^{(2)}+$$
$$+\frac{|e_2e_1|}{m_2c}\left[\frac{1}{\breve{\rho}_2^{\ 2}}+\frac{\bar{c}\omega_3\left(\breve{\rho}_2+\breve{\rho}_3+3\Delta\right)}{c^3\Delta^2}+\frac{\omega_2^{\ 2}R_2}{c^2}\right](\Phi_2+\Upsilon_2)e^{\mu(t-T_k)}+$$
$$+\frac{e_2e_1}{m_2c^4}\frac{u_2^{(2)}}{\tau_{21}^{\ 3}}-3\frac{e_2e_1}{m_2c^4}\frac{x_2^{(2)}\dot{\tau}_{21}}{\tau_{21}^{\ 4}}+$$
$$+\frac{e_2e_3}{m_2c^3}\left(\frac{\dot{\xi}_{\cdot2}^{(23)}}{\tau_{23}^{\ 3}}\frac{\left\langle\vec{\xi}^{(23)},\vec{\dot{u}}^{(3)}\right\rangle}{c^3}-\frac{3\xi_2^{(23)}\dot{\tau}_{23}}{\tau_{23}^{\ 4}}\frac{\left\langle\vec{\xi}^{(23)},\vec{\dot{u}}^{(3)}\right\rangle}{c^3}+\frac{\xi_2^{(23)}}{\tau_{23}^{\ 3}}\frac{\left\langle\vec{\dot{\xi}}^{(23)},\vec{\dot{u}}^{(3)}\right\rangle+\left\langle\vec{\xi}^{(23)},\vec{\ddot{u}}^{(3)}\right\rangle}{c^3}-\frac{\ddot{u}_2^{(3)}}{c\tau_{23}}+\frac{\dot{u}_2^{(3)}\dot{\tau}_{23}}{c\tau_{23}^{\ 2}}\right)-\frac{e_2^2}{m_2c^3}\dddot{u}_2^{(2)}+$$
$$+\left[\frac{e_2e_1}{m_2c^4}\frac{x_2^{(2)}}{\tau_{21}^{\ 3}}+\frac{e_2e_3}{m_2c^3}\left(\frac{\xi_2^{(23)}}{\tau_{23}^{\ 3}}\frac{\left\langle\vec{\xi}^{(23)},\vec{\dot{u}}^{(3)}\right\rangle}{c^3}-\frac{\dot{u}_2^{(3)}}{c\tau_{23}}\right)-\frac{e_2^2}{m_2c^3}\ddot{u}_2^{(2)}\right](\Phi_2+\Upsilon_2)e^{\mu(t-T_k)}+$$
$$+\frac{e_2e_1}{m_2c^4}\frac{u_3^{(2)}}{\tau_{21}^{\ 3}}-3\frac{e_2e_1}{m_2c^4}\frac{x_3^{(2)}\dot{\tau}_{21}}{\tau_{21}^{\ 4}}+$$
$$+\frac{e_2e_3}{m_2c^3}\left(\frac{\dot{\xi}_{\cdot3}^{(23)}}{\tau_{23}^{\ 3}}\frac{\left\langle\vec{\xi}^{(23)},\vec{\dot{u}}^{(3)}\right\rangle}{c^3}-\frac{3\xi_3^{(23)}\dot{\tau}_{23}}{\tau_{23}^{\ 4}}\frac{\left\langle\vec{\xi}^{(23)},\vec{\dot{u}}^{(3)}\right\rangle}{c^3}+\frac{\xi_3^{(23)}}{\tau_{23}^{\ 3}}\frac{\left\langle\vec{\dot{\xi}}^{(23)},\vec{\dot{u}}^{(3)}\right\rangle+\left\langle\vec{\xi}^{(23)},\vec{\ddot{u}}^{(3)}\right\rangle}{c^3}-\frac{\ddot{u}_3^{(3)}}{c\tau_{23}}+\frac{\dot{u}_3^{(3)}\dot{\tau}_{23}}{c\tau_{23}^{\ 2}}\right)-\frac{e_2^2}{m_2c^3}\dddot{u}_3^{(2)}+$$
$$+\left[\frac{e_2e_1}{m_2c^4}\frac{x_3^{(2)}}{\tau_{21}^{\ 3}}+\frac{e_2e_3}{m_2c^3}\left(\frac{\xi_3^{(23)}}{\tau_{23}^{\ 3}}\frac{\left\langle\vec{\xi}^{(23)},\vec{\dot{u}}^{(3)}\right\rangle}{c^3}-\frac{\dot{u}_3^{(3)}}{c\tau_{23}}\right)-\frac{e_2^2}{m_2c^3}\ddot{u}_3^{(2)}\right]\Upsilon_2e^{\mu(t-T_k)};$$

$$\left|\frac{d^2B_{r_2}^{(k)}}{dt^2}\right|=\left|\frac{dG_{r_2}}{dt}\right|\le\frac{|e_2e_1|}{m_2}\left[\frac{\widetilde{c}}{c^4\Delta^3}+\frac{3\rho_2}{c^4\Delta^4}+\frac{8\omega_3\bar{c}}{c^4\Delta^2}+\frac{2\omega_3^{\ 2}\bar{c}}{c^4\Delta}+\frac{\omega_2^{\ 3}R_2}{c^3}\right]e^{\mu(t-T_k)}+\frac{|e_2e_1|}{m_2c}\left[\frac{1}{\breve{\rho}_2^{\ 2}}+\frac{\bar{c}\omega_3\left(\breve{\rho}_2+\breve{\rho}_3+3\Delta\right)}{c^3\Delta^2}+\frac{\omega_2^{\ 2}R_2}{c^2}\right](\Phi_2+\Upsilon_2)e^{\mu(t-T_k)}+$$
$$+\frac{|e_2e_1|}{m_2}\left[\frac{\bar{c}}{c^4\Delta^3}+\frac{3\rho_2}{c^4\Delta^4}+\frac{8\omega_3\bar{c}}{c^4\Delta^2}+\frac{2\omega_3^{\ 2}\bar{c}}{c^4\Delta}+\frac{\omega_2^{\ 3}R_2}{c^3}\right]e^{\mu(t-T_k)}+\frac{|e_2e_1|}{m_2c}\left[\frac{1}{\breve{\rho}_2^{\ 2}}+\frac{\bar{c}\omega_3\left(\breve{\rho}_2+\breve{\rho}_3+3\Delta\right)}{c^3\Delta^2}+\frac{\omega_2^{\ 2}R_2}{c^2}\right](\Phi_2+\Upsilon_2)e^{\mu(t-T_k)}+$$
$$+\frac{|e_2e_1|}{m_2}\left[\frac{\widetilde{c}}{c^4\Delta^3}+\frac{3\rho_2}{c^4\Delta^4}+\frac{8\omega_3\bar{c}}{c^4\Delta^2}+\frac{2\omega_3^{\ 2}\bar{c}}{c^4\Delta}\frac{\omega_2^{\ 3}R_2}{c^3}\right]e^{\mu(t-T_k)}+\frac{|e_2e_1|}{m_2c}\left[\frac{1}{\breve{\rho}_2^{\ 2}}+\frac{\bar{c}\omega_3\left(\breve{\rho}_2+\breve{\rho}_3+3\Delta\right)}{c^3\Delta^2}+\frac{\omega_2^{\ 2}R_2}{c^2}\right]\Upsilon_2e^{\mu(t-T_k)}\le\omega_2^{\ 2}R_2e^{\mu(t-T_k)}.$$

Appendix E. Upper Bounds for Partial Derivatives

Since $\varphi_{23} = \varphi_2(t) - \varphi_3(t-\tau_{23})$ and $\left|\left\langle \overset{\rightarrow}{\xi}^{(23)}, \overset{\cdot\rightarrow}{u}^{(3)} \right\rangle\right| \leq \bar{c}c\tau_{23}\omega_3 e^{\mu T}$, one obtains

$$\left|\frac{\partial G_{r_2}}{\partial \rho_2}\right| \leq \frac{|e_2 e_1|}{m_2}\left|\frac{2}{c\rho_2{}^3} + \frac{1}{c^3}\frac{\bar{c}c\tau_{23}\omega_3 e^{\mu(t-T_k)}}{c^3\tau_{23}{}^3}\right| \leq \frac{|e_2 e_1|}{m_2}\left(\frac{2}{c\breve{\rho}_2{}^3} + \frac{\bar{c}\omega_3}{c^5\Delta^2}\right)e^{\mu T} \equiv \frac{\partial \overline{G}_{r_2}}{\partial \rho_2};$$

$$\left|\frac{\partial G_{r_2}(r_2,\phi_2,\eta_2,r_3,\phi_3,\eta_3)}{\partial r_2}\right| = 0;\left|\frac{\partial G_{r_2}}{\partial \rho_3}\right| \leq \frac{|e_2 e_3|\omega_3}{m_2}\frac{\bar{c}+c(\Phi_3+Y_3)\Delta}{c^6\Delta^2}e^{\mu T} \equiv \frac{\partial \overline{G}_{r_2}}{\partial \rho_3};$$

$$\left|\frac{\partial G_{r_2}}{\partial \dot{r}_3}\right| \leq \frac{|e_2 e_3|}{m_2}\frac{2\left(\breve{\rho}_2+\breve{\rho}_3\right)^2+\Delta^2}{c^6\Delta^3} \equiv \frac{\partial \overline{G}_{r_2}}{\partial r_3};$$

$$\left|\frac{\partial G_{r_2}}{\partial \varphi_2}\right| \leq \frac{|e_2 e_3|}{m_2}\omega_3\bar{c}\left(\frac{\breve{\rho}_3 e^{\mu T}}{c^5\Delta^2} + \frac{\breve{\rho}_2+\breve{\rho}_3}{c^3} + \frac{\breve{\rho}_2}{c^3\Delta^3} + \frac{1}{c^4\Delta}\right) \equiv \frac{\partial \overline{G}_{r_2}}{\partial \varphi_2};$$

$$\left|\frac{\partial G_{r_2}}{\partial \varphi_3}\right| \leq \frac{|e_2 e_1|}{m_2}\left[\frac{\breve{\rho}_2+\breve{\rho}_3}{c^3}\frac{\bar{c}\omega_3 e^{\mu T}}{c^2\Delta^2} + \frac{\breve{\rho}_2+\breve{\rho}_3}{c^3} + \frac{\breve{\rho}_2\omega_3\bar{c}}{c^3\Delta^3} + \frac{\omega_3\bar{c}}{c^4\Delta}\right] \equiv \frac{\partial \overline{G}_{r_2}}{\partial \varphi_3};$$

$$\left|\frac{\partial G_{r_2}(r_2,\phi_2,\eta_2,r_3,\phi_3,\eta_3)}{\partial \phi_2}\right| = 0;\ \left|\frac{\partial G_{r_2}(r_2,\phi_2,\eta_2,r_3,\phi_3,\eta_3)}{\partial \phi_3}\right| = 0;$$

$$\left|\frac{\partial G_{r_2}}{\partial \lambda_2}\right| \leq \frac{|e_2 e_3|}{m_2}e^{\mu T}\bar{c}\omega_3\left(\frac{\breve{\rho}_2+\breve{\rho}}{c^3}\frac{\Delta+\breve{\rho}_2}{c^3\Delta^3} + \frac{1}{c^4\Delta}\right) = \frac{\partial \overline{G}_{r_2}}{\partial \lambda_2};\ \left|\frac{\partial G_{r_2}(r_2,\phi_2,\eta_2,r_3,\phi_3,\eta_3)}{\partial \eta_2}\right| = 0;$$

$$\left|\frac{\partial G_{r_2}}{\partial \lambda_3}\right| \leq \frac{|e_2 e_1|}{m_2}\frac{\bar{c}\omega_3}{c^4\Delta}\left(\frac{(\breve{\rho}_2+\breve{\rho}_3)(\Delta e^{\mu T}+2\breve{\rho}_2)}{c^2\Delta^2}+1\right) \equiv \frac{\partial \overline{G}_{r_2}}{\partial \lambda_3};$$

$$\left|\frac{\partial G_{r_2}(r_2,\phi_2,\eta_2,r_3,\phi_3,\eta_3)}{\partial \eta_3}\right| = 0;\left|\frac{\partial G_{r_2}}{\partial \dot{r}_2}\right| \leq \frac{e_2^2}{m_2 c^3} \equiv \frac{\partial \overline{G}_{r_2}}{\partial \dot{r}_2}.$$

The upper bounds for $\frac{\partial G_{r_3}}{\partial \rho_2},\ldots,\frac{\partial G_{r_3}}{\partial \dot{r}_3},\ldots,\frac{\partial G_{\eta_3}}{\partial \dot{r}_3}$ can be obtained in a similar way.

References

1. Angelov, V.G. The electromagnetic three-body problem with radiation terms—Derivation of equations of motion (I). *Results Nonlinear Anal.* **2020**, *3*, 45–58.
2. Angelov, V.G. The electromagnetic three-body problem with radiation terms—Existence-uniqueness of periodic orbit (II). *Results NonlinearAnal.* **2020**, *3*, 137–158.
3. Angelov, V.G. The N-body problem in classical electrodynamics. *Phys. Essays* **1993**, *6*, 204–211. [CrossRef]
4. Angelov, V.G. On the original Dirac equations with radiation term. *Lib. Math.* **2011**, *31*, 57–86.
5. Driver, R.D. A neutral system with state-dependent delay. *J. Diff. Equ.* **1984**, *54*, 73–86. [CrossRef]
6. Pauli, W. Relativitatstheorie. In *Encyklopedie der Mathematischen Wissenschaften*; Band. 5, Heft 4, Art. 19; Springer: Berlin/Heidelberg, Germany, 1921.
7. Synge, J.L. On the electromagnetic two-body problem. *Proc. Roy. Soc.* **1940**, *A177*, 118–139.
8. Angelov, V.G. Fixed point theorems for (Φ, j_1, j_2,..., j_m) contractive mappings in uniform spaces. *Bull. Allahabad Math. Soc.* **2023**, *2*. *in press*.
9. Bauer, G.; Deckert, D.-A.; Durr, D.; Hinrichs, G. On irreversibility and radiation in classical electrodynamics of point particles. *J. Stat. Phys.* **2014**, *154*, 610. [CrossRef]
10. Deckert, D.-A.; Durr, D.; Vona, N. Delay equations of the Wheeler-Feynman type. *J. Math. Sci.* **2014**, *202*, 623–636. [CrossRef]
11. Bauer, G.; Deckert, D.-A.; Durr, D. On the existence of dynamics in Wheeler-Feynman electromagnetism. *J. Appl. Math. Phys.* **2012**, *64*, 1087–1124. [CrossRef]
12. de Luca, J. Variational electrodynamics of atoms. *Prog. Electromagn. Res. B* **2013**, *53*, 1–11. [CrossRef]
13. Simaciu, C.C. Classical model of electron in stochastic electrodynamics. *Rev. Mex. Fis.* **2001**, *47*, 392–394.
14. Abouelmagd, E.I.; García Guirao, J.L.; Llibre, J. Periodic orbits of quantised restricted three-body problem. *Universe* **2023**, *9*, 149. [CrossRef]
15. Pan, S.; Hou, X. Analysis of resonance transition periodic orbits in the circular restricted three-body problem. *Appl. Sci.* **2022**, *12*, 8952. [CrossRef]
16. Abouelmagd, E.I.; Guirao, J.L.G.; Llibre, J. On the periodic orbits of the perturbed two- and three-body problems. *Galaxies* **2023**, *11*, 58. [CrossRef]
17. Abouelmagd, E.I.; Alhowaity, S.; Diab, Z.; Guirao, J.L.G.; Shehata, M.H. On the periodic solutions for the perturbed spatial quantized hill problem. *Mathematics* **2022**, *10*, 614. [CrossRef]
18. Cseh, J.; Dang, P.; Szilágyi, S.; Lévai, G. Orbital regularity of exoplanets and the symmetries of the kepler problem. *Symmetry* **2023**, *15*, 2114. [CrossRef]
19. Angelov, V.G. Periodic, transition and escape trajectories for 3D-Kepler 2-body problem of classical electrodynamics. *Phys. Sci. Int. J.* **2023**, *27*, 31–72. [CrossRef]
20. Angelov, V.G. *A Method for Analysis of Transmission Lines Terminated by Nonlinear Loads*; Nova Science: New York, NY, USA, 2014.
21. Pohl, R.W. *Optik und Atomphysik*; Springer: Berlin/Heidelberg, Germany, 1963.
22. Bohr, N. *Atomic Theory and Description of Nature*; Cambridge University Press: Teddington, UK, 1934.

23. Sommerfeld, A. *Atombau und Spektrallinien*; Friedrich Vieweg & Sohn, Braunschweig: Berlin, Germany, 1951.
24. Angelov, V.G. Spin three-body problem of classical electrodynamics with radiation terms (I)—Derivation of spin equations. *Results Nonlinear Anal.* **2021**, *4*, 1–20. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

AppliedMath

MDPI

Article

New Bivariate Copulas via Lomax Distribution Generated Distortions

Fadal Abdullah Ali Aldhufairi [1,*] and Jungsywan H. Sepanski [2]

[1] Department of Mathematics, King Khalid University, Asir-Abha 61421, KSA, Saudi Arabia
[2] Department of Statistics, Acturial & Data Sciences, Additional Central Michigan University, Mt. Pleasant, MI 48859, USA; sepan1jh@cmich.edu
* Correspondence: aldhu1fa@cmich.edu

Abstract: We develop a framework for creating distortion functions that are used to construct new bivariate copulas. It is achieved by transforming non-negative random variables with Lomax-related distributions. In this paper, we apply the distortions to the base copulas of independence, Clayton, Frank, and Gumbel copulas. The properties of the tail dependence coefficient, tail order, and concordance ordering are explored for the new families of distorted copulas. We conducted an empirical study using the daily net returns of Amazon and Google stocks from January 2014 to December 2023. We compared the popular Clayton, Gumbel, Frank, and Gaussian copula models to their corresponding distorted copula models induced by the unit-Lomax and unit-inverse Pareto distortions. The new families of distortion copulas are equipped with additional parameters inherent in the distortion function, providing more flexibility, and are demonstrated to perform better than the base copulas. After analyzing the data, we have found that the joint extremes of Amazon and Google stocks are more likely for high daily net returns than for low daily net returns.

Keywords: Archimedean; concordance; Clayton copula; distortion; Frank copula; Gumbel copula; Kendall's τ; Lomax distribution; tail dependence coefficient; tail order

check for updates

Citation: Aldhufairi, F.A.A.; Sepanski, J.H. New Bivariate Copulas via Lomax Distribution Generated Distortions. *AppliedMath* **2024**, *4*, 641–665. https://doi.org/10.3390/appliedmath4020035

Received: 29 March 2024
Revised: 7 May 2024
Accepted: 8 May 2024
Published: 17 May 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Descriptions and measurements of correlation and dependence between risks and losses have been important in various fields, such as finance, risk management, and actuarial studies. Bivariate copulas of different dependence structures can model the tail dependence in extreme risks Ref. [1] but be independent of the marginal distributions of the risks. Ref. [2] numerically and graphically illustrated the types of relationships that various copula-based measures of association can detect. Ref. [3] addressed the mathematics of copula functions illustrated with a finance application to financial topics in derivative pricing and credit risk analysis. The seminal paper by [4] demonstrated their practical applications, such as the estimation of joint-life mortality and multi-decrement models. Extreme or tail losses tend to occur together; see [5]. Ref. [6] employed copulas to study the effect of tail dependence and tailedness by quantifying extreme risks. Refs. [7,8] applied copula modeling to investigate the increasing hydroclimatic extremes associated with a warming climate; see also [9] for flood and hydrological models. Overall, copula modeling has shown to be an effective tool in analyzing dependent structures between variables.

Sklar's theorem [10] states that the joint cumulative distribution function (cdf) can be expressed as the product of the marginals and the copula, and conversely, the copula can be uniquely determined if we know the cdf and marginals. For instance, the Gaussian copula is derived from the bivariate Gaussian distribution and can also be used to generate new bivariate probability distributions via (A1) in the appendix; see [11] for summaries of the methods of constructing copulas. Necessary preliminaries on bivariate copulas corralled in Appendix A can be found in [11,12].

Most recently, Ref. [13] extended the traditional one-parameter Archimedean copulas by integrating the log-gamma-generated margins. Ref. [14] proposed a class of bivariate independence copula transformation of the form $C(u,v) = uvf((1-u),(1-v))$, where f is a twice differentiable. Let $\Pi(u,v) = uv$ be the independence copula. Generalizing the FGM copula given by $\pi(u,v) + \alpha xy(1-u)(1-v)$, Refs. [15,16] constructed new copulas of the form $\Pi(u,v) + Q(u,v;\alpha)$, where Q is a perturbation function involving trigonometric, hyperbolic, logarithmic, or exponential functions. Ref. [7] used a truncation of the log-concave half-logistic distribution function F as a multiplicative Archimedean generator to construct the copulas of the form $F^{-1}\big(F(u)F(v)\big)$.

In this paper, we are concerned with the construction of new bivariate copulas via distortion functions. A function T is called a distortion function if it is continuous and increasing on the unit interval $I = [0,1]$, with $T(0) = 0$ and $T(1) = 1$. A new family of copulas born of the distortion is given by

$$C_T(u,v) = T\big(C\big(T^{-1}(u), T^{-1}(v)\big)\big), u, v \in [0,1], \tag{1}$$

T is termed as admissible distortion if (1) is a copula. If the initial copula is Archimedean with generator ψ, then C_T is Archimedean with generator $\psi \circ T^{-1}$; see [17,18]. Theorem 3.3.3 in [11] (p. 96) shows that T is admissible if and only if T is increasing and convex; see also [19]. This key result dictates the convexity requirement and opens the door for explorations of admissible distortion functions. Ref. [20] showed that T is admissible if $T \circ \exp : (-\infty, 0) \to [0,1]$ is log-convex and suggested several distortion functions. Ref. [21] proposed to apply the distortion to the copula function only, marginals only, or both. The induced copula in (1) is a result of distortions to both the copula and marginals.

Refs. [22,23] constructed new families of copulas via beta and Kumaraswamy cdf distortions.Ref. [24] employed the unit-Lomax distortion. Ref. [25] studied families of copulas generated by a unit-Weibull distortion. Ref. [26] investigated the properties of unit-Gompertz distorted copulas and applied them to analyze the anthropocentric data. The unit-Lomax, unit-Weibull, and unit-Gompertz distortions are derived from an exponential transformation of the Lomax, Weibull, and Gompertz random variables.

Motivated by the fact that the cdf of a continuous random variable with unit interval support meets the definition of a distortion function, we propose a transformation that converts a non-negative random variable to one with unit interval support, which, consequently, establishes a distortion function. The distortion can then be used to generate new families of copulas by distorting existing ones. Similar to the other constructions of new copulas in the literature, the aim is to obtain new families of copulas that may account for a wider range of tail dependence values. With the parameters in the distortion cdf, the distorted copulas have additional parameters in addition to those in the existing copulas, making them more flexible.

The paper is organized as follows. Section 2 begins with the proposed mechanism for generating new distortion functions and admissible parameter spaces of distortions to be studied further. Section 3 provides Archimedean generators for the new families of distorted copulas when the base copulas are independence, Clayton, Frank, and Gumbel. The family of distorted independence copulas is presented to serve as a validation of the results obtained in this paper. Sections 4 and 5 investigate the properties of tail dependence, tail order, concordance order, and Kendall's tau. Section 6 contains the numerical results of a simulation study and empirical application, followed by concluding remarks. Copula preliminaries and derivations of tail orders and concordance ordering are included in the Appendices A and B.

2. Proposed Method

Let Y be a non-negative continuous random variable with cdf F. Consider the following transformation of the random variable Y:

$$X = \frac{Y}{\theta + Y}, \theta > 0. \tag{2}$$

The random variable X has a support of the unit interval, and its cdf is given by

$$G(x) = P(X \leq x) = P\left(Y \leq \frac{\theta x}{1-x}\right) = F\left(\frac{\theta x}{1-x}\right). \tag{3}$$

The cdf G and its quantile or inverse function may both serve as distortion functions to develop new copulas. If the variable Y assumes any value on the real number line, one may consider a transformation of its absolute value; that is, $|Y|/(1+|Y|)$, whose cdf is not as straightforward as (3).

We below demonstrate the method with F being the non-negative Lomax and inverse Lomax distributions, and derive the parameter space on which each of the distortions is convex. All the parameters in the generating distribution function F are assumed to be positive. The prime symbol, such as G' or G'', denotes the derivative of a function.

Example 1. *Unit-Lomax (UL) distortion and its quantile (QUL). Let Y be a Lomax or Pareto Type II random variable with a cdf given by $F(y) = 1 - [\beta/(\beta+y)]^{\alpha}$, where $y > 0$. In this case, the transformation in (2) produces the distortion given by*

$$G(x) = 1 - \left[\frac{\beta(1-x)}{\beta(1-x)+\theta x}\right]^{\alpha}, \tag{4}$$

where $0 < x < 1$. Note that, for example,

$$\left[\frac{2(1-x)}{2(1-x)+1x}\right]^{\alpha} = \left[\frac{4(1-x)}{4(1-x)+2x}\right] = \left[\frac{8(1-x)}{4(1-x)+4x}\right]^{\alpha}. \tag{5}$$

Equation (5) demonstrates that multiple values of the parameters (β, θ) give rise to the same distortion in (4), and thus the parameters cannot be uniquely identified. Therefore, to solve this problem, with reparametrization, we consider the distortion G_L (UL) and its inverse Q_L (QUL) given by

$$G_L(x) = 1 - \left[\frac{(1-x)}{(1-x)+\theta x}\right]^{\alpha}; \quad Q_L(x) = \frac{1}{1+\theta[(1-x)^{-1/\alpha}-1]^{-1}}.$$

Lemma 1. *The distortion $G_L(x) = 1 - \left[(1-x)\big((1-x)+\theta x\big)^{-1}\right]^{\alpha}$ is convex on I if $0 < \theta \leq 1$ and $0 < \alpha \leq 1$.*

Proof. Let $A(x) = (1-x)+\theta x$, $B(x) = (1-x)[A(x)]^{-1}$, and then $G_L(x) = 1 - [B(x)]^{\alpha}$. For simplicity, the argument notation of (x) is dropped. Note that $(B)^{-1}(A)^{-1} = (1-x)^{-1}$. The derivatives are given by

$$\begin{aligned} B' &= -\theta A^{-2}; \quad G_L' = \alpha\theta B^{\alpha-1}A^{-2} \\ G_L'' &= -\alpha(\alpha-1)\theta^2 B^{\alpha-2}A^{-4} - 2\alpha\theta B^{\alpha-1}A^{-3}(\theta-1) \\ &= -\alpha\theta B^{\alpha-1}A^{-3}\left[\frac{(\alpha-1)\theta}{1-x} + 2(\theta-1)\right]. \end{aligned} \tag{6}$$

If $0 < \theta \leq 1$ and $0 < \alpha \leq 1$, then $G_L''(x) \geq 0$ for all $x \in I$. □

Lemma 2. *The distortion $Q_L(x) = 1/\left[1+\theta\big((1-x)^{-1/\alpha}-1\big)^{-1}\right]$ is convex on I if $\theta \geq 1$ and $\alpha \geq 1$.*

Proof. Since Q_L is the inverse function of G_L, and G_L is an increasing function, by (6), we obtain this lemma. □

Example 2. *Unit-inverse Pareto (UIP) distortion and its quantile (QUP). Consider the inverse Pareto random variable, defined to be the reciprocal of a Lomax random variable, with a cdf given by* $F(y) = [y/(\beta + y)]^{\alpha}$, *where* $y > 0$. *In this case, (3) gives*

$$G(x) = \left[\frac{\theta x}{\beta(1-x) + \theta x}\right]^{\alpha},$$

where $0 < x < 1$. *For the same reason explained by (5), we propose the distortion* G_P *(UIP) and its inverse* Q_P *(QUP) given by*

$$G_P(x) = \left[\frac{\theta x}{(1-x) + \theta x}\right]^{\alpha}; \quad Q_P(x) = \frac{1}{1 + \theta(x^{-1/\alpha} - 1)}.$$

Lemma 3. *The distortion* $G_P(x) = \left[\frac{\theta x}{(1-x) + \theta x}\right]^{\alpha}$ *is convex on* I *if* $0 < \theta \leq 1$ *and* $0 < \alpha \leq 1$.

Proof. Let $A(x) = (1-x) + \theta x$ and $B(x) = \theta x[A(x)]^{-1}$. Then $[B(x)A(x)]^{-1} = 1/(\theta x)$ and $G_P(x) = [B(x)]^{\alpha}$. For simplicity, the argument notation of (x) is dropped. The relevant derivatives are given by

$$\begin{aligned} B' &= \theta(A)^{-1} - \theta x A^{-2}(\theta - 1) = \theta A^{-2}; \quad G_P' = \alpha\theta B^{\alpha-1}A^{-2} \\ G_P'' &= \alpha(\alpha-1)\theta^2 B^{\alpha-2}A^{-4} - 2\alpha\theta B^{\alpha-1}A^{-3}(\theta-1) \\ &= \alpha\theta B^{\alpha-1}A^{-3}x^{-1}[(\alpha-1) - 2(\theta-1)x]. \end{aligned} \tag{7}$$

If $0 < \theta \leq 1$ and $\alpha \geq 1$, then the second derivative $G_P'' \geq 0$ for all $x \in I$. □

Lemma 4. $Q_P(x) = [1 + \theta(x^{-1/\alpha} - 1)]^{-1}$ *is convex on* I *if* $\theta \geq 1$ *and* $0 < \alpha \leq 1$.

Proof. This lemma follows since Q_P is the inverse function of G_P and G_P is increasing and concave by (7) when $\theta \geq 1$ and $0 < \alpha \leq 1$. □

In summary, applying the proposed transformation in (2) to non-negative Lomax-related random variables, we are bestowed with four new admissible distortions tabulated in Table 1. Two functional forms are displayed as they would come in handy for calculations. Note that the admissible parameter spaces are only a sufficient condition for C_T in (1) to be a copula. The dual power distortion is a special case of G_L with $\theta = 1$. The power distortion is a special case of G_P with $\theta = 1$. When $\theta = \alpha = 1$, it is important to note that all the distortions in Table 1 become the identity function. This means that when these distortions are applied to a base copula, the resulting family of copulas offers a greater flexibility when fit to data because it includes the base copula as a special case.

Remark 1. *Let* Y *be a Lomax random variable. We also considered the cdf of* $Y^{1/\tau}$, *i.e., the Burr distribution, given by* $F(y) = 1 - (1 + y^{\tau})^{-\alpha}$, *as the generating cdf. However, there does not exist a parameter space on which the resulting distortion is convex. Another generating cdf candidate is the exponentiated Lomax cdf of the form* $\left[1 - F\left(\gamma/(\gamma + y)\right)^{\alpha}\right]^{\gamma}$, *which is more complex and will be investigated in the future.*

Remark 2. *Instead of (2), the transformation of* $X = 1/(1 + Y)$ *also gives rise to a distortion cdf. The cdf of* $X = 1/[1 + Y]$ *is given by, for* $0 < x < 1$,

$$G(x) = P(X \leq x) = 1 - F\left(Y \geq \frac{1 - X}{X}\right).$$

In this case, when Y *has a Lomax distribution in Example 1, we derive the* G_P. *When* Y *has a Lomax distribution in Example 2, we derive the* G_L.

Table 1. Proposed distortions and admissible parameter spaces.

Distortion	Function Form		Convex Parameter Space
U-Lomax (UL)	$G_L(x)$	$= 1 - \left[\frac{1-x}{1-x+\theta x}\right]^{\alpha}$	$0 < \theta \leq 1, 0 < \alpha \leq 1$
		$= 1 - \frac{1}{\{1+\theta[(1-x)^{-1}-1]\}^{\alpha}}$	
Quantile U-Lomax (QL)	$Q_L(x)$	$= \frac{(1-x)^{-1/\alpha}-1}{\theta+(1-x)^{-1/\alpha}-1}$	$\theta \geq 1, \alpha \geq 1$
		$= \frac{1}{1+\theta[(1-x)^{-1/\alpha}-1]^{-1}}$	
U-Inverse Pareto (UIP)	$G_P(x)$	$= \left[\frac{\theta x}{1-x+\theta x}\right]^{\alpha}$	$0 < \theta \leq 1, \alpha \geq 1.$
		$= \frac{1}{[1+\theta^{-1}(x^{-1}-1)]^{\alpha}}$	
Quantile U-Inverse Pareto (QUP)	$Q_P(x)$	$= \frac{x^{1/\alpha}}{x^{1/\alpha}+\theta(1-x^{1/\alpha})}$	$\theta \geq 1, 0 < \alpha \leq 1$
		$= \frac{1}{1+\theta(x^{-1/\alpha}-1)}$	

Remark 3. *There are a multitude of cdfs that may be used as the generating cdf F in (3). For instance, an exponential distribution with mean β. In this case, (3) gives a candidate distortion given by* $1 - \exp\left(-x/[\beta(1-x)]\right)$. *Another example, a generalized Pareto with a pdf given by*

$$f(y) = \frac{\Gamma(\alpha+\tau)}{\Gamma(\alpha)\Gamma(\tau)} \frac{\theta^{\tau} y^{\tau-1}}{(y+\theta)^{\alpha+\tau}}$$

where $x > 0$, $\alpha > 0$, $\tau > 0$, *and* $\theta > 0$.

3. Families of New Distorted Copulas

The general form of copulas constructed by a distortion T is in (1). Archimedean copulas are widely used [27] and are recognized for their flexibility in modeling dependence within multivariate random variables. We formulate the Archimedean generators resulting from the distortions in Section 3.1 when the base copulas are independence, Clayton, Frank, and Gumbel; see Table A1 in Appendix A. The base copulas, except the independence copula, have a single parameter, and the families of the distorted copulas contain more parameters and are less restrictive. Let r denote the parameter in the base copula.

3.1. Archimedean Generators

Example 3 (Independence Copula). *The Archimedean generators are*

- *UL-independence generator:* $\log\left(\theta\left[(1-u)^{-1/\alpha}-1\right]^{-1}+1\right)$.
- *QUL-independence generator:* $-\log\left(1-\left[1+\theta\left((1-u)^{-1}-1\right)\right]^{-\alpha}\right)$.
- *UIP-independence generator:* $\log\left(1+\theta(u^{-1/\alpha}-1)\right)$.
- *QUP-independence generator:* $\log\left(1+\theta^{-1}(u^{-1}-1)\right)^{\alpha}$.

Example 4 (Clayton Copula). *In this case, we obtain the following:*

- *UL-Clayton generator:* $\left\{\left[\theta\left((1-u)^{-1/\alpha}-1\right)^{-1}+1\right]^{-r}-1\right\}/r$.
- *QUL-Clayton generator:* $\left\{\left[1-\left(1+\theta((1-u)^{-1}-1)\right)^{-\alpha}\right]^{-r}-1\right\}/r$.
- *UIP-Clayton generator:* $\left\{\left[1+\theta\left(u^{-1/\alpha}-1\right)\right]^{-r}-1\right\}/r$.
- *QUP-Clayton generator:* $\left\{\left[1+\theta^{-1}(u^{-1}-1)\right]^{\alpha r}-1\right\}/r$.

Example 5 (Frank Copula). *When the base copula C is Frank, we obtain the following:*

- *UL-Frank generator:* $-\log\Big(1+\Big[e^{-r[\theta((1-u)^{-1/\alpha}-1)^{-1}+1]^{-1}}-1\Big]/(e^{-r}-1)\Big)$.
- *QUL-Frank generator:* $-\log\Big(1+\Big[e^{-r[1-(1+\theta((1-u)^{-1}-1))^{-\alpha}]}-1\Big]/(e^{-r}-1)\Big)$.
- *UIP-Frank generator:* $-\log\Big(1+\Big[e^{-r[1+\theta(u^{-1/\alpha}-1)]^{-1}}-1\Big]/(e^{-r}-1)\Big)$.
- *QUP-Frank generator:* $-\log\Big(1+\Big[e^{-r[1+\theta^{-1}(u^{-1}-1)]^{-\alpha}}-1\Big]/(e^{-r}-1)\Big)$.

Example 6 (Gumbel Copula). *For Gumbel copula, we obtain the following:*

- *UL-Gumbel generator:* $\Big[-\log\Big(\theta((1-u)^{-1/\alpha}-1)^{-1}+1\Big)\Big]^{r}$.
- *QUL-Gumbel generator:* $\big[-\log\big(1-(1+\theta((1-u)^{-1}-1))^{-\alpha}\big)\big]^{r}$.
- *UIP-Gumbel generator:* $\Big[\log\Big(1+\theta(u^{-1/\alpha}-1)\Big)\Big]^{r}$.
- *QUP-Gumbel generator:* $\Big[\log\big(1+\theta^{-1}(u^{-1}-1)\big)^{\alpha}\Big]^{r}$.

3.2. Distortions of Independence Copula

We stage here the copula families constructed by distortions of the independent copula aiming to demonstrate the versatility of distortion and validate the results presented in this paper. The distortions proposed distort the parameter-free independent copula into diverse families of copulas adorned with two parameters and better flexibility.

Note that $G_L^{-1}=Q_L$; see Table 1. For the UL-independence copula, $C(u,v)=uv$,

$$\begin{aligned} C_{G_L}(u,v) &= 1-\left(\frac{1}{1+\theta\big\{\big[1-C\big(G_L^{-1}(u),G_L^{-1}(u)\big)\big]^{-1}-1\big\}}\right)^{\alpha} \\ &= 1-\left(\frac{\theta+(1-u)^{-1/\alpha}+(1-v)^{-1/\alpha}-2}{\theta+(1-u)^{-1/\alpha}(1-v)^{-1/\alpha}-1}\right)^{\alpha}. \end{aligned} \tag{8}$$

When $\alpha=\theta=1$, $C_{G_L}(u,v)=uv$. Let $\bar{u}=(1-u)$ and $\bar{v}=(1-v)$. When $\theta=1$,

$$C_{G_L}(u,v) = 1-(\bar{u}^{1/\alpha}+\bar{v}^{1/\alpha}-\bar{u}^{1/\alpha}\bar{v}^{1/\alpha})^{\alpha}, \tag{9}$$

which is the Joe or B5 copula. The Joe copula is given by $C(u,v)=1-\big((1-u)^r+(1-v)^r-(1-u)^r(1-v)^r\big), r\in[1,\infty)$. It has $\kappa_L=2$ and $\lambda_U=2-2^{\alpha}, 0<\alpha\leq 1$, which endorses Proposition 1 in Section 4. The Joe copula in (9) is negatively ordered by α. When $\alpha=1$, $C_{G_L}(u,v)=uv/[1-(1-\theta)\bar{u}\bar{v}]$, which is the Ali–Mikhail–Haq (AMH) copula [11] and is negatively ordered by the parameter θ. The AMH copula family has lower and upper tail dependence coefficients of zero, and is given by $C(u,v)=uv/[1-r(1-u)(1-v)], r\in[-1,1]$.

For the QUL-independence copula, from Table 1,

$$\begin{aligned} C_{Q_L}(u,v) &= 1-\frac{1}{\theta+[1-C\big(Q_L^{-1}(u),Q_L^{-1}(u)\big)]^{-1/\alpha}-1}, \text{ where} \\ C\big(Q_L^{-1}(u),Q_L^{-1}(u)\big) &= \left(1-\left[\frac{1-u}{1-u+\theta u}\right]^{\alpha}\right)\left(1-\left[\frac{1-v}{1-v+\theta v}\right]^{\alpha}\right). \end{aligned} \tag{10}$$

When $\theta=1$, (10) produces the Joe copula in (9) with a different parametrization from the one in [12]. When $\alpha=1$, (10) yields

$$1-\frac{\bar{u}\bar{v}+\theta u\bar{v}+\theta\bar{u}v}{\theta\big(\bar{u}\bar{v}+\theta u\bar{v}+\theta\bar{u}v+\theta uv\big)}.$$

For the UIP-independence copula, from Table 1,

$$C_{G_P}(u,v) = \left(\frac{1}{1+\theta^{-1}\{[C(G_P^{-1}(u),G_P^{-1}(u))]^{-1}-1\}}\right)^{\alpha} \tag{11}$$

$$= \left[1+(u^{-1/\alpha}-1)+(v^{-1/a}-1)+\theta(u^{-1/\alpha}-1)(v^{-1/\alpha}-1)\right]^{-\alpha}$$

$$= uv\{1-(1-\theta)[(1-u^{1/\alpha})(1-v^{1/a})]\}^{-\alpha}, \tag{12}$$

which is the BB10 copula [12] with $\kappa_L = \kappa_U = 2$. The BB10 copula is given by $C(u,v) = uv[1-r_1(1-u^{1/r_2})(1-v^{1/r_2})]^{-r_2}$, where $r_1 \in [0,1)$ and $r_2 > 0$. As shown in Proposition 3, UIP-distorted copulas have the same upper and lower tail orders as the base copula.

For the QUP-independence copula, from Table 1,

$$C_{Q_P}(u,v) = \frac{1}{1+\theta[C(G_P^{-1}(u),G_P^{-1}(u))^{-1/\alpha}-1]} \tag{13}$$

$$= \left[1+\frac{(1-u)(1-v)+\theta u(1-v)+\theta v(1-u)}{\theta uv}\right]^{-1}$$

$$= uv\left[1-\frac{\theta-1}{\theta}(1-u)(1-v)\right]^{-1},$$

which is the AMH copula or a special case of the BB10 copula.

4. Tail Dependence Coefficients and Tail Orders

In this section, we investigate the tail dependence coefficients, tail orders, and concordances for the new families of copulas emerging from the four distortions in Table 1. The lengthy derivations of tail orders are stationed in Appendix B.

Let $t(u) = dT(u)/du$ and assume that the lower tail dependence (ltd) coefficient λ_L of the base copula C and $\lim_{\tau\to 0+} t\big(C(u,u)\big)/t(u)$ exist. By definition in (A3) and L'Hopital's rule, the ltd coefficient for a T distortion-induced copula is given by

$$\lambda_{T,L} = \lim_{u\to 0^+}\frac{T\big(C(T^{-1}(u),T^{-1}(u))\big)}{u}$$

$$= \lim_{u\to 0^+}\frac{T\big(C(u,u)\big)}{T(u)} = \lim_{u\to 0^+}\frac{t\big(C(u,u)\big)}{t(u)}\frac{dC(u,u)}{du}. \tag{14}$$

Since $\lim_{u\to 1^-} T(u) = 1$, with the substitution of $v = T^{-1}(u)$, the upper tail dependence (utd) coefficient of C_T is given by

$$\lambda_{T,U} = 2-\lim_{u\to 1^-}\frac{1-T\big(C(T^{-1}(u),T^{-1}(u))\big)}{1-u}$$

$$= 2-\lim_{v\to 1^-}\frac{1-T(C(v,v))}{1-T(v)} = 2-\lim_{v\to 1^-}\frac{t\big(C(v,v)\big)}{t(v)}\frac{dC(v,v)}{dv}. \tag{15}$$

Below, we assume that the ltd coefficient $\lambda_L = 0$ when $\kappa_L > 1$ and the utd coefficient $\lambda_U = 0$ when $\kappa_U > 1$ for the base copula C. Furthermore, we assume that $C(u,u) \sim u^{\kappa_L}\ell(u)$ as $u \to 0^+$ and $\overline{C}(1-u,1-u) \sim u^{\kappa_U}\ell_*(u)$ as $u \to 0^+$ for some slowly varying functions ℓ and ℓ_* at 0^+. Let the subscript denote a property owner, e.g., the subscript T in $\lambda_{T,U}$ is used to denote the ltd coefficient of a $T-$distorted copula.

Proposition 1 (Unit-Lomax Distortion). *Let $C_{G_L}(u,v)$ be the G_L-distorted copula defined in (8), where $0 < \theta \le 1$ and $0 < \alpha \le 1$. Then,*

(i) *$\kappa_{G_L,L} = \kappa_L$ and $\lambda_{G_L,L} = \lambda_L$.*

(ii) *$\kappa_{G_L,U} = \kappa_U$ when $\alpha = 1$ and $\kappa_{G_L,U} = 1$ when $0 < \alpha < 1$. And $\lambda_{G_L,U} = 2-(2-\lambda_U)^{\alpha}$.*

Proof. The tail orders are shown in (A9) and (A17) in Appendix B. The ltd coefficient, by (14) and L'Hopital's rule, is given by

$$\begin{aligned}\lambda_{G_L,L} &= \lim_{u\to 0^+}\left\{1-\left(\frac{1-C(u,u)}{1-C(u,u)+\theta C(u,u)}\right)^{\alpha}\right\}\Big/\left\{1-\left[\frac{1-u}{1-u+\theta u}\right]^{\alpha}\right\}\\ &= \lim_{u\to 0^+}\left\{\frac{[1-C(u,u)]}{[1-C(u,u)+\theta C(u,u)]}\frac{[(1-u)+\theta u]}{(1-u)}\right\}^{\alpha-1}\\ &\qquad\times\frac{C'(u,u)}{[1-C(u,u)+\theta C(u,u)]^2}\frac{[(1-u)+\theta u]^2}{1}=\lambda_L\end{aligned}$$

since $\lim_{u\to 0^+} C(u,u)/u = \lim_{u\to 0^+} dC(u,u)/du = \lambda_L$. The utd coefficient, by (15) and L'Hopital's rule, we obtain that

$$\begin{aligned}\lambda_{G_L,U} &= 2-\lim_{u\to 1-}\left\{\frac{1-C(u,u)}{1-C(u,u)+\theta C(u,u)}\right\}^{\alpha}\Big/\left[\frac{1-u}{1-u+\theta u}\right]^{\alpha}\\ &= 2-\lim_{u\to 1-}\left\{\frac{1-u+\theta u}{[1-C(u,u)+\theta C(u,u)]}\right\}^{\alpha}\left[\frac{1-C(u,u)}{1-u}\right]^{\alpha}=2-(2-\lambda_U)^{\alpha}\end{aligned}$$

since $\lim_{u\to 1-}[1-C(u,u)]/(1-u) = 2-\lambda_U$. □

Proposition 2 (Quantile Unit-Lomax Distortion). *Let $C_{Q_L}(u,v)$ be the Q_L-distorted copula defined in (10), where $\theta \geq 1$ and $\alpha \geq 1$. Then,*

(i) *$\kappa_{Q_L,L} = \kappa_L$ and $\lambda_{Q_L,L} = \lambda_L$.*

(ii) *$\kappa_{Q_L,U} = \kappa_U$ when $\alpha = 1$ and $\kappa_{Q_L,U} = 1$ when $\alpha > 1$. And $\lambda_{Q_L,U} = 2-(2-\lambda_U)^{1/\alpha}$.*

Proof. The tail orders are shown in (A10) and (A18) in Appendix B. By L'Hopital's rule,

$$\begin{aligned}\lambda_{Q_L,L} &= \lim_{u\to 0^+}\frac{\left(1-C(u,u)\right)^{-1/\alpha}-1}{\left(1-C(u,u)\right)^{-1/\alpha}-1+\theta}\Big/\frac{(1-u)^{-1/\alpha}-1}{(1-u)^{-1/\alpha}-1+\theta}\\ &= \lim_{u\to 0^+}\frac{[(1-u)^{-1/\alpha}-1+\theta]^2}{[\left(1-C(u,u)\right)^{-1/\alpha}-1+\theta]^2}\frac{\left(1-C(u,u)\right)^{-1/\alpha-1}C'(u,u)}{(1-u)^{-1/\alpha-1}}=\lambda_L\end{aligned}$$

since $\lim_{u\to 0^+} C(u,u)/u = \lim_{u\to 0^+} dC(u,u)/du = \lambda_L$. By (15) and L'Hopital's rule, the ltd coefficient is given by

$$\begin{aligned}\lambda_{Q_L,U} &= 2-\lim_{u\to 1-}\frac{\theta}{\left(1-C(u,u)\right)^{-1/\alpha}-1+\theta}\Big/\frac{\theta}{(1-u)^{-1/\alpha}-1+\theta}\\ &= 2-\lim_{u\to 1-}\left[\frac{1-C(u,u)}{1-u}\right]^{1/\alpha+1}\left[\frac{dC(u,u)}{du}\right]^{-1}=2-(2-\lambda_U)^{1/\alpha}\end{aligned}$$

since $\lim_{u\to 1-}[1-C(u,u)]/[1-u] = 2-\lambda_U$. □

Proposition 3 (Unit-Inverse Pareto Distortion). *Let $C_{G_P}(u,v)$ be the G_P-distorted copula defined in (11), where $0<\theta\leq 1$ and $\alpha \geq 1$. Then,*

(i) *$\kappa_{G_P,L} = \kappa_L$ and $\lambda_{G_P,L} = (\lambda_L)^a$.*

(ii) *$\kappa_{G_P,U} = \kappa_U$ and $\lambda_{G_P,U} = \lambda_U$.*

Proof. The tail orders are shown in (A11) and (A19) in Appendix B. For the ltd, with the help of L'Hopital's rule, we obtain that

$$\lambda_{G_P,L} = \lim_{u\to 0^+} \frac{\{C(u,u)/[1-C(u,u)+\theta C(u,u)]\}^\alpha}{\{u/[1-u+\theta u]\}^\alpha} = \lim_{u\to 0^+} \left(\frac{C(u,u)}{u}\right)^\alpha \left\{\frac{1-C(u,u)+\theta C(u,u)}{1-u+\theta u}\right\}^{-\alpha} = (\lambda_L)^\alpha.$$

The utd coefficient, by (15) and L'Hopital's rule, is given by

$$\lambda_{G_P,U} = 2 - \lim_{u\to 1-} \frac{1-\{C(u,u)/[1-C(u,u)+\theta C(u,u)]\}^\alpha}{1-\{u/[1-u+\theta u]\}^\alpha} = 2 - \lim_{u\to 1-} \left(\frac{C(u,u)}{u}\right)^{(\alpha-1)} \left\{\frac{1-C(u,u)+\theta C(u,u)}{1-u+\theta u}\right\}^{-\alpha-1} \frac{dC(u,u)}{du} = \lambda_U$$

since $\lim_{u\to 1-} C(u,u) = 1$ and $\lim_{u\to 1-} dC(u,u)/du = 2-\lambda_U$. □

Proposition 4 (Quantile Unit-Inverse Pareto Distortion). *Let $C_{Q_P}(u,v)$ be the Q_P-distorted copula defined in (13), where $\theta \geq 1$ and $0 < \alpha \leq 1$. Then,*

(i) *$\kappa_{Q_P,L} = \kappa_L$ and $\lambda_{Q_P,L} = (\lambda_L)^{1/\alpha}$.*

(ii) *$\kappa_{Q_P,U} = \kappa_U$. If $\kappa_U = 1$, then $\lambda_{Q_P,U} = \lambda_U$.*

Proof. The tail orders are derived in (A13) and (A20) in Appendix B. For the ltd coefficient,

$$\lambda_{Q_P,L} = \lim_{u\to 0^+} \frac{1/\left[1+\theta\left(C(u,u)^{-1/\alpha}-1\right)\right]}{1/[1+\theta(u^{-1/\alpha}-1)]} = \lim_{u\to 0^+} \frac{1+\theta(u^{-1/\alpha}-1)}{1+\theta\left(C(u,u)^{-1/\alpha}-1\right)} = \lim_{u\to 0^+} \left(\frac{C(u,u)}{u}\right)^{1+1/\alpha} \frac{du}{dC(u,u)} = (\lambda_L)^{1/\alpha}.$$

The utd coefficient, by (15) and L'Hopital's rule, is given by

$$\lambda_{Q_P,U} = 2 - \lim_{u\to 1-} \frac{\theta[C(u,u)^{-1/\alpha}-1]}{1+\theta[C(u,u)^{-1/\alpha}-1]} \Big/ \frac{\theta(u^{-1/\alpha}-1)}{1+\theta(u^{-1/\alpha}-1)} = 2 - \lim_{u\to 1-} \frac{1+\theta(u^{-1/\alpha}-1)}{1+\theta[C(u,u)^{-1/\alpha}-1]} \frac{C(u,u)^{-1/\alpha}-1}{(u^{-1/\alpha}-1)} = 2 - \lim_{u\to 1-} \frac{1+\theta(u^{-1/\alpha}-1)}{1+\theta[C(u,u)^{-1/\alpha}-1]} \frac{[C(u,u)]^{-1/\alpha-1}}{u^{-1/\alpha-1}} \frac{dC(u,u)}{du} = \lambda_U$$

since $\lim_{u\to 1-} C(u,u)/u = 1$ and $\lim_{u\to 1-} dC(u,u)/du = 2-\lambda_U$. □

The results of the propositions are summarized in Table 2. The utd coefficients of UL- and QUL-distorted copulas differ from those of base copulas when $\alpha \neq 1$, while the ltd coefficients remain unchanged. The UL and QUL distortions render new copulas with upper tail dependence regardless of whether the base copula has it or not. Conversely, the UIP and QUP distortions form copulas with different ltd coefficients when $\alpha \neq 1$ from the base copula, while the utd coefficients remain the same. The Clayton copula has zero upper tail dependence. However, based on Table 2, by applying UL and QUL distortions to a Clayton copula, we create a new family of copulas that are more flexible in the sense that they can accommodate upper tail dependence values ranging from 0 to 1. This same conclusion can be applied to the Frank copula.

Table 2. Tail dependence coefficients of the distorted copulas.

	UL ($0 < \alpha \leq 1$)	**QUL ($\alpha \geq 1$)**	**UIP ($\alpha \geq 1$)**	**QUP ($0 < \alpha \leq 1$)**
$\lambda_{T,L}$	λ_L	λ_L	$(\lambda_L)^{\alpha}$	$(\lambda_L)^{1/\alpha}$
$\lambda_{T,U}$	$2-(2-\lambda_U)^{\alpha}$	$2-(2-\lambda_U)^{1/\alpha}$	λ_U	λ_U

We next construct the density contour plots for UL- and UIP-distorted copulas in Figures 1 and 2. Contour plots and observations for QUL- and QUP-distorted copulas are displayed in Figures A1 and A2 in Appendix B.1, respectively. The joint pdf of a copula is $\partial C(u,v)/\partial u \partial v$. Note, if a contour plot is elongated along one direction, it signals a strong dependence in that direction. A circular contour shape indicates independence between variables. For example, the Clayton has lower tail dependence and no upper tail dependence, therefore, one expects to see a elongated or tightened contour line on the lower left-hand side. Let r be the parameter in the base copula. When $\theta = \alpha = 1$, both the UL and UIP distortions deliver us the base copulas shown in the first column of both figures.

While the Frank copula is tail-independent, the Clayton and Gumbel copulas are characterized by asymmetric tail dependence with only utd and ltd, respectively. As summarized in Table 2, UL distortions may yield new families of copulas with upper tail dependence. Copulas constructed using the UL distortion exhibit upper tail dependence and maintain the same patterns of lower tail behaviors as the base copula. Notably, UL-distorted Clayton copulas have both lower and upper tails.

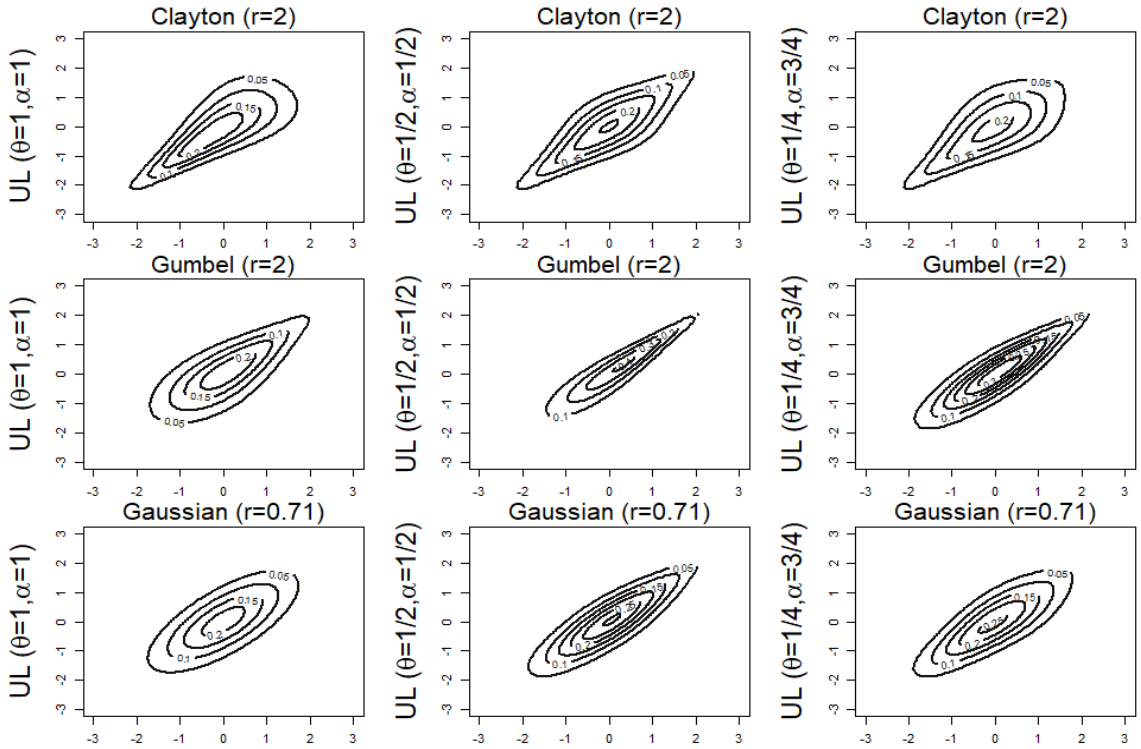


Figure 1. Density contour plots of the UL-Clayton, UL-Gumbel, and UL-Gaussian copulas with standard normal margins and parameters $(\theta, \alpha) = (1,1), (1/2, 1/2)$, and $(1/4, 3/4)$. For Clayton, Gumbel, and Gaussian, the parameter r is selected so that the value of Kendall's tau is $1/2$.

As revealed in Table 2 and Figure 2, UIP-distorted copulas have similar characteristics to the base copulas in the first column. If the base copula has lower tail dependence, the family introduced an additional parameter to the ltd coefficient but retains the upper tail dependence of the base copula.

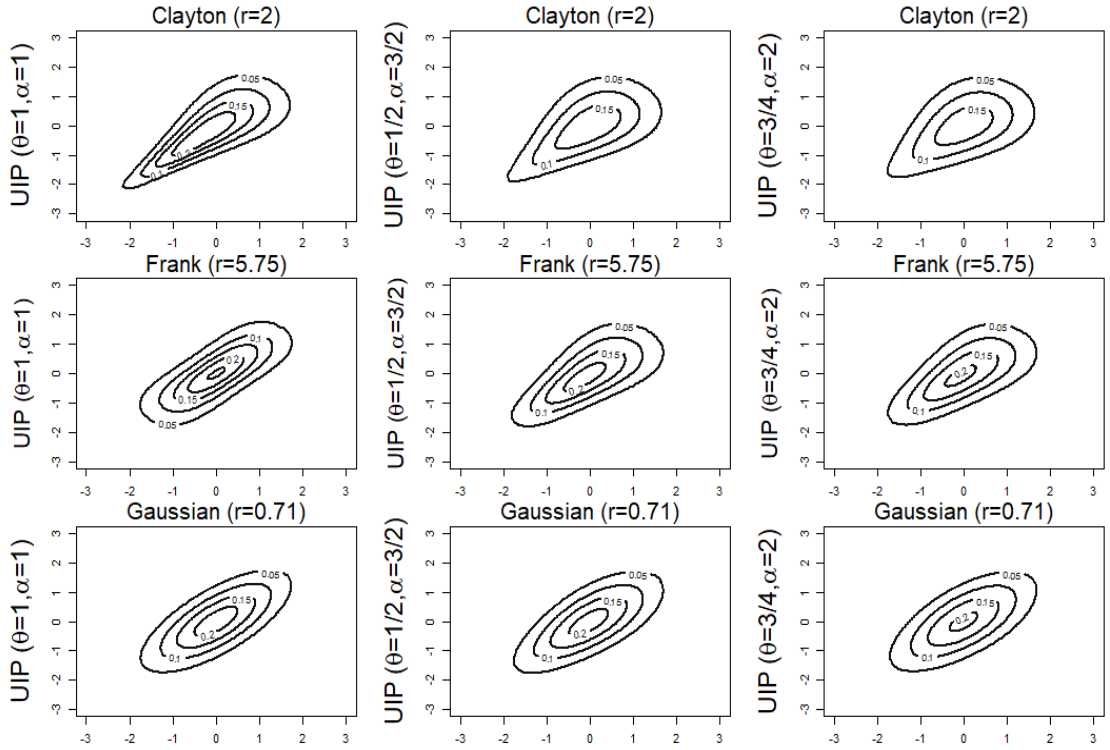


Figure 2. Density contour plots of the UIP-Clayton, UIP-Gumbel, and UIP-Gaussian copulas with standard normal margins and parameters $(\theta, \alpha) = (1,1), (1/2,1/2)$, and $(1/4,3/4)$. For Clayton, Frank, and Gaussian, the parameter r is selected so that the value of Kendall's tau is $1/2$.

5. Concordance Order and Kendall's Tau

In this section, we examine the concordance order and Kendall's tau. We provide formulas for Kendall's tau of the UL- and UIP-distorted copulas by using the Formula (16) and the Archimedean generators provided in Section 3. The formulas can be readily adapted to write programs to compute Kendall's tau values at various parameter values.

5.1. Concordance Ordering

A family of copula functions $C_r(u,v) = C(u,v;r)$ with parameter r is positively ordered, denoted by $C_{r_1} \prec C_{r_2}$, if $C_{r_1}(u,v) \leq C_{r_r}(u,v)$, and is negatively ordered, denoted by $C_{r_1} \succ C_{r_2}$, if $C_{r_1}(u,v) \geq C_{r_r}(u,v)$ for all $r_1 \leq r_2$ and $u,v \in I$. According to the definition of a concordance measure, if a family of copulas is ordered by a parameter, its Kendall's tau is either nonincreasing or nondecreasing in the parameter; see [11].

One can compute the first derivative with respect to a parameter to determine if copula C is positively or negatively ordered by the parameter, which can be a daunting task. If a copula is of Archimedean class, in addition to Theorem A1 in Appendix C, we can utilize the following corollary to examine the concordance orderings in the parameters; see [11] or [28].

Theorem 1 ([11]). *Let C_1 and C_2 be Archimedean copulas with generators ψ_1 and ψ_2, respectively. Then $C_1 \prec C_2$ holds if one of the following conditions is satisfied: (i) $\psi_1 \circ \psi_2^{-1}$ is concave; (ii) ψ_1/ψ_2 is nondecreasing on I; and (iii) ψ_1 and ψ_2 are continuously differentiable on I and ψ_1'/ψ_2' is nondecreasing on I.*

If the base copula C is positively ordered, then, for $r_1 \leq r_2$, $C(T^{-1}(u), T^{-1}(v); r_1) \leq C(T^{-1}(u), T^{-1}(v); r_2)$ for all $u,v \in I$. Since T is increasing, $T\big(C(T^{-1}(u), T^{-1}(v); r_1)\big) \leq T\big(C(T^{-1}(u), T^{-1}(v); r2)\big)$. Similar arguments can be applied to a negatively ordered base copula. That is, a family of distortion copulas built by admissible distortions preserves

the concordance order in the parameter of the base copula if the family of base copulas is ordered by the parameter; see also [24].

If the base copula is Archimedean with generator ψ, which is free of the parameters θ and α, the new family of T-distorted copulas can be written as

$$T\big(C(T^{-1}(u), T^{-1}(v))\big) = T \circ \psi^{-1}\big(\psi(T^{-1}(u)) + \psi(T^{-1}(v))\big),$$

with Archimedean with a generator given by $\psi(T^{-1}(u))$, a continuous, strictly decreasing, convex function such that $\psi(T^{-1}(1)) = 0$. Assume below the base copula is Archimedean with generator ψ. Let T_1 and T_2 be the T distortion evaluated at parameter values r_1 and r_2, where $r_1 \leq r_2$, respectively.

Corollary 1. *Let C_{T_1} and C_{T_2} be Archimedean copulas with generators $\psi \circ T_1^{-1}$ and $\psi \circ T_2^{-1}$, respectively. If one of the following conditions holds: (i) $\psi \circ T_1^{-1} \circ T_2 \circ \psi^{-1}$ is concave; (ii) $\psi \circ T_1^{-1}/\psi \circ T_2^{-1}$ is nondecreasing on $(0,1)$; or (iii) ψ is continuously differentiable on I and $(\psi^{'} \circ T_1^{-1}/\psi^{'} \circ T_2^{-1})(T_2^{'} \circ T_2^{-1}/T_1^{'} \circ T_1^{-1})$ is nondecreasing on I holds, then $C_{T_1} \prec C_{T_2}$.*

5.2. Kendall's Tau

For the T distortion-induced copulas in (1), by substituting $T^{-1}(u) = x$ and $T^{-1}(v) = y$ in the definition of Kendall's tau in (A5), then its Kendall's tau can be expressed as

$$\tau = 1 - 4\int_0^1 \int_0^1 [t(C(x,y))]^2 C_{1|2}(x|y) C_{2|1}(y|x) dxdy,$$

where $t(v) = dT(v)/dv$, $C_{2|1}(u,v) = \partial C(u,v)/\partial u$ and $C_{1|2}(u,v) = \partial C(u,v)/\partial v$. Numerical integration methods will be required to compute Kendall's tau. Define $\Psi(u) = \psi(T^{-1}(u))$. By (A6), Kendall's tau for a distorted copula is given by

$$\tau \quad = \quad 1 + 4\int_0^1 \frac{\Psi(u)}{\Psi'(u)} du = 1 + 4\int_0^1 \frac{\psi(v)}{\psi'(v)} t^2(v) dv. \tag{16}$$

We next present the explicit formulas of Kendall's tau for UL and UIP distortions when the base copulas are independence, Clayton, Gumbel, and Frank copulas. The Archimedean generators for the base copulas are reported in Table A1 in Appendix A.

- UL-independence copula: $1 - 4\int_0^1 \frac{\alpha[(1-u)^{-1/\alpha} - 1]\log\left(\theta[(1-u)^{-1/\alpha} - 1]^{-1} - 1\right)}{\theta[(1-u)^{-1/\alpha - 1}][\theta + (1-u)^{-1/\alpha} - 1]^{-1}} du.$
- UL-Clayton copula: $1 - 4\int_0^1 \frac{\alpha\theta u(1-u^r)}{r(1-u)^2[1+\theta(-1+(1-u)^{-1})]^{\alpha+1}} du.$
- UL-Gumbel copula: $1 - 4\int_0^1 \frac{\alpha\theta u(-\log u)}{r(1-u)^2[1+\theta(-1+(1-u)^{-1})]^{\alpha+1}} du.$
- UL-Frank copula: $1 - 4\int_0^1 \frac{\alpha\theta(1-e^{ru})}{r(1-u)^2[1+\theta(-1+(1-u)^{-1})]^{\alpha+1}} \log\left(\frac{e^{-ru}-1}{e^{-r}-1}\right) du.$
- UIP-independence copula: $1 - 4\int_0^1 \frac{\alpha[1+\theta(u^{-1/a}-1)]\log\left(1+\theta(u^{-1/a}-1)\right)}{\theta u^{-1/\alpha - 1}} du.$
- UIP-Clayton copula: $1 - 4\int_0^1 \frac{\alpha u(1-u^r)}{r\theta u^2[1+\theta^{-1}(u^{-1}-1)]^{\alpha+1}} du.$
- UIP-Gumbel copula: $1 - 4\int_0^1 \frac{\alpha u(-\log u)}{r\theta u^2[1+\theta^{-1}(u^{-1}-1)]^{\alpha+1}} du.$
- UIP-Frank copula: $1 - 4\int_0^1 \frac{\alpha(1-e^{ru})}{r\theta u^2[1+\theta^{-1}(u^{-1}-1)]^{\alpha+1}} \log\left(\frac{e^{-ru}-1}{e^{-r}-1}\right) du.$

Example 7. *We here illustrate only the ordering in the parameter θ for UL and UIP distortions of the independence copula in Appendix C due to the page limit. While not all mathematically shown, Kendall's tau surface plots in Figure 3 indicate that UL, QUL, UIP, and QUP distortions of the*

independence copula result in new families that are negatively ordered in the parameter θ and α. For the family of UIP-independence copulas in (12), the BB10 copula is negatively ordered by the parameter θ, see [12]. The family of QUP-independence copulas in (13) does not depend on the parameter α, and hence the flat lines along the α axis.

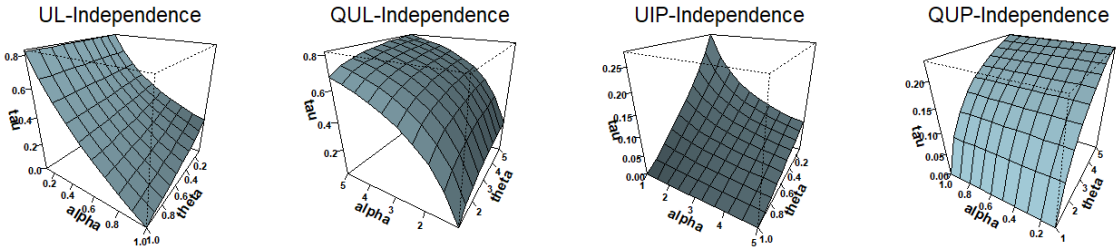


Figure 3. Surface plots of Kendall's tau for UL-, QL-, QIP-, and QUP-independence copulas.

Not all families of distortion-generated copulas are ordered by concordance. For example, the UL-distorted Clayton family. When a family of copulas is ordered by a parameter, we expect Kendall's tau values to increase or decrease with the parameter. Figure 4 exhibits Kendall's tau surface plots when distortions are applied to the Clayton copula with parameter r. These plots indicate that the distortion copulas, just like the base Clayton copula, are positively ordered by the parameter r. The UL-Clayton copula with $r = 15$ and the QUL-Clayton copula with $r = 20$ are not ordered by the parameter α. Regarding the UIP-Clayton copulas, the concordance order fails in the parameter α at $r = 25$ and at $\theta = 1/10$, as well as in the parameter θ when $\alpha = 25$. Furthermore, depending on the α or r value, the QUL-Clayton copula can be positively or negatively ordered by the parameter θ when $r = 2$ and $\alpha = 3/4$.

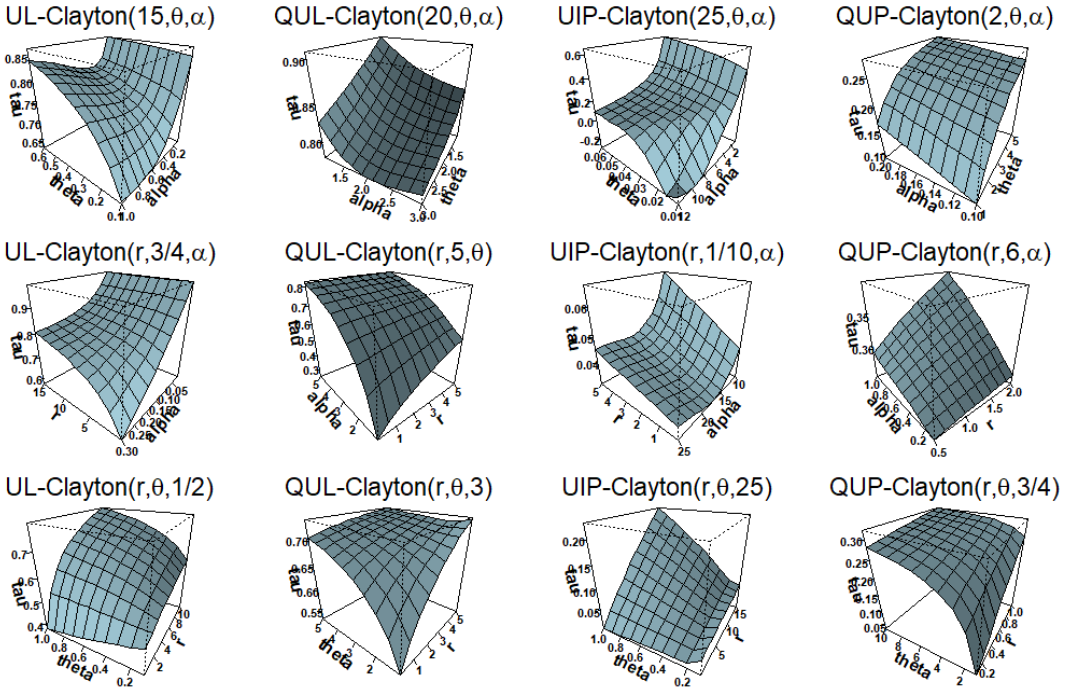


Figure 4. Surface plots of Kendall's tau constructed at various combinations of parameter values for the UL-Clayton, QL-Clayton, UIP-Clayton, and QUP-Clayton copula.

6. Numerical Results

In this section, we run a simulation study to inspect how the newly minted families of UL distortion copulas perform when fit to data generated from the beloved Clayton,

Gumbel, Gaussian, and Frank copulas, and vice versa. Additionally, the copula models are applied to a bivariate dataset consisting of the daily return rates of Amazon and Google stocks.

6.1. A Simulation Study

A general algorithm to generate draws from a bivariate copula C is the conditional distribution approach, as described by [19,24]. It consists of two steps: (i) generate two independent uniform random values (u_1, v) and (ii) solve $C(u_2|u_1) - v = 0$ for u_2, where $C(u|v) = \partial C(u,v)/\partial v$. The desired pair is (u_1, u_2). Using this algorithm, we generated 2000 bivariate pseudo-observations from the Clayton, Gumbel, Gaussian, and Frank copulas. We also simulated the same number of pseudo-observations from the families of UL-distorted copulas, where the four copulas served as base copulas. We do not present the figures for QIP, QUL, and QUP distortions, as the conclusions are similar to those from UL distortion.

The values of parameters in the base copulas are selected so they have Kendall's tau value of 0.5. The UL distortion has parameter values of $\alpha = 1/2$ and $\theta = 1/2$. We used the pseudo-likelihood estimation method [12] that maximizes (17) to fit the base copula and distorted copulas to the data. We then computed the empirical probabilities using the estimated copula models and constructed the probability–probability (PP) plots of the estimated probability distribution against the theoretical one. In the first row, the four base copulas are approximated by their UL-distorted counterparts, UL-Clayton, UL-Gumbel, UL-Gaussian, and UL-Frank copulas, and vice versa in the second row. The solid black line is the one that resulted from fitting the data to the copula model, from which observations were generated.

As another way of comparison, we also calculated the maximum distance in Table 3 between the theoretical and empirical probabilities at each data pair of (u_1, u_2). The univariate Kolmogorov–Smirnov (KS) test came to mind, and for a sample size of 2000, the 95% critical value of 0.03 is used as an ad hoc threshold.

Based on Figure 5 and Table 3, the Clayton copula, which is represented in red in the second row of Figure 5, shows greater deviations from the 45-degree line when fit to the data generated from UL-distorted copulas. According to Lemma 1, the UL-distorted copula has an upd coefficient of $2 - (2 - \lambda_U)^\alpha$, which may be attuned to zero. Furthermore, it has zero ltd when the base copula has zero ltd. The Clayton copula does not have upper tail dependence, so one would expect it to perform poorly when fit to the data generated from the UL-distorted family. In contrast, e.g., the UL-Gumbel appears to do well when fit to the data generated from the Frank and Gaussian copulas. In general, the performance of a copula depends on its tail dependence characteristics, and the results show that the UL-distorted copulas are more flexible as they have extra parameters.

Table 3. Maximum distance between the theoretical and empirical copula distributions.

Distorted Copula	Approximating			
	Clayton by	Gumbel by	Frank by	Gaussian by
UL-Clayton	0.014	0.021	0.035	0.026
UL-Gumbel	0.028	0.012	0.019	0.008
UL-Frank	0.021	0.012	0.011	0.024
UL-Gaussian	0.031	0.021	0.031	0.016
Base Copula	**Approximating**			
	UL-Clayton by	UL-Gumbel by	UL-Frank by	UL-Gaussian by
Clayton	0.035	0.047	0.042	0.042
Gumbel	0.050	0.013	0.034	0.033
Frank	0.046	0.024	0.030	0.032
Gaussian	0.021	0.019	0.017	0.022

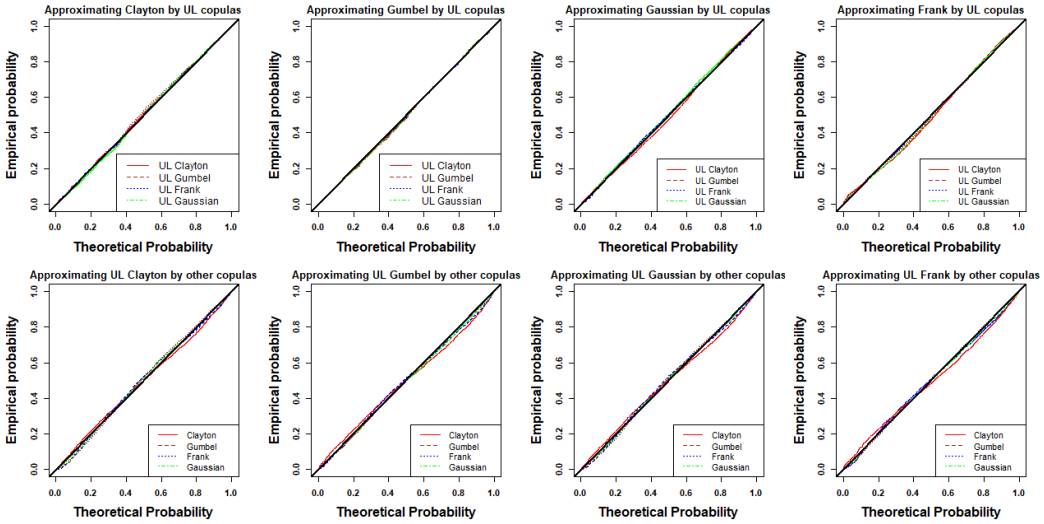


Figure 5. PP plots of UL and base copulas.

6.2. Empirical Application

We fit the proposed families of copula models to a bivariate dataset of daily return rates on Amazon and Google stocks. Historical data for the daily open, close, high, and low prices, and the adjusted closing price for stocks can be downloaded from Yahoo Finance. The adjusted closing price accounts for any splits and dividend distributions. We downloaded the data for Amazon and Google stocks for the period from January 2014 to December 2023, which amounts to a sample size of 2516 daily data points. The daily net return rates in percentages were then calculated based on the adjusted closing price. To calculate the return rate for today, the difference between today's price and yesterday's price is divided by yesterday's price.

Table 4 displays the summary statistics for both variables. The sample Pearson correlation and Kendall's tau are 0.71 (p-value < 0.001) and 0.49 (p-value < 0.001), respectively, both of which are significantly different from zero. Compared to Google, the center tendency measures for Amazon are smaller, but there is no significant difference in means. The Amazon daily rate return is significantly more variable based on the F test and is more skewed judging from the skewness measures and histograms in Figure 6.

Table 4. Descriptive statistics of the dataset.

	Mean	Sd	Min	1st Qu	Median	3rd Qu	Max	Skew
Amazon	−0.01	1.60	−8.56	−0.86	−0.01	0.91	8.24	−0.09
Google	0.04	1.35	−5.76	−0.67	0.05	0.84	6.65	−0.20

Let $\{x_i, y_i\}_{i=1}^n$ denote the bivariate observations. The pseudo-observations or scaled empirical distributions $\{u_i, v_i\}_{i=1}^n$ are defined to be $u_i = \sum_{j=1}^n J(x_j \le x_i)/(n+1)$ and $v_i = \sum_{j=1}^n J(y_j \le y_i)/(n+1)$, where $J(\cdot)$ is the indicator function. Figure 6 contains the scatter plots of y_i versus x_i and v_i versus u_i. Based on the histograms, the return rates for both stocks are concentrated around their center, which is also reflected in the resulting scatter plot. However, the pseudo-observations computed using a scaled empirical distribution are uniformly distributed. Therefore, one would expect a more evenly dispersed scatter plot.

The maximum pseudo-likelihood estimation introduced by [18] is used to estimate the parameters. It maximizes the pseudo-log-likelihood function, i.e., the log-likelihood with the copula functions evaluated at pseudo-observations, given by

$$L(\gamma) = \sum_{i=1}^{n} \log c_T(u_i, v_i; \gamma), \tag{17}$$

where c_T is the copula pdf in (A1) and γ is the parameter vector.

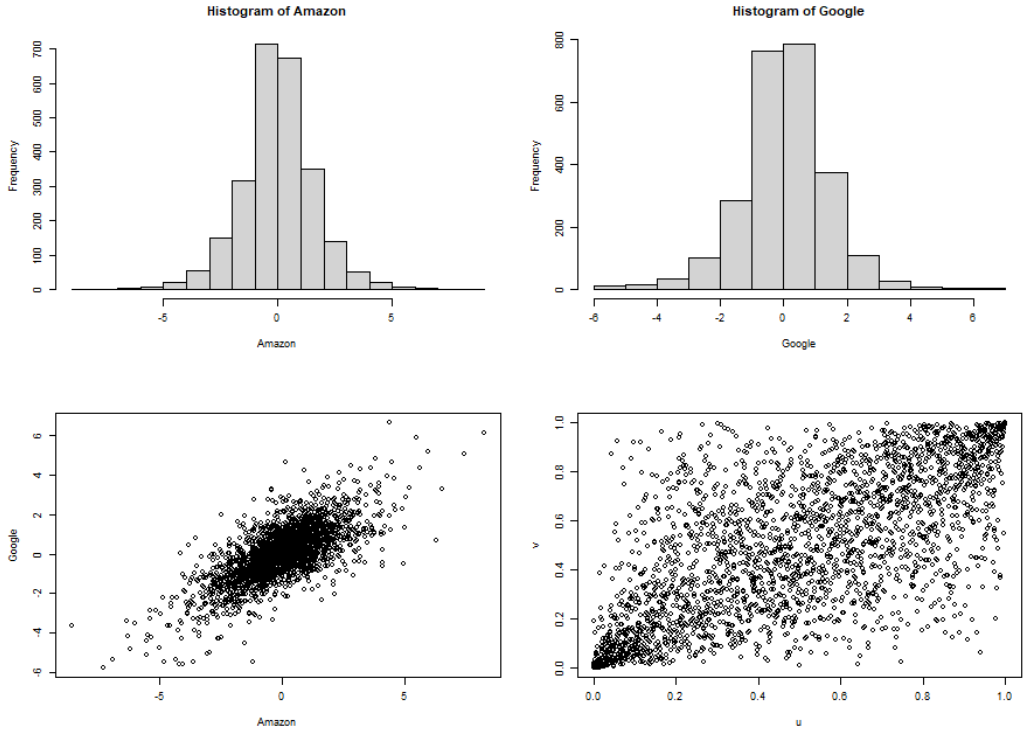


Figure 6. Histograms of daily return rates for Amazon and Google stocks, and scatter plots of Amazon against Google return rates and their corresponding pseudo-observations.

We did not fit a marginal distribution to each of the net returns. Our primary objective was to compare the new families of the distorted copula with the base copulas. Let r be the parameter in the base copula. Table 5 below reports the estimates with the estimated standard error in the parentheses, the maximum pseudo-likelihood (MPL), and the AIC values. All the parameter estimates fall within admissible spaces. Kendall's tau estimates $\widehat{\tau}$ were computed by plugging parameter estimates into either the theoretical Kendall's tau formula or (16).

Based on the scatter plots, it appears that there is a weak dependence in both the lower and upper tails between the daily net returns of the two stocks. Table 5 shows that among the base copulas considered, the Gaussian copula performs the best in terms of MPL and AIC, followed by Gumbel. The estimated Kendall's tau calculated from the Gaussian copula produces the closest match to the sample Kendall's tau between Google and Amazon.

The Frank copula, which is supposedly suitable for data with weak tail dependence, performs the worst. Note that the Gumbel copula with a parameter value of 1 represents the independent copula where Kendall's tau is equal to 0. It performs better than the Clayton copula, which suggests that there might be a stronger upd than ltd. Furthermore, a distortion copula that can accommodate a wider range of upper tail dependence, e.g., UL-distorted copulas, may do well in fitting this net return dataset.

Table 5 indicates that the UL-distorted copula model outperforms the corresponding base copula. The fitted UL-Clayton copula model has the largest AIC value, with the estimated lower and upper tail dependence coefficients of 0.44 and 0.50. It is less satisfactory than other UL-distorted copula models, probably due to weak lower tail dependence in the data. Both the UL-distorted Frank and Gaussian copulas have an estimated upper tail dependence coefficient of 0.39 and perform better than the Gaussian copula in terms of MPL and AIC.

The UIP-Clayton and UIP-Frank copulas do not exhibit tail dependence behaviors and their performance is worse than their base copulas. The UL-Gumbel and UIP-Gumbel copulas are the best performers, with UL-Gumbel being slightly better than UIP-Gumbel in terms of MPL and AIC. The UIP-Gumbel model produces an estimated Kendall's tau closer to the sample Kendall's tau. Both models have upper tail dependence, but not lower tail tail dependence.

According to the copula models employed in this application, there is a moderate linear correlation between the daily net returns of Amazon and Google stocks. Additionally, there appears asymmetrical in the extreme co-movements; that is, joint extremes are more likely for high daily net return values than for low daily net return values.

Table 5. MPL, AIC, $\widehat{\tau}$, parameter estimates and their standard deviation in parentheses $\widehat{r}$, $\widehat{\theta}$, $\widehat{\alpha}$ for the base, UL-distorted, and UIP-distorted copula models.

Family	MPL	$\widehat{\tau}$	AIC	$\widehat{r}$	$\widehat{\theta}$	$\widehat{\alpha}$
Clayton	756.0	0.488	−1510.1	1.907(0.055)	–	–
Gumbel	758.7	0.461	−1515.5	1.856(0.033)	–	–
Frank	754.3	0.491	−1506.7	5.582(0.165)	–	–
Gaussian	838.6	0.493	−1675.3	0.699(0.009)	–	–
UL-Clayton	877.3	0.450	−1748.6	4.855(0.192)	0.000(0.174)	0.562(0.029)
UL-Gumbel	906.4	0.461	−1806.9	1.256(0.043)	0.008(0.004)	0.873(0.056)
UL-Frank	900.0	0.476	−1794.0	4.223(0.319)	0.084(0.023)	0.640(0.019)
UL-Gaussian	901.9	0.484	−1797.9	0.527(0.034)	0.028(0.012)	0.686(0.026)
UIP-Clayton	748.3	0.353	−1490.6	1.413(0.097)	0.999(0.030)	1.296(0.108)
UIP-Gumbel	906.4	0.494	−1806.8	1.392(0.036)	0.010(0.006)	1.174(0.097)
UIP-Frank	748.5	0.481	−1494.9	9.281(0.443)	0.999(0.060)	3.484(0.043)
UIP-Gaussian	887.1	0.472	−1768.3	0.852(0.073)	0.022(0.062)	1.995(0.194)

7. Concluding Remarks

The framework advanced in the paper originates from the fact that a cumulative distribution function with unit interval support is a distortion function. It employs a transformation of a non-negative random variable into a variable with the support of the unit interval. The additional parameters in the distortion allow for more modeling flexibility. As demonstrated in Section 3.2, distortion of the independence copula creates a new family of copulas that includes the base copula and other existing copulas as its members and accommodates a wider range of tail dependence behaviors that the independence copula would never dream of having.

The tail behavior of a copula model is a crucial factor in determining whether it can adequately fit the data. The use of UL and QUL distortions can morph a family of base copulas without upper tail dependence into a new family of copulas with upper tail dependence. The upper tail dependence coefficient of the UL- and QUL-distorted copulas involves more parameters than the one of the base copula. The distortions can ultimately lead to a better accommodation of the upper tail dependence when compared to the base copula. The tail behaviors in the families of the UIP- and QUP-distorted copulas are similar to the ones in the base copula. However, they can accommodate better the lower tail dependence when compared to the base copulas.

We are not certain whether a more complicated generating cdf or distortion, e.g., one with more than two parameters, would result in a new family of copulas with both upper and lower tail dependence when applied to a base copula with no tail dependence behavior. The framework proposed in this article opens the door to a world of new distortions. Due to the page length limit, further exploration of the concordance ordering of the new family of distorted copulas will be pursued in more detail. The distortions of multivariate copulas of higher dimensions may also be of interest. Unlike the distortions of bivariate copulas, the distortions of multivariate copulas require more care and will be explored in the future.

Author Contributions: All authors contributed equally. All authors have read and agreed to the published version of the manuscript

Funding: This research received no external funding.

Institutional Review Board Statement: The study did not require ethical approval.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are included in the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Preliminaries

Let $F(x,y)$ be the joint cumulative probability distribution (cdf) of continuous random variables (X,Y), and $F_X(x)$ and $F_Y(y)$ be the respective marginal cdfs of X and Y. By Sklar's Theorem [10], there exists a unique copula C, satisfying the boundary and monotonicity conditions, such that

$$F(x,y) = P(X \le x, Y \le y) = C\big(F_X(x), F_Y(y)\big), \quad x,y \in (-\infty,\infty). \tag{A1}$$

The joint probability density function (pdf), denoted by $f(x,y)$, of (X,Y) is

$$f(x,y) = c\big(F_X(x), F_Y(y)\big) f_X(x) f_Y(y), \tag{A2}$$

where f_X and f_Y are the respective pdfs of X, and Y and $c(u,v)$ is the copula pdf such that $c(u,v) = \partial^2 C(u,v)/\partial u \partial v$. The Equation (A2) implies that a joint bivariate probability distribution can be separated into univariate marginals and a dependence structure, where the dependence structure is represented by a copula.

A copula C by definition has the following properties: (i) $C(u,0) = C(0,v) = 0$, $(u,v) \in I^2$ where $I = [0,1]$; (ii) $C(u,1) = u$ and $C(1,v) = v$, $(u,v) \in I^2$; and (iii) $C(u_2,v_2) - C(u_2,v_1) - C(u_1,v_2) + C(u_1,v_1) \ge 0$, for $u_1 \le u_2$, $v_1 \le v_2$, and (u_1,u_2), (v_1,v_2) in I^2.

If C is an Archimedean copula with a strict generator $\psi(\cdot)$ such that $\psi(0) = \infty$, it can be expressed as $C(u,v) = \psi^{-1}(\psi(u) + \psi(v))$, where ψ^{-1} is the inverse of ψ. The function $\psi : [0,1] \to [1,\infty)$ is convex, continuous, and strictly decreasing with $\psi(1) = 0$ and $\psi(0) = \infty$. Archimedean copulas are popular because they admit explicit formulas and can accommodate higher dimensions with only one parameter. Table A1 highlights some prominent bivariate Archimedean copulas, their generators, and their tail dependence behaviors. Let Π be the independence copula.

Table A1. Important Archimedean copulas and their generators.

Name	Copula Function	Generator $\psi(t)$	κ_L or λ_L	κ_U or λ_U
Π	uv	$-\log(t)$	$\kappa_L = 2$	$\kappa_U = 2$
Clayton	$(u^{-r} + v^{-r} - 1)^{-1/r}, r > 0$	$(t^{-r} - 1)/r$	$\lambda_L = 2^{-1/r}$	$\kappa_U = 2$
Frank	$\frac{-1}{r}\log\left[1 + \frac{(e^{-ru}-1)(e^{-rv}-1)}{e^{-r}-1}\right]$, $r \ne 0$	$-\log\left(\frac{e^{-rt}-1}{e^{-r}-1}\right)$	$\kappa_L = 2$	$\kappa_U = 2$
Gumbel	$e^{\left[-((-\log(u))^r + (-\log(v))^r)^{1/r}\right]}, r > 1$	$\big[-\log(t)\big]^r$	$\kappa_L = 2^{1/r}$	$\lambda_U = 2 - 2^{1/r}$

The lower and upper tail dependence parameters, λ_L and λ_U, are given by

$$\begin{aligned} \lambda_L &= \lim_{u\to 0^+} P(F_X(X) < u \mid F_Y(Y) < u) = \lim_{u\to 0^+} \frac{C(u,u)}{u}, \\ \lambda_U &= \lim_{u\to 0^+} P(F_X(X) > u \mid F_Y(Y) > u) = \lim_{u\to 1^-} \frac{\overline{C}(u,u)}{(1-u)} = 2 - \lim_{u\to 1^-} \frac{1 - C(u,u)}{1-u}, \end{aligned} \tag{A3}$$

where $\overline{C}(u,v) = P(U > u, V > v) = 1 - u - v + C(u,v)$. They measure the probability that a random variable reaches extreme values given another variable attains extremes.

A non-negative function f is said to be slowly varying if for $s > 0$, $\lim_{x\to 0} f(sx)/f(x) = 1$. Let f_1 and f_2 be two functions. If $\lim_{u\to u_0} f_1(u)/f_2(u) = 1$, we denote it by $f_1(u) \sim f_2(u)$ as $u \to u_0$. For a bivariate copula if

$$C(u,u) \sim u^{\kappa_L}\ell(u), u \to 0^+,$$

where $\ell(u)$ is slowly varying at 0^+, then κ_L is referred to as the lower tail order of C. The survival copula is given by

$$\widehat{C}(u,v) = u + v - 1 + C(1-u, 1-v). \tag{A4}$$

The upper tail order κ_U meets the condition that

$$\widehat{C}(u,u) \sim u^{\kappa_U}\ell_*(u), u \to 0^+$$

for some slowly varying function $\ell_*(u)$. Note that $1 \le \kappa_L, \kappa_U \le 2$. If $\kappa_L > 1$ and $\kappa_U > 1$, then $\lambda_L = \lambda_U = 0$. If $\kappa_L = 1, \lambda_L = \lim_{u\to 0^+} \ell(u)$; and if $\kappa_U = 1, \lambda_U = \lim_{u\to 0^+} \ell^*(u)$. When $\kappa_L = 2$ and $\ell(u) \to q$ as $u \to 0^+$, for some positive q, the variables are near independent in the lower tail. If $1 < \kappa_L < 2$, the variables are positively associated and have intermediate tail dependence. Similar conclusions can be made for the upper tail dependence; see [12,29] for more details. Kendall's tau value of a bivariate copula can be expressed as

$$\tau = 1 - 4\int_0^1 \int_0^1 \frac{\partial C}{\partial u}(u,v)\frac{\partial C}{\partial v}(u,v)\, dudv. \tag{A5}$$

For an Archimedean copula Kendall's tau can also be calculated by

$$\tau = 1 + 4\int_0^1 \frac{\psi(u)}{\psi'(u)}du. \tag{A6}$$

Appendix B. Contour Plots and Derivations of Tail Orders

Appendix B.1. Contour Plots

Figure A1 shows that the upper tail dependence tails of QUL-Clayton, QUL-Gumbel, and QUL-Gaussian appear more pronounced than the corresponding base copula for the selected α values, similar to those of UL-Clayton, UL-Gumbel, and UL-Gaussian, respectively. The UL- and QUL-distorted copulas have no lower tail dependence when the base copulas have no lower tail dependence. While the parameters are selected so that $\tau = 0.5$, the contour plots in Figures 1 and A1 show various shapes and sparsities.

Table 2 indicates that the upper tail dependence coefficient of the UIP- and QUP-distorted copulas is the same as that of the base copula. However, the lower tail dependence coefficients are $(\lambda_L)^\alpha$ where $\alpha \ge 1$ and $(\lambda_L)^\alpha$ where $0 < \alpha \le 1$, respectively. That is, the UIP distortion with $\alpha = 3/2$ in Figure 2 and the QUP distortion with $\alpha = 1/2$ in Figure A2 have similar if not the same shapes on the upper right side. Asymmetry is present in QUP-Frank and UIP-Frank when θ and α do not equal 1. The contour plots for UIP-Gaussian and QUP-Gaussian show little change.

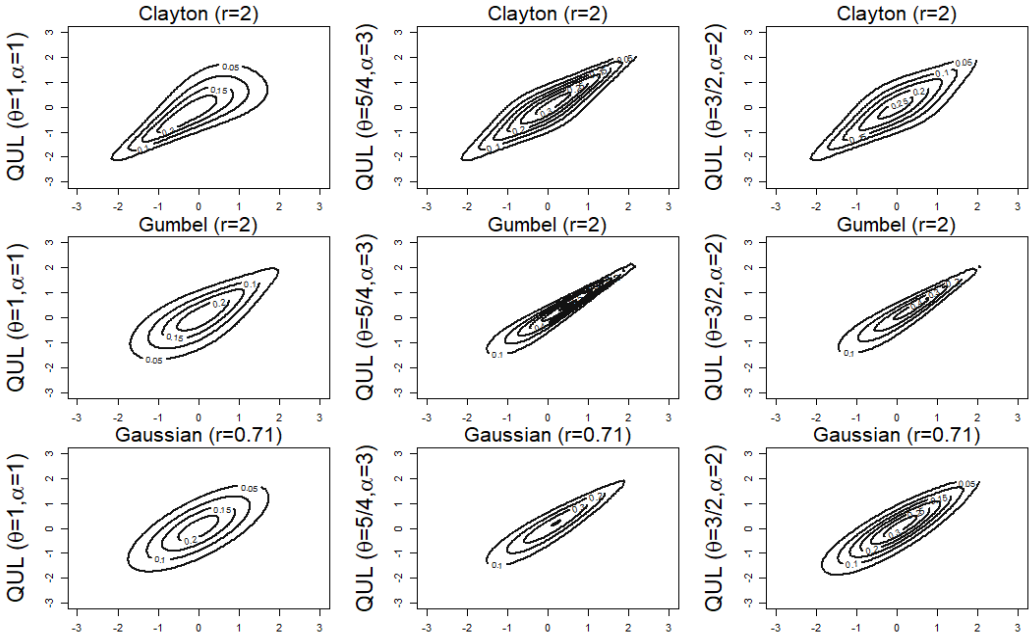


Figure A1. Density contour plots of the QUL-Clayton, QUL-Gumbel, and QUL-Gaussian copulas with standard normal margins and parameters $(\theta, \alpha) = (1,1), (3/2, 1/2)$, and $(2, 3/4)$. For Clayton, Gumbel, and Gaussian, the parameter r is selected so that the value of Kendall's tau is $1/2$.

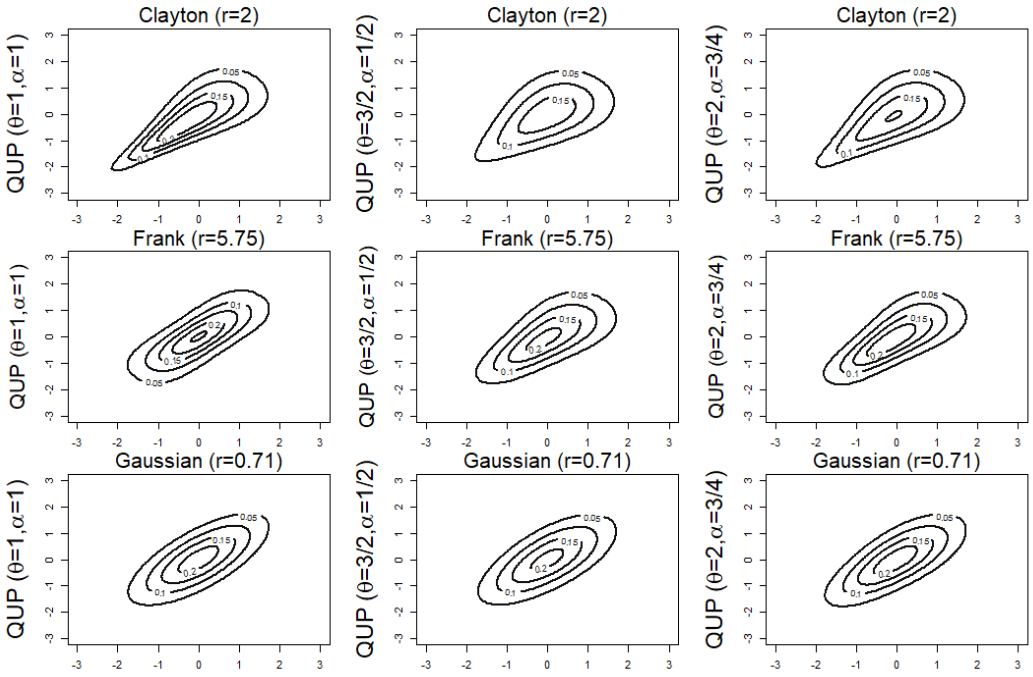


Figure A2. Density contour plots of the QUP-Clayton, QUP-Frank, and QUP-Gaussian copulas with standard normal margins and parameters $(\theta, \alpha) = (1,1), (3/2, 1/2)$, and $(2, 3/4)$. For Clayton, Frank, and Gaussian, the parameter r is selected so that the value of Kendall's tau is $1/2$.

Appendix B.2. Derivations of Tail Orders

The following Talyor's series approximations are key workhorses for calculating the tail orders of proposed distortions. For two constants a and b, we have that

$$(1+u)^a \sim 1+au, \ \frac{1}{1+au} \sim 1-au, \ \frac{u}{1+au} \sim u, \ \frac{u}{b+au} \sim \frac{u}{b} \text{ as } u \to 0. \tag{A7}$$

Assume $C(u,u) \sim u^{\kappa_L}\ell(u)$ as $u \to 0^+$ and $\overline{C}(1-u,1-u) \sim u^{\kappa_U}\ell_*(u)$ as $u \to 0^+$ for some slowly varying functions ℓ and ℓ_* at 0^+.

Appendix B.3. Lower Tail Orders

Note that $G_L^{-1} = Q_L$ and $G_P^{-1} = Q_P$. From (A7) and Table 1, we obtain that, as $u \to 0^+$,

$$\begin{aligned} G_L(u) &= 1 - \left(1 - \frac{\theta u}{1-(1-\theta)u}\right)^{\alpha} \sim 1 - (1-\theta u)^{\alpha} \sim \alpha\theta u; \\ Q_L(u) &= \frac{(1-u)^{-1/\alpha} - 1}{\theta + (1-u)^{-1/\alpha} - 1} \sim \frac{u/\alpha}{\theta + (1-u)^{-1/\alpha} - 1} \sim u/(\alpha\theta); \\ G_P(u) &\sim \theta^{\alpha}u^{\alpha}; \quad Q_P(u) \sim u^{1/\alpha}/\theta. \end{aligned} \tag{A8}$$

For UL-distorted copulas, since $C\big(G_L^{-1}(u), G_L^{-1}(u)\big) \to 0$ as $u \to 0^+$, applying (A8) yields

$$\begin{aligned} G_L\Big(C\big(G_L^{-1}(u), G_L^{-1}(u)\big)\Big) &= 1 - \left[\frac{1 - C\big(G_L^{-1}(u), G_L^{-1}(u)\big)}{1 - C\big(G_L^{-1}(u), G_L^{-1}(u)\big) + \theta C\big(G_L^{-1}(u), G_L^{-1}(u)\big)}\right]^{\alpha} \\ &\sim \alpha\theta C\big(G_L^{-1}(u), G_L^{-1}(u)\big) \sim \alpha\theta[Q_L(u)]^{\kappa_L}\ell\big(Q_L(u)\big) \\ &\sim u^{\kappa_L}(\alpha\theta)^{1-\kappa_L}\ell\big(Q_L(u)\big) \text{ as } u \to 0^+. \end{aligned} \tag{A9}$$

By (A8), $\ell\big(Q_L(u)\big) \sim \ell\big(u/(\alpha\theta)\big)$ and by definition, $(\alpha\theta)^{1-\kappa_L}\ell\big(Q_L(u)\big)$ is slowly varying. For the QUL-distorted copulas, Q_L, by Table 1, (A7) and (A8), we have

$$\begin{aligned} Q_L\Big(C\big(Q_L^{-1}(u), Q_L^{-1}(u)\big)\Big) &= \frac{\big[1 - C\big(Q_L^{-1}(u), Q_L^{-1}(u)\big)\big]^{-1/\alpha} - 1}{\theta + \big[1 - C\big(Q_L^{-1}(u), Q_L^{-1}(u)\big)\big]^{-1/\alpha} - 1} \\ &\sim (\alpha\theta)^{-1}C\big(Q_L^{-1}(u), Q_L^{-1}(u)\big) \sim (\alpha\theta)^{-1}[G_L(u)]^{\kappa_L}\ell\big(Q_L^{-1}(u)\big) \\ &\sim u^{\kappa_L}(\alpha\theta)^{\kappa_L - 1}\ell\big(G_L(u)\big) \text{ as } u \to 0^+. \end{aligned} \tag{A10}$$

By (A8), $\ell\big(G_L(u)\big) \sim \ell(\alpha\theta u)$ and by definition, $(\alpha\theta)^{\kappa_L - 1}\ell\big(G_L(u)\big)$ is slowly varying. For the UIP-distorted copulas, since $C\big(G_P^{-1}(u), G_P^{-1}(u)\big) \to 0$ as $u \to 0^+$ and by (A8),

$$\begin{aligned} G_P\Big(C\big(G_P^{-1}(u), G_P^{-1}(u)\big)\Big) &= \left[\frac{\theta C\big(G_P^{-1}(u), G_P^{-1}(u)\big)}{1 - C\big(G_P^{-1}(u), G_P^{-1}(u)\big) + \theta C\big(G_P^{-1}(u), G_P^{-1}(u)\big)}\right]^{\alpha} \\ &\sim \theta^{\alpha}\big[C\big(G_P^{-1}(u), G_P^{-1}(u)\big)\big]^{\alpha} \sim \theta^{\alpha}\big\{[Q_P(u)]^{\kappa_L}\ell\big(Q_P(u)\big)\big\}^{\alpha} \\ &\sim u^{\kappa_L}\theta^{\alpha(1-\kappa_L)}\big[\ell\big(Q_P(u)\big)\big]^{\alpha} \text{ as } u \to 0^+. \end{aligned} \tag{A11}$$

Since $G_P(u) \sim \theta^{\alpha}u^{\alpha}$ and ℓ is slowly varying at 0^+, for $s, \alpha > 0$,

$$\begin{aligned} &\lim_{u\to 0+} \frac{\ell(su)}{\ell(u)} = \lim_{u\to 0+}\left[\frac{\ell(su)}{\ell(u)}\right]^{\alpha} = 1; \\ &\lim_{u\to 0+} \frac{\ell(G_P(su))}{\ell(G_P(u))} = \lim_{u\to 0+} \frac{\ell(\theta^a s^{\alpha}u^{\alpha})}{\ell(\theta^a u^{\alpha})} = \lim_{v\to 0+} \frac{\ell(s^{\alpha}v)}{\ell(v)} = 1. \end{aligned} \tag{A12}$$

Therefore, $\big[\ell\big(Q_P(u)\big)\big]^{\alpha}$ is slowly varying.
For the QUP-distorted copulas, $C\big(Q_P^{-1}(u), Q_P^{-1}(u)\big) \to 0$ and by (A8), we obtain

$$\begin{aligned} Q_P\Big(C\big(Q_P^{-1}(u),Q_P^{-1}(u)\big)\Big) &= \frac{1}{1+\theta\big\{\big[C\big(Q_P^{-1}(u),Q_P^{-1}(u)\big)\big]^{-1/\alpha}-1\big\}} \\ &\sim \theta^{-1}\big[C\big(Q_P^{-1}(u),Q_P^{-1}(u)\big)\big]^{1/\alpha} \sim \theta^{-1}\big\{[G_P(u)]^{\kappa_L}\ell\big(G_P(u)\big)\big\}^{1/\alpha} \\ &\sim u^{\kappa_L}\theta^{\kappa_L-1}\big[\ell\big(G_P(u)\big)\big]^{1/\alpha} \text{ as } u\to 0^+. \end{aligned} \tag{A13}$$

By similar arguments in (A12), $\big[\ell\big(G_P(u)\big)\big]^{1/\alpha}$ can be shown to be slowly varying.

Appendix B.4. Upper Tail Orders

By the approximations in (A7), as $u\to 0^+$,

$$\begin{aligned} G_L(1-u) &= 1-\left[\frac{u}{\theta+(1-\theta)u}\right]^{\alpha} \sim 1-\big(u/\theta\big)^{\alpha}, \\ Q_L(1-u) &= \frac{u^{-1/\alpha}-1}{\theta+u^{-1/\alpha}-1} = 1-\frac{\theta u^{1/\alpha}}{1+(\theta-1)u^{1/\alpha}} \sim 1-\theta u^{1/\alpha}, \\ G_P(1-u) &= \left[\frac{\theta(1-u)}{u+\theta(1-u)}\right]^{\alpha} = \left[1-\frac{u}{\theta+(1-\theta)u}\right]^{\alpha} \sim 1-\alpha u/\theta, \\ Q_P(1-u) &= \frac{(1-u)^{1/\alpha}}{\theta+(1-\theta)(1-u)^{1/\alpha}} \sim \frac{1-u/\alpha}{\theta+(1-\theta)(1-u/\alpha)} \\ &\sim 1-\frac{\theta u/\alpha}{1-u/\alpha+\theta u/a} \sim 1-\theta u/\alpha. \end{aligned} \tag{A14}$$

By (A4), if $\widehat{C}(u,u)\sim u^{\kappa_U}\ell_*(u)$ as $u\to 0^+$, $C(1-u,1-u)\sim 1-2u+u^{\kappa_U}\ell_*(u)$, as $u\to 0^+$. Therefore, by (A14),

$$\begin{aligned} C\big(G_L^{-1}(1-u),G_L^{-1}(1-u)\big) &\sim C\big(1-\theta u^{1/\alpha},1-\theta u^{1/\alpha}\big) \\ &\sim 1-2\theta u^{1/\alpha}+(\theta u^{1/\alpha})^{\kappa_U}\ell_*(\theta u^{1/\alpha}) = 1-k_{G_L}(u), \\ C\big(Q_L^{-1}(1-u),Q_L^{-1}(1-u)\big) &\sim C\big(1-(u/\theta)^{\alpha},1-(u/\theta)^{\alpha}\big) \\ &\sim 1-2(u/\theta)^{\alpha}+(u/\theta)^{\alpha})^{\kappa_U}\ell_*((u/\theta)^{\alpha}) = 1-k_{Q_L}(u), \\ C\big(G_P^{-1}(1-u),G_P^{-1}(1-u)\big) &\sim C\big(1-\theta u/\alpha,1-\theta u/\alpha\big) \\ &\sim 1-2\theta u/\alpha+(\theta u/\alpha)^{\kappa_U}\ell_*(\theta u/\alpha) = 1-k_{G_P}(u), \\ C\big(Q_P^{-1}(1-u),Q_P^{-1}(1-u)\big) &\sim C\big(1-\alpha u/\theta,1-\alpha u/\theta\big) \\ &\sim 1-2\alpha u/\theta+(\alpha u/\theta)^{\kappa_U}\ell_*(\alpha u/\theta) = 1-k_{Q_P}(u). \end{aligned} \tag{A15}$$

Note that the remainder terms $k_{G_L}(u), k_{Q_L}(u), k_{G_P}(u)$, and $k_{Q_P}(u)$ go to 0 as $u\to 0^+$.

For UL-distorted copulas, by (A14) and then (A7), $G_L\big(C\big(G_L^{-1}(1-u),G_L^{-1}(1-u)\big)\big)$ is given by

$$\begin{aligned} &1-\left[\frac{1-C\big(G_L^{-1}(1-u),G_L^{-1}(1-u)\big)}{1-C\big(G_L^{-1}(1-u),G_L^{-1}(1-u)\big)+\theta C\big(G_L^{-1}(1-u),G_L^{-1}(1-u)\big)}\right]^{\alpha} \\ &\sim 1-\left[\frac{k_{G_L}(u)}{\theta+(1-\theta)k_{G_L}(u)}\right]^{\alpha} \sim 1-\big[k_{G_L}(u)/\theta\big]^{\alpha}. \end{aligned} \tag{A16}$$

Therefore, by (A4), for $0<\alpha\le 1$, and $\kappa_U\ge 1$, as $u\to 0^+$,

$$\hat{G}_L\Big(C\big(G_L^{-1}(1-u),G_L^{-1}(1-u)\big)\Big)=2u-1+G_L\Big(C\big(G_L^{-1}(1-u),G_L^{-1}(1-u)\big)\Big)$$
$$\sim 2u-\big[k_{G_L}(u)/\theta\big]^{\alpha}=2u-[2u^{1/\alpha}-\theta^{\kappa_U-1}u^{\kappa_U/\alpha}\ell_*(\theta u^{1/\alpha})]^{\alpha}$$
$$\sim u\big\{2-2^{\alpha}[1-2^{-1}\theta^{\kappa_U-1}u^{(\kappa_U-1)/\alpha}\ell_*(\theta u^{1/\alpha})]^{\alpha}\big\}$$
$$\sim u\big\{2-2^{\alpha}[1-\alpha 2^{-1}\theta^{\kappa_U-1}u^{(\kappa_U-1)/\alpha}\ell_*(\theta u^{1/\alpha})]\big\}$$
$$\sim \theta^{\kappa_U-1}u^{\kappa_U}\ell_*(\theta u) \text{ if } \alpha=1. \tag{A17}$$

For QUL-distorted copulas, by (A15), $Q_L\big(C\big(Q_L^{-1}(1-u),Q_L^{-1}(1-u)\big)\big)$ is given by

$$\frac{\big[1-C\big(Q_L^{-1}(1-u),Q_L^{-1}(1-u)\big)\big]^{-1/\alpha}-1}{\theta+\big[1-C\big(Q_L^{-1}(1-u),Q_L^{-1}(1-u)\big)\big]^{-1/\alpha}-1}=\frac{\big[k_{Q_L}(u)\big]^{-1/\alpha}-1}{\theta+\big[k_{Q_L}(u)\big]^{-1/\alpha}-1}$$
$$=1-\frac{\theta\big[k_{Q_L}(u)\big]^{1/\alpha}}{\theta\big[k_{Q_L}(u)\big]^{1/\alpha}+1-\big[k_{Q_L}(u)\big]^{1/\alpha}}\sim 1-\theta\big[k_{Q_L}(u)\big]^{1/\alpha}.$$

Therefore, for $\alpha\geq 1$, and $\kappa_U\geq 1$, as $u\to 0^+$,

$$\hat{Q}_L\big(C\big(G_L^{-1}(1-u),G_L^{-1}(1-u)\big)\big)=2u-1+Q_L\big(C\big(Q_L^{-1}(1-u),Q_L^{-1}(1-u)\big)\big)$$
$$\sim 2u-\theta\big[k_{Q_L}(u)\big]^{1/\alpha}=2u-\theta\big[2(u/\theta)^{\alpha}-(u/\theta)^{\alpha})^{\kappa_U}\ell_*((u/\theta)^{\alpha})\big]^{1/\alpha}$$
$$=2u-2^{1/\alpha}u\big[1-2^{-1}(u/\theta)^{\alpha(\kappa_U-1)}\ell_*((u/\theta)^{\alpha})\big]^{1/\alpha}$$
$$\sim u\Big\{2-2^{1/\alpha}\big[1-2^{-1}\alpha^{-1}(u/\theta)^{\alpha(\kappa_U-1)}\ell_*((u/\theta)^{\alpha})\big]\Big\}$$
$$\sim (u/\theta)^{\kappa_U}\ell_*(u/\theta) \text{ if } \alpha=1. \tag{A18}$$

For UIP-distorted copulas, by (A7) and (A15), $G_P\big(C\big(G_P^{-1}(1-u),G_P^{-1}(1-u)\big)\big)$ is given by

$$\left[\frac{\theta C\big(G_P^{-1}(1-u),G_P^{-1}(1-u)\big)}{1-C\big(G_P^{-1}(1-u),G_P^{-1}(1-u)\big)+\theta C\big(G_P^{-1}(1-u),G_P^{-1}(1-u)\big)}\right]^{\alpha}$$
$$=\left[\frac{\theta[1-k_{G_P}(u)]}{k_{G_P}(u)]+\theta[1-k_{G_P}(u)]}\right]^{\alpha}=\left[1-\frac{k_{G_P}(u)}{k_{G_P}(u)]+\theta[1-k_{G_P}(u)]}\right]^{\alpha}$$
$$\sim \big[1-k_{G_P}(u)/\theta\big]^{\alpha}\sim 1-\alpha\theta^{-1}k_{G_P}(u).$$

Therefore, for $\alpha\geq 1$, and $\kappa_U\geq 1$, as $u\to 0^+$,

$$\hat{G}_P\big(C\big(G_P^{-1}(1-u),G_P^{-1}(1-u)\big)\big)=2u-1+G_P\big(C\big(G_P^{-1}(1-u),G_P^{-1}(1-u)\big)\big)$$
$$\sim 2u-\alpha\theta^{-1}k_{G_P}(u)=2u-\alpha\theta^{-1}[2\theta u/\alpha-(\theta u/\alpha)^{\kappa_U}\ell_*(\theta u/\alpha)]$$
$$\sim \alpha\theta^{-1}(\theta u/\alpha)^{\kappa_U}\ell_*(\theta u/\alpha). \tag{A19}$$

For QIP-distorted copulas, by (A7) and (A15), $Q_P\big(C\big(Q_P^{-1}(1-u),Q_P^{-1}(1-u)\big)\big)$ is given by

$$\frac{[C\big(Q_P^{-1}(1-u),Q_P^{-1}(1-u)\big)]^{1/\alpha}}{[C\big(Q_P^{-1}(1-u),Q_P^{-1}(1-u)\big)]^{1/\alpha}+\theta\big\{1-[C\big(Q_P^{-1}(1-u),Q_P^{-1}(1-u)\big)]^{1/\alpha}\big\}}$$
$$=\frac{[1-k_{Q_P}(u)]^{1/\alpha}}{[1-k_{Q_P}(u)]^{1/\alpha}+\theta\big\{1-[1-k_{Q_P}(u)]^{1/\alpha}\big\}}$$
$$=\frac{1-k_{Q_P}(u)/\alpha}{1-k_{Q_P}(u)/\alpha+\theta k_{Q_P}(u)/\alpha}=\frac{\alpha-k_{Q_P}(u)}{\alpha-k_{Q_P}(u)+\theta k_{Q_P}(u)}$$
$$=1-\frac{\theta k_{Q_P}(u)}{\alpha-k_{Q_P}(u)+\theta k_{Q_P}(u)}\sim 1-\theta k_{Q_P}(u)/\alpha.$$

Therefore,

$$\hat{Q}_P\Big(C\big(Q_P^{-1}(1-u),Q_P^{-1}(1-u)\big)\Big) = 2u-1+Q_P\Big(C\big(Q_P^{-1}(1-u),Q_P^{-1}(1-u)\big)\Big)$$
$$\sim 2u-\theta k_{Q_P}(u)/\alpha = 2u-\theta\alpha^{-1}[2\alpha u/\theta-(\alpha u/\theta)^{\kappa_U}\ell_*(\alpha u/\theta)]$$
$$\sim u^{\kappa_U}(\alpha/)^{\kappa_U-1}\ell_*(\alpha u/\theta). \tag{A20}$$

By applying similar arguments as in (A12), we can show that ℓ_* in (A17)–(A20) are slowly varying.

Appendix C. Concordance Ordering

A function f is subadditive if $f(x+y)\le f(x)+f(y)$ for all x and y in its domain. A function f is superadditive if $f(x+y)\ge f(x)+f(y)$ for all x and y in its domain.

Theorem A1 ([11]). *let C_1 and C_2 be two Archimedean copulas generated by ψ_1 and ψ_2. Then $C_1\prec C_2$ if and only if $\psi_1\circ\psi_2^{-1}$ is subadditive.*

Furthermore, $C_1\succ C_2$ if only if $\psi_1\circ\psi_2^{-1}$ is superadditive.

Here, we show by using Theorem A1 that the families of UL- and UIP-independence copulas are ordered in the parameter θ. For the UL-independence copula, $\Psi_1\circ\Psi_2^{-1}=\psi\circ T_1^{-1}\circ T_2\circ\psi^{-1}$, where $\psi(t)=-\log(t)$ and $T(u)=G_L(u)=Q_L^{-1}(u)$. Let $h(u)=T_1^{-1}\circ T_2$. For $\theta_1<\theta_2\le 1$, by Table 1, we derive that

$$h(u)=Q_{L1}\circ G_{L2}(u)=\frac{1}{1+\theta_1(1-u)/(\theta_2 u)}=\frac{\theta_2 u}{(\theta_2-\theta_1)u+\theta_1}$$
$$\Psi_1\circ\Psi_2^{-1}(x+y)=-\log\left(\frac{\theta_2^2 e^{-(x+y)}}{\theta_2\big[(\theta_2-\theta_1)e^{-(x+y)}+\theta_1\big]}\right) \tag{A21}$$
$$\Psi_1\circ\Psi_2^{-1}(x)+\Psi_1\circ\Psi_2^{-1}(y)=-\log\left(\frac{\theta_2 e^{-x}}{\big[(\theta_2-\theta_1)e^{-x}+\theta_1\big]}\frac{\theta_2 e^{-y}}{\big[(\theta_2-\theta_1)e^{-y}+\theta_1\big]}\right) \tag{A22}$$

We wish to compare (A21) and (A22) to determine if $\Psi_1\circ\Psi_2^{-1}$ is subadditive or superadditive. Since $e^{-(x+y)}+1\ge e^{-x}+e^{-y}$ for $x,y\ge 0$, we obtain that

$$\frac{\theta_2}{\theta_2-\theta_1}\ge -e^{x+y}+e^{-x}+e^{-y}+\frac{\theta_1}{\theta_2-\theta_1}$$
$$\text{and } \theta_2 e^{-(x+y)}+\frac{\theta_1\theta_2}{\theta_2-\theta_1}\ge(\theta_2-\theta_1)e^{-(x+y)}+\theta_1 e^{-x}+\theta_1 e^{-y}+\frac{\theta_1^2}{(\theta_2-\theta_1)}.$$

Therefore, (A21) is greater than or equal to (A22). That is, $\Psi_1\circ\Psi_2^{-1}$ is superadditive. Hence, the family of UL-independence copulas is negatively ordered by the parameter θ. Since the UIP distortion gives rise to the same $h(u)$, the family of UIP-independence copulas is also negatively ordered by the parameter θ.

References

1. Genest, C.; Ghoudi, K.; Rivest, L.P. Understanding Relationships Using Copulas by Edward Frees and Emiliano Valdez. *N. Am. Actuar. J.* **1998**, *2*, 143–149. [CrossRef]
2. Ades, M.; Provost, S.B.; Zang, Y. Four Measures of Association and Their Representations in Terms of Copulas. *AppliedMath* **2024**, *4*, 363–382. [CrossRef]
3. Cherubini, U.; Luciano, E.; Vecchiato, W. *Copula Methods in Finance*; John Wiley & Sons: Hoboken, NJ, USA, 2004. [CrossRef]
4. Frees, E.W.; Valdez, E.A. Understanding Relationships Using Copulas. *N. Am. Actuar. J.* **1998**, *2*, 1–25. [CrossRef]
5. Embrechts, P.; McNeil, A.; Straumann, D. Correlation and Dependence in Risk Management: Properties and Pitfalls. In *Risk Management: Value at Risk and Beyond*; Dempster, M.A.H., Ed.; Cambridge University Press: Cambridge, UK, 2002; pp. 176–223.
6. Ji, L.; Tan, K.S.; Yang, F. Tail Dependence And Heavy Tailedness In Extreme Risks. *Insur. Math. Econ.* **2021**, *99* , 282–293. [CrossRef]

7. Alzaid, A.A.; Alhadlaq, W.M.A. New Family of Archimedean Copulas: The Half-Logistic Family of Copulas. *Mathematics* **2024**, *12*, 101. [CrossRef]
8. Tootoonchi, F.; Sadegh, M.; Haerter, J.O.; Raty, O.; Grabs, T.; Teutschbein, C. Copulas for hydroclimatic analysis: A practice-oriented overview. *WIREs Water* **2022**, *9*, e1579. [CrossRef]
9. Chen, L.; Guo, S. *Copulas and Its Application in Hydrology and Water Resources*; Springer: Berlin/Heidelberg, Germany, 2019; 290p; ISBN 978-981-13-0573-3.
10. Sklar, M. Fonctions de Repartition a n Dimensions et Leurs Marges. *Ann. l'ISUP* **1959**, *8*, 229–231.
11. Nelsen, R.B. *An Introduction to Copulas*; Springer: New York, NY, USA, 2006.
12. Joe, H. *Dependence Modeling with Copulas*; CRC Press: Boca Raton, FL, USA, 2015.
13. Yang, Y.; Li, S. On a family of log-gamma-generated archimedean copulas. *N. Am. Actuar. J.* **2022**, *26*, 123–142. [CrossRef]
14. Manstavičius, M.; Bagdonas, G. A class of bivariate independence copula transformations. *Fuzzy Sets Syst.* **2022**, *428*, 58–79. [CrossRef]
15. Chesneau, C. A Collection of New Trigonometric- and Hyperbolic-FGM-Type Copulas. *AppliedMath* **2023**, *3*, 147–174. [CrossRef]
16. Chesneau, C. A Note on Two New Three-parameter Logarithic and Exponential Copulas. *Asia Pac. J. Math.* **2023**, *10*, 8.
17. Di Bernardino, E.; Rulliere, D. On Certain Transformations of Archimedean Copulas: Application to the Non-parametric Estimation of their Generators. *Depend. Model.* **2013**, *1*, 1–36. [CrossRef]
18. Genest, C.; Kilani, G.; Rivest, L.P. A Semiparametric Estimation Procedure of Dependence Parameters in Multivariate Families of Distributions. *Biometrika* **1995**, *82*, 543–552. [CrossRef]
19. Morillas, P.M. A Method To Obtain New Copulas From A Given One. *Metrika* **2005**, *61*, 169–184. [CrossRef]
20. Durante, F.; Foschi, R.; Sarkoci, P. Distorted Copulas: Constructions and Tail Dependence. *Commun. Stat. Theory Methods* **2010**, *39*, 2288–2301. [CrossRef]
21. Valdez, E.A.; Xiao, Y. On The Distortion of a Copula and Its Margins. *Scand. Actuar. J.* **2011**, *4*, 292–317. [CrossRef]
22. Samanthi, R.G.M.; Sepanski, J. A Bivariate Extension Of The Beta Generated Distribution Derived From Copulas. *Commun. Stat. Theory Methods* **2019**, *48*, 1043–1059. [CrossRef]
23. Samanthi, R.G.M.; Sepanski, J. On Bivariate Kumaraswamy-Distorted Copulas. *Commun. Stat. Theory Methods* **2020**, *51*, 2477–2495. [CrossRef]
24. Aldhufairi, F.A.A.; Samanthi, R.G.M.; Sepanski J.H. New Families of Bivariate Copulas via Unit Lomax Distortion. *Risks* **2020**, *8*, 106. [CrossRef]
25. Aldhufairi, F.A.A.; Sepanski, J.H. New Families of Bivariate Copulas via Unit Weibull Distortion. *J. Stat. Distrib. Appl.* **2020**, *7*, 1–20. [CrossRef]
26. Aldhufairi, F.A.A. Utilizing a Unit Gompertz Distorted Copula to Model Dependence in Anthropometric Data. *Commun. Stat. Appl. Methods* **2023**, *30*, 467–483. [CrossRef]
27. Kularatne, T.D.; Li, J.; Pitt, D. On the use of archimedean copulas for insurance modeling. *Ann. Actuar. Sci.* **2021**, *15*, 57–81. [CrossRef]
28. Schweizer, B.; Sklar, A. *Probabilistic Metric Spaces*; Dover Publications: North-Holland, NY, USA, 1983.
29. Hua, L.; Joe, H. Intermediate Tail Dependence: A Review and Some New Results. In *Stochastic Orders in Reliability and Risk Lecture Notes in Statistics*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 291–311.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

AppliedMath

MDPI

Article

SIRS Epidemic Models with Delays, Partial and Temporary Immunity and Vaccination

Benito Chen-Charpentier

Department of Mathematics, University of Texas at Arlington, Arlington, TX 76019, USA; bmchen@uta.edu

Abstract: The basic reproduction, or reproductive number, is a useful index that indicates whether or not there will be an epidemic. However, it is also very important to determine whether an epidemic will eventually decrease and disappear or persist as an endemic. Different infectious diseases have different behaviors and mathematical models used to simulated them should capture the most important processes; however, the models also involve simplifications. Influenza epidemics are usually short-lived and can be modeled with ordinary differential equations without considering demographics. Delays such as the infection time can change the behavior of the solutions. The same is true if there is permanent or temporary immunity, or complete or partial immunity. Vaccination, isolation and the use of antivirals can also change the outcome. In this paper, we introduce several new models and use them to find the effects of all the above factors paying special attention to whether the model can represent an infectious process that eventually disappears. We determine the equilibrium solutions and establish the stability of the disease-free equilibrium using various methods. We also show that many models of influenza or other epidemics with a short duration do not have solutions with a disappearing epidemic. The main objective of the paper is to introduce different ways of modeling immunity in epidemic models. Several scenarios with different immunities are studied since a person may not be re-infected because he/she has total or partial immunity or because there were no close contacts. We show that some relatively small changes, such as in the vaccination rate, can significantly change the dynamics; for example, the existence and number of the disease-free equilibria. We also illustrate that while introducing delays makes the models more realistic, the dynamics have the same qualitative behavior.

Keywords: influenza; epidemic models; differential equations; basic reproduction number; partial immunity

check for updates

Citation: Chen-Charpentier, B. SIRS Epidemic Models with Delays, Partial and Temporary Immunity and Vaccination. *AppliedMath* **2024**, *4*, 666–689. https://doi.org/10.3390/appliedmath4020036

Academic Editor: Carlo Bianca

Received: 7 May 2024
Revised: 17 May 2024
Accepted: 23 May 2024
Published: 28 May 2024

Copyright: © 2024 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Influenza, or flu, is a contagious disease spread by contact with an infected individual. Seasonal influenza affects between 3% and 11% of the population in the United States every year [1]. The flu peaks from December to February of each year. Although some patients die from seasonal flu, it is not a particularly deadly disease with 0.2 deaths per 100,000 population [2]. Other contagious diseases that are short-lived are the common cold, Respiratory Syncytial Virus Disease and Streptococcus throat. There are many papers dealing with mathematical models for epidemics, both via direct contact and vector transmitted. Mathematical modeling of epidemics started in the early 20th century. One of the first efforts is [3] and the seminal paper of Kermack and McKendrick [4] is widely recognized as introducing the first compartmental epidemic model based on ordinary differential equations (ODEs). The use of ODEs is a good modeling assumption if the different populations are well-mixed and there is little or no spatial dependence. Kermack and McKendrick introduced the SIR (susceptible-infected-removed) model, which assumes that the population can be in one of three states: susceptible, infected or infective, and removed (recovered with permanent immunity or deceased). The SIR model and its many variations including the SIS (susceptible-infective-susceptible) and SIRS (susceptible-infective-recovered-susceptible)

models have been very influential in understanding epidemics and their control. Among the modifications made to the original SIR model of Kermack and McKendrick, are the introduction of demographics (not important for seasonal flu), vaccinations, different age groups, use of anti-virals and isolation. One important modification is the addition of delays. Delays appear in epidemics due to the time it takes certain processes to occur, like the time it takes a susceptible to become infected after a contact with and infected individual, or the time it takes vaccines to become effective after being administered. The two main ways of introducing delays into a model based on ordinary differential equations are by using delay differential equations (DDEs) or by introducing and additional exposed or latent population. DDEs may have different dynamics than ODEs [5,6]. Among the possible differences are the introduction of oscillations, introduction of discontinuities in the time derivatives, non-uniqueness and different stability. While there is a wide range of literature about models of epidemics or disease transmission using delay differential equations, most books on DDEs have sparse coverage of contagious disease spread, and books on epidemics only have little material on DDEs. Two exceptions are the following books [7,8]. SIR models consider that after being infected, individuals recover and achieve permanent and total immunity from the disease. For seasonal influenza and other contagious diseases the immunity is temporal which leads to SIRS models in which the recovered individuals lose immunity completely after a certain time period. However, there are other possibilities: the individuals whom recovered lose total immunity but keep partial immunity forever; the individuals whom recovered lose total immunity and achieve partial immunity for a time period; some individuals whom recovered keep total and permanent immunity while others only keep partial permanent immunity, and so on. What happens depends on the particular infection and population. Vaccines may also confer total permanent immunity, or total immunity that wanes to partial immunity, or other scenarios.

One of the most important uses of mathematical models of contagious diseases is to predict whether there will be an epidemic or not. The basic reproduction number R_0 defined as the number of secondary infections produced by a newly infected individual at the start of the infective process is a useful index to determine whether the epidemic will die or not. It is usually calculated using the next generation matrix method [9,10]. It also gives the stability of the disease-free equilibrium (DFE) of the model. In case that the number of infectives grows initially, it is also very important to determine if the epidemic will eventually die or not. In this paper, we will develop mathematical models of short-lived contagious diseases liked seasonal influenza that include different assumptions about total, partial, permanent and temporal immunity. We will also add vaccinations with different immunity effects, and also add the use of anti-virals and isolation. The models will include delayed versions, using DDEs and also exposed populations. For these models we will study whether the epidemic persists or not. Ref. [11] compares some basic models for influenza that we will use as a starting point. The rest of the paper is organized as follows. In Section 2, we present several new models, that are extensions to the SIRS model, with different hypotheses about total and partial immunity and vaccination, determine their equilibria, and analyze the stability of their disease-free equilibria. In Section 3, numerical simulations and results are presented, including the effects of introducing delays. We also show that some relatively small changes in the models can significantly change the dynamics. Finally, in Section 4, the discussions and conclusions about the models and results are presented.

2. Materials and Methods

In this section, we introduce several new mathematical models of short-term epidemics with different types of immunities. We also add vaccination, the use of anti-virals and isolation or quarantine. To determine if an epidemic develops according to the model used, the equilibrium points need to be found. If there is an equilibrium point with no infectives and susceptibles equal to the total population, and if the number of infectives grows with time, then there will be an epidemic. An equilibrium with zero infectives is

known as a disease-free equilibrium (DFE). There are different methods to determine if the number of infectives grows: the differential equation for the infectives has a positive right-hand side; the DFE is locally unstable; using the basic reproduction number R_0 gives $R_0 > 1$; the infective compartments subsystem of equations is locally unstable; the R_0 index calculated using the next generation matrix method is greater than 1 [10,12]. In case there is an epidemic, it is very important to establish if it will eventually die. This can be calculated by determining if there is a DFE for positive time that is locally asymptotically stable. This may happen for a finite time or only as time tends to infinity. This determination will be conducted in the following subsections for different models with no demographics. The first subsection is a review of the susceptible-infective-susceptible model (SIS), the susceptible-infective-recovered model (SIR) and the susceptible-infective-recovered-susceptible model (SIRS) with different immunities.

2.1. SIS, SIR and SIRS Models

A very simple epidemic model, for a total constant population N, consisting of only two population compartments, susceptibles S and infectives I is the SIS model. It assumes that a susceptible gets infected with probability $\beta/N > 0$ by a contact with an infective, and then recovers with no immunity and returns to the susceptible compartment. The common cold is an example of a contagious disease that can be modeled by a SIS system of equations. The total population is usually considered constant which assumes that the disease is fast compared to the demographic effects. An almost equivalent assumption, since the death rate is very small compared to the time frame of the disease, is the number of births is equal to the number of deaths with no migration. The SIS model with no demographics is [13,14]

$$\begin{aligned} \frac{dS}{dt} &= -\frac{\beta}{N}SI + \gamma I \\ \frac{dI}{dt} &= \frac{\beta}{N}SI - \gamma I, \end{aligned} \tag{1}$$

where $\gamma > 0$ is the recovery rate and $N > 0$ is the total population. Since the population is constant, $S(t) = N - I(t)$ and we can eliminate one equation

$$\frac{dI}{dt} = \frac{\beta}{N}(N - I)I - \gamma I. \tag{2}$$

The SIS model has two equilibrium states, $I = 0$ and $I = (1 - \frac{\gamma}{\beta})N$. The last one is physical only for $\frac{\gamma}{\beta} \leq 1$. The basic reproduction number (also called basic reproductive number) R_0 is the number of secondary infectives generated by a single infective at the beginning of the infection process. For the SIS model it can be calculated as the product of the probability of a new infective, β/N, times the number of susceptibles at the beginning of the infective process, N, times the average lifetime of an infection, $1/\gamma$. Therefore $R_0 = \beta/\gamma$. If $R_0 > 1$ the number of infectives increases away from the DFE, and the other equilibrium makes biological sense since all the states are non-negative. If $R_0 < 1$ the infection disappears and the DFE equilibrium is the only one that makes biological sense. $R_0 > 1$ implies that the DFE is locally asymptotically unstable so at least one the eigenvalues of the linearized system about the DFE has positive real part. To determine if the infection is growing it is only necessary to look at the infected compartment, $I(t)$. The next generation matrix method [9,10] is widely used to calculate the basic reproduction number in more complex models in epidemiology. Applying this method to the SIS model there is only one infective compartment, I, so the Jacobian J of this system is a 1×1 matrix. Decomposing this matrix as $J = T + \Sigma$, with T the transmission and Σ the transition matrices, produces $T = \beta S/N$ and $\Sigma = -\gamma$. The next generation matrix is $K = -T\Sigma^{-1} = \beta S/(\gamma N)$. At the beginning of the infection $S = N - 1$, so the spectral radius of K, which is R_0, can be approximated by $R_0 = \beta/\gamma$. The local stability of the DFE can also be established by finding an index that states when the eigenvalue of the linearized system consisting of the

disease compartment evaluated at the DFE has negative real part. A straightforward way of finding whether the number of infectives grows at the start of the epidemic is to the evaluate the differential equation for I at the initial state with $S = N$:

$$\frac{dI}{dt} = \frac{\beta}{N} NI - \gamma I = (\beta - \gamma) I,$$

which is positive for $\beta/\gamma > 1$. Note that since at the start of the epidemic $I = 1$, $S = N - 1$. However, for large N, S can be taken equal to N. Another way of finding the stability of the DFE at the start of the infection can is by calculating the eigenvalues of the Jacobian matrix corresponding to the full system of equations, but the formulas are usually complicated or cannot be determined analytically. Using the characteristic polynomial and the Routh-Hurwitz criterion is another possibility [13,15].

Two delays may be added to the SIS model, τ the time in takes the disease to develop in a susceptible individual after being infected by a contact with an infective individual, and the time it takes for a newly infected individual to recover. This second delay is usually included implicitly since $1/\gamma$ is the average recovery time. There are two ways to introduce the first delay. The first way is given by SIS Delay Model I:

$$\begin{aligned} \frac{dS}{dt} &= -\frac{\beta}{N} S(t-\tau) I(t-\tau) + \gamma I(t) \\ \frac{dI}{dt} &= \frac{\beta}{N} S(t-\tau) I(t-\tau) - \gamma I(t). \end{aligned} \tag{3}$$

The assumption is that a contact at $t - \tau$ produces a new infective at time t. The second way is given by SIS Delay Model II:

$$\begin{aligned} \frac{dS}{dt} &= -\frac{\beta}{N} S(t) I(t-\tau) + \gamma I(t) \\ \frac{dI}{dt} &= \frac{\beta}{N} S(t) I(t-\tau) - \gamma I(t). \end{aligned} \tag{4}$$

The assumption now is that an individual susceptible after contact with an infective turns into an infective but takes it a time τ to actually become infected with the disease and be able to infect. Both models make biological sense. Another approach to studying epidemic models is to consider them as chemical reactions, as shown, for example, in [16]. As an illustration, the reaction between an infective and a susceptible produces two infectives. For the reaction to happen both species should be at the same time, which is a caveat for Delay Model II. However, since both ways of introducing the infection delay are used in many papers, we include both to see if they produce the same dynamics qualitatively. The dynamics may be different since their characteristic equations are different. For other epidemic models that introduce the delay in the infection as was performed in Delay Model 1 see, for example, [17–19]. And for other applications using the delay as in Delay Model 2 see [20,21].

For the SIS model, the equilibrium solutions of both delay models are the same as those of the ODE model. To determine the stability of the DFE for model (3), let $I = i$ and $S = N - s$. Then the linearized differential equation for i is $\frac{di}{dt} = \beta i - \gamma i$. By taking $i = \exp(-r\tau)$ we obtain the characteristic equation

$$r = \beta \exp(-r\tau) - \gamma.$$

This equation is transcendental and there is no exact solution, but $r > 0$ only for $\beta > \gamma$. So we have the same R_0 as the for ODE model. A similar calculation for model (4) yields the same result. This method has been applied to other models by [20,22].

Another way of introducing the time it takes the infection to develop is to add another compartment for the exposed or latent population (E). After a contact with an infective, the susceptible individual goes into the exposed compartment and spends an average time equal to the infection time there. Exposed individuals cannot infect. The SIS model with an exposed population is

$$\begin{aligned} \frac{dS}{dt} &= -\frac{\beta}{N}SI + \gamma I \\ \frac{dE}{dt} &= \frac{\beta}{N}SI - \frac{1}{\tau}E \\ \frac{dI}{dt} &= \frac{1}{\tau}E - \gamma I. \end{aligned} \tag{5}$$

The DFE equilibrium of (5) is the same as that of (1). Its stability can be calculated using the next generation matrix method and gives $R_0 = \beta/\gamma$. So the stability is the same as that of (1). Exposed populations have been used in many epidemic models such as those in [23,24].

The SIR model of Kermack and McKendrick [4] considers that and infected recovers with total and permanent immunity. It adds the recovered (R) compartment to the SIS model:

$$\begin{aligned} \frac{dS}{dt} &= -\frac{\beta}{N}SI \\ \frac{dI}{dt} &= \frac{\beta}{N}SI - \gamma I \\ \frac{dR}{dt} &= \gamma I. \end{aligned} \tag{6}$$

Since the total population is constant, one equation can be eliminated using $R(t) = N - S(t) - I(t)$. The new system of equations is

$$\begin{aligned} \frac{dS}{dt} &= -\frac{\beta}{N}SI \\ \frac{dI}{dt} &= \frac{\beta}{N}SI - \gamma I. \end{aligned}$$

The only steady solution of (6) is $I = 0, S + R = N$. At the start of the infection $S(0) = N$. The second equation of (6) states that at the initial time, $t = 0$, $\frac{dI}{dt} > 0$ for $\beta > \gamma$. So for $\beta/\gamma > 1$ the DFE is unstable, and $I(t)$ has its maximum value when $S = N\gamma/\beta$. Furthermore, the solution as t tends to infinity is $S(\infty) = N - R(0) - \frac{\gamma N}{\beta}\ln(\frac{S(\infty)}{S(0)})$ and $S(\infty) > 0$ [13]. So if there is an epidemic it does not disappear.

The delay due to the time of infection can be introduced in the two ways as conducted for the SIS model. The equilibrium solution of both delay models is the DFE with the same stability as for the no delay model. The same stability result for the DFE of the no delay ODE model is true for the SIR model with the exposed population.

A model for propagation of many different diseases in a population is the SIRS model with no demographics [13]. It consists of three populations: susceptibles ($S(t)$), infectives or infected ($I(t)$) and recovered ($R(t)$). A susceptible turns into an infective with a given probability after a contact with an infective. An infective recovers with a rate γ. A recovered loses immunity at a rate ν and converts into a susceptible. The total population is assumed constant, $S(t) + I(t) + R(t) = N$, and the disease is assumed to be short enough so births and deaths need not be considered. The infection coefficient will be denoted by β/N. All parameters are positive. The system of ordinary differential equations (ODEs) describing the model is [13,14]

$$\begin{aligned}\frac{dS}{dt} &= -\frac{\beta}{N}SI + \nu R\\ \frac{dI}{dt} &= \frac{\beta}{N}SI - \gamma I\\ \frac{dR}{dt} &= \gamma I - \nu R.\end{aligned} \tag{7}$$

Since the total population is constant, one equation can be eliminated using $R(t) = N - S(t) - I(t)$. The new system of equations is

$$\begin{aligned}\frac{dS}{dt} &= -\frac{\beta}{N}SI + \nu(N - S - I)\\ \frac{dI}{dt} &= \frac{\beta}{N}SI - \gamma I.\end{aligned} \tag{8}$$

The SIRS model has two equilibrium solutions: the DFE, $S = N, I = 0, R = 0$, and the endemic equilibrium, $S = \gamma N/\beta, I = \frac{(\beta-\gamma)N\nu}{\beta(\gamma+\nu)}, R = N - S - I$. The basic reproduction number is $R_0 = \beta/\gamma$.

There are several different methods of introducing delays in the infection into the model: The first one which we will call SIRS Model I [17–19] is given by

$$\begin{aligned}\frac{dS}{dt} &= -\frac{\beta}{N}S(t-\tau)I(t-\tau) + \nu(N - S(t) - I(t))\\ \frac{dI}{dt} &= \frac{\beta}{N}S(t-\tau)I(t-\tau) - \gamma I(t),\end{aligned} \tag{9}$$

where the hypothesis is that an infective has contact with a susceptible and it takes a time τ for this susceptible to turn infective. It is not counted as a susceptible until that time. A second model is the SIRS Model II [25] described by:

$$\begin{aligned}\frac{dS}{dt} &= -\frac{\beta}{N}S(t)I(t-\tau) + \nu(N - S(t) - I(t))\\ \frac{dI}{dt} &= \frac{\beta}{N}S(t)I(t-\tau) - \gamma I(t),\end{aligned} \tag{10}$$

where the assumption is that after a contact between a susceptible and an infective, a time τ passes before the susceptible turns into an infective and is then able to infect another susceptible.

Instead of considering a recovered population, where individuals spend an average time $1/\nu$ before losing immunity, infective individuals after stopping being infective have to wait a delay time $1/\nu$ before turning infectives. In [26], this is performed in addition to keeping the recovered compartment.

The two models with delays introduced above have the same equilibrium solutions but the actual time dependent solutions will be different. There is also the question of the influence of the initial conditions given in the history interval. For an epidemic they can always be given as the equilibrium solutions with no disease. That is, the natural assumption that the epidemic starts at $t = 0$ can be used.

A model that takes into account the dependence in the past, caused by the delay in the transmission, by including an exposed population, E, is [27,28]

$$\begin{aligned} \frac{dS}{dt} &= -\frac{\beta}{N}SI + \nu R \\ \frac{dE}{dt} &= \frac{\beta}{N}SI - \sigma E \\ \frac{dI}{dt} &= \sigma E - \gamma I \\ \frac{dR}{dt} &= \gamma I - \nu R. \end{aligned} \tag{11}$$

The additional parameter is $\sigma = 1/\tau$. Models of this form are usually called SEIRS models. It has the same DFE and R_0 as the non-delay model.

2.2. Models with Different Immunities

The three models given in Section 2.1 have different types of immunity. The SIS model has no immunity, the SIR model has total and permanent immunity, and the SIRS model has total but temporal immunity. However, for different infections there may be different types of immunity such as partial and permanent or partial and temporal. Some papers considering immunity for influenza are [29–31]. Ref. [32] includes a reinfected class with different immunity. Ref. [33] models partial immunity of the recovered population and [34] includes that recovered lose immunity but not all infectives recover. Ref. [35] adds the effect of reinfection of influenza. Ref. [36] also takes into account waning immunity. Ref. [37] models the loss of immunity with time-dependent rates. We will consider several new models with different immunity based on the SIRS model. The first model assumes all infective individuals get partial but permanent immunity. Similar models are given by [34,38]. Susceptibles get infected and then move into susceptibles with partial immunity S_p who can be reinfected with infection rate β_p/N. The equations of the model are

$$\begin{aligned} \frac{dS}{dt} &= -\frac{\beta}{N}SI \\ \frac{dI}{dt} &= \frac{\beta}{N}SI + \frac{\beta_p}{N}S_pI - \gamma I \\ \frac{dS_p}{dt} &= -\frac{\beta_p}{N}S_pI + \gamma I. \end{aligned} \tag{12}$$

Since the total population is constant, one equation can be eliminated using $S_p(t) = N - S(t) - I(t)$. The system has two equilibrium solutions: the DFE $I = 0, S + S_p = N$, and $S = 0, S_p = \frac{\gamma N}{\beta_p}, I = N - S_p$. At the start of the infection, $S_p = 0$ so the equation for I is the same as that in system (6), and R_0 has the same value $R_0 = \beta/\gamma$. The two delay models are constructed as before. The first model is

$$\begin{aligned} \frac{dS}{dt} &= -\frac{\beta}{N}S(t-\tau)I(t-\tau) \\ \frac{dI}{dt} &= \frac{\beta}{N}S(t-\tau)I(t-\tau) + \frac{\beta_p}{N}S_p(t-\tau)I(t-\tau) - \gamma I \\ \frac{dS_p}{dt} &= -\frac{\beta_p}{N}S_p(t-\tau)I(t-\tau) + \gamma I, \end{aligned} \tag{13}$$

and the second is

$$\begin{aligned}\frac{dS}{dt} &= -\frac{\beta}{N}S(t)I(t-\tau)\\ \frac{dI}{dt} &= \frac{\beta}{N}S(t)I(t-\tau) + \frac{\beta_p}{N}S_p(t)I(t-\tau 1) - \gamma I\\ \frac{dS_p}{dt} &= -\frac{\beta_p}{N}S_p(t)I(t-\tau) + \gamma I.\end{aligned} \tag{14}$$

They have the same equilibrium points as the no-delayed model. The model with the exposed population is constructed similarly and is given by

$$\begin{aligned}\frac{dS}{dt} &= -\frac{\beta}{N}SI\\ \frac{dE}{dt} &= \frac{\beta}{N}SI + \frac{\beta_p}{N}S_pI - \sigma E\\ \frac{dI}{dt} &= \sigma E - \gamma I\\ \frac{dS_p}{dt} &= -\frac{\beta_p}{N}S_pI + \gamma I.\end{aligned} \tag{15}$$

All these three models also have the same R_0 as (12).

The second partial immunity model assumes that all infectives get partial immunity but that it is only temporal. Susceptibles get infected, then go to susceptibles with partial immunity and then back to susceptibles with no immunity. The model is given by the following system of ODEs

$$\begin{aligned}\frac{dS}{dt} &= -\frac{\beta}{N}SI + \nu S_p\\ \frac{dI}{dt} &= \frac{\beta}{N}SI + \frac{\beta_p}{N}S_pI - \gamma I\\ \frac{dS_p}{dt} &= -\frac{\beta_p}{N}S_pI + \gamma I - \nu S_p.\end{aligned} \tag{16}$$

Again since the total population is constant, one equation can be eliminated. The system has two equilibrium solutions: the DFE $I = 0, S = N, S_p = 0$, and a second equilibrium with $S > 0, S_p > 0$ and $I > 0$. At the start of the infection, $S_p = 0$ so the equation for I is the same as that in system (6) and $R_0 = \beta/\gamma$. The two delay models are similarly constructed as before. They have the same equilibrium points as the non-delayed model. The model with the exposed population has this population added in the same form. All these three models also have the same R_0 as (12).

The third model assumes that all infectives get total immunity but that it is only temporal, and then get permanent partial immunity. Susceptibles get infected, then recover with total temporal immunity, then get partial permanent immunity as susceptibles with partial immunity. The model is given by

$$\begin{aligned}\frac{dS}{dt} &= -\frac{\beta}{N}SI\\ \frac{dI}{dt} &= \frac{\beta}{N}SI + \frac{\beta_p}{N}S_pI - \gamma I\\ \frac{dR}{dt} &= \gamma I - \nu R\\ \frac{dS_p}{dt} &= -\frac{\beta_p}{N}S_pI + \nu R.\end{aligned} \tag{17}$$

Again since the total population is constant, one equation can be eliminated. The system has two equilibrium solutions: the DFE $I = 0, S + S_p = N$, and a second equilibrium $S = 0, S_p = \frac{\gamma N}{\beta_p}, I = \frac{(\beta_p - \gamma)N\nu}{\beta_p(\gamma+\nu)}$ and $R = N - S_p - I$. This second equilibrium solution only makes sense for $\beta_p > \gamma$. At the start of the infection, $S_p = 0$ so the equation for I is the same as that in system (6) and $R_0 = \beta/\gamma$. However, if $\beta_p < \gamma$, the epidemic will eventually disappear.

The fourth model assumes that all infectives get total immunity but that it is only temporal, and then get partial immunity which is also temporal. Susceptibles get infected, then recover with total temporal immunity, then get partial temporal immunity as susceptibles with partial immunity, that then convert to susceptibles. The model is given by

$$\begin{aligned} \frac{dS}{dt} &= -\frac{\beta}{N}SI + \nu_1 S_p \\ \frac{dI}{dt} &= \frac{\beta}{N}SI + \frac{\beta_p}{N}S_p I - \gamma I \\ \frac{dR}{dt} &= \gamma I - \nu R \\ \frac{dS_p}{dt} &= -\frac{\beta_p}{N}S_p I + \nu R - \nu_1 S_p. \end{aligned} \tag{18}$$

Again since the total population is constant, one equation can be eliminated. The system has two equilibrium solutions: the DFE $I = 0, S + S_p = N$ and a second endemic equilibrium $S > 0, S_p > 0, I > 0$ and $R = N - S_p - I$. At the start of the infection, $S_p = 0$ so the differential equation for I is the same as that in system (6) and $R_0 = \beta/\gamma$.

2.3. Models with Loss of Immunity and Two Groups

Different individuals lose immunity at different rates. In this subsection for simplicity we will include two groups of individuals, but the models can be extended to more groups. Ref. [34] has a model with some recovered losing immunity. The first proposed model assumes that all infected get permanent immunity but the two groups lose it at different rates, ν and ν_1. One group has a stronger immune system due to among other things to previous exposure to a previous variation of influenza or vaccinations on previous seasons [29–31].

$$\begin{aligned} \frac{dS}{dt} &= -\frac{\beta}{N}SI + \nu f R + \nu_1(1-f)R \\ \frac{dI}{dt} &= \frac{\beta}{N}SI - \gamma I \\ \frac{dR}{dt} &= \gamma I - \nu f R - \nu_1(1-f)R. \end{aligned} \tag{19}$$

Here, $0 \leq f \leq 1$ is the fraction of the population in the first group. The two models with delay in the infection are given by

$$\begin{aligned} \frac{dS}{dt} &= -\frac{\beta}{N}S(t-\tau)I(t-\tau) + \nu f R + \nu_1(1-f)R \\ \frac{dI}{dt} &= \frac{\beta}{N}S(t-\tau)I(t-\tau) - \gamma I \\ \frac{dR}{dt} &= \gamma I - \nu f R - \nu_1(1-f)R, \end{aligned} \tag{20}$$

and

$$\begin{aligned}\frac{dS}{dt} &= -\frac{\beta}{N}S(t)I(t-\tau) + \nu f R + \nu_1(1-f)R\\ \frac{dI}{dt} &= \frac{\beta}{N}S(t)I(t-\tau) - \gamma I\\ \frac{dR}{dt} &= \gamma I - \nu f R - \nu_1(1-f)R.\end{aligned} \tag{21}$$

For both delay models, the differential equation at the beginning of the infection process for the infective compartment is the same as that of the SIS model (1), therefore the same procedure can be used to determine that $R_0 = \beta/\gamma$. Also, for both models the equilibrium points are the same as those of the ODE model (19).

The model with an exposed population is

$$\begin{aligned}\frac{dS}{dt} &= -\frac{\beta}{N}SI + \nu f R + \nu_1(1-f)R\\ \frac{dE}{dt} &= \frac{\beta}{N}SI - \sigma E\\ \frac{dI}{dt} &= \sigma E - \gamma I\\ \frac{dR}{dt} &= \gamma I - \nu f R - \nu_1(1-f)R.\end{aligned} \tag{22}$$

It has the same DFE as (19). The stability of the DFE at the beginning of the infection is still given by $R_0 = \beta/\gamma$.

In the next model, a fraction f of the infectives gets total and permanent immunity, and the rest gets partial and permanent immunity. The ODEs of the model are

$$\begin{aligned}\frac{dS}{dt} &= -\frac{\beta}{N}SI\\ \frac{dI}{dt} &= \frac{\beta}{N}SI + \frac{\beta_p}{N}S_pI - \gamma I\\ \frac{dR}{dt} &= \gamma f I\\ \frac{dS_p}{dt} &= -\frac{\beta_p}{N}S_pI + \gamma(1-f)I.\end{aligned} \tag{23}$$

The next model assumes that a fraction f gets total and temporal immunity and the rest get partial and temporal immunity. The system of equations is

$$\begin{aligned}\frac{dS}{dt} &= -\frac{\beta}{N}SI + \nu R + \nu_1 S_p\\ \frac{dI}{dt} &= \frac{\beta}{N}SI + \frac{\beta_p}{N}S_pI - \gamma I\\ \frac{dR}{dt} &= \gamma f I - \nu R\\ \frac{dS_p}{dt} &= -\frac{\beta_p}{N}S_pI + \gamma(1-f)I - \nu_1 S_p.\end{aligned} \tag{24}$$

In the next model all infectives get total immunity for a finite time, then lose it different rates ν and ν_1, and get partial and permanent immunity with different rates of infection β_1 and β_2.

$$\begin{aligned}
\frac{dS}{dt} &= -\frac{\beta}{N}SI \\
\frac{dI}{dt} &= \frac{\beta}{N}SI + \frac{\beta_1}{N}S_1I + \frac{\beta_2}{N}S_2I - \gamma I \\
\frac{dR}{dt} &= \gamma I - \nu f R - \nu_1(1-f)R \\
\frac{dS_1}{dt} &= -\frac{\beta_1}{N}S_1I + \nu f R \\
\frac{dS_2}{dt} &= -\frac{\beta_2}{N}S_2I + \nu_1(1-f)R.
\end{aligned} \tag{25}$$

2.4. Models with Vaccination and Immunity Loss

Vaccination is a very effective tool for achieving immunity to seasonal influenza. Usually the effect of the vaccine wanes with time. We consider that the immunity is permanent if it lasts until the next flu season with a new variant of the virus. The vaccinated population will be denoted by V. There are different ways of introducing the vaccination rate. The first way is that it is proportional to the number of susceptibles S [39–41]. Refs. [42–44] also consider the waning of the vaccine. Ref. [45] has a basic vaccination model based on the SIR model. A second way is to assume that only a percentage p_r of susceptibles is willing to be vaccinated [46] which is a more realistic assumption [47]. This leads to a vaccination rate of the form $r(p_rS - V)$. It can also be considered proportional to the total population or just constant [48,49]. Ref. [50] considers vaccination, partial immunity and reinfection. Refs. [40,51] also include quarantine. Our first model with vaccination assumes infectives recover and get total and temporal immunity for an average length of time $1/\nu$. Vaccinated individuals get total immunity but it wanes with time at a rate ω. The model is

$$\begin{aligned}
\frac{dS}{dt} &= -\frac{\beta}{N}SI + \omega V + \nu R - rS \\
\frac{dV}{dt} &= rS - \omega V \\
\frac{dI}{dt} &= \frac{\beta}{N}SI - \gamma I \\
\frac{dR}{dt} &= \gamma I - \nu R.
\end{aligned} \tag{26}$$

In this model the rate of vaccination is proportional to the number of susceptibles. Since the equation for the infective population is the same as that for (6), the value of R_0 is still β/γ. If the vaccines does not wane, $\omega = 0$, the system has only one equilibrium point, $S = 0, I = 0, R = 0, V = N$. If $\omega \neq 0$ there are two equilibrium points, $I = 0, R = 0$, $S = (\omega N/(\omega + r)$, $V = rS/\omega$ and $S = \gamma N/\beta, V = rS/\omega, R = \gamma I/\nu$ and I given by the solution to $I = N - S - V - R$. The first delay model is obtained by replacing $S(t)$ and $I(t)$ in the infection terms by $S(t-\tau)$ and $I(t-\tau)$. The second one by only replacing the $I(t)$ term. Both have the same equilibrium points and R_0 as (26). The system with an exposed population has the same DFE equilibrium points since $I = 0$ implies $E = 0$. The R_0 calculated by the next generation matrix method is still $R_0 = \beta/\gamma$.

The second model with vaccination considers that the infective population gets total immunity for a fixed time and then gets partial permanent immunity. The vaccinated population gets total but temporal immunity, and then it wanes to partial permanent immunity. There is a new compartment, S_p, representing susceptibles with partial immunity.

$$\begin{aligned}
\frac{dS}{dt} &= -\frac{\beta}{N}SI - rS \\
\frac{dS_p}{dt} &= -\frac{\beta_p}{N}S_pI + \omega V + \nu R \\
\frac{dV}{dt} &= rS - \omega V \\
\frac{dI}{dt} &= \frac{\beta}{N}SI + \frac{\beta_p}{N}S_pI - \gamma I \\
\frac{dR}{dt} &= \gamma I - \nu R.
\end{aligned} \tag{27}$$

This model has a DFE $I = 0, S = 0, R = 0, V = 0, S_p = N$. There is also an endemic equilibrium with $S = 0, V = 0$. Since at the beginning of the infection, $S_p = 0$, using the differential equation for I, $R_0 = \beta/\gamma$. The two delay models are obtained modifying (27) in a similar way as was performed for previous models. The equilibrium states and R_0 are the same as those of (27). The corresponding equation with an exposed population also has the same R_0.

If the vaccination rate is modified to $r(p_rS - V)$ instead of rS, to make the model agree with not everybody wanting to be vaccinated, the modified model is

$$\begin{aligned}
\frac{dS}{dt} &= -\frac{\beta}{N}SI - r(p_rS - V) \\
\frac{dS_p}{dt} &= -\frac{\beta_p}{N}S_pI + \omega V + \nu R \\
\frac{dV}{dt} &= r(p_rS - V) - \omega V \\
\frac{dI}{dt} &= \frac{\beta}{N}SI + \frac{\beta_p}{N}S_pI - \gamma I \\
\frac{dR}{dt} &= \gamma I - \nu R.
\end{aligned} \tag{28}$$

This model does not have a steady solution with $I = 0$ for arbitrary values of the parameters. This an important difference in the dynamics with those of the other models.

The next model considers that all vaccinated get total immunity for a time, then a fraction f_v keeps the total immunity, while the rest gets partial permanent immunity. All infectives get total temporal immunity after they recover.

$$\begin{aligned}
\frac{dS}{dt} &= -\frac{\beta}{N}SI - r(p_rS - V) + \nu R \\
\frac{dS_p}{dt} &= -\frac{\beta_p}{N}S_pI + \omega(1 - f_v)V \\
\frac{dV}{dt} &= r(p_rS - V) - \omega(1 - f_v)V \\
\frac{dI}{dt} &= \frac{\beta}{N}SI + \frac{\beta_p}{N}S_pI - \gamma I \\
\frac{dR}{dt} &= \gamma I - \nu R.
\end{aligned} \tag{29}$$

If $\omega = 0$ the DFE is $I = 0, R = 0, , S_p = 0, V = p_rS, S = N/(1 + p_r)$. For $\omega \neq 0$ the DFE is $I = 0, R = 0, V = 0, S = 0, S_p = N$. Since $S_p(0) = 0$, we have for both case $R_0 = \beta/\gamma$. The two delay models and the exposed population models are derived from (29) as was performed for the previous models. All three models have the same DFE and R_0 as (29).

As a final and more complete model we will add the use of antivirals to reduce the infection time and isolation to reduce the number of active infectives. There is a new compartment, the quarantined or isolated population Q. Infectives recover with total and temporal immunity, vaccinated get total and temporal immunity and then partial permanent immunity. A fraction f_a of the infected gets an antiviral and exit the compartment at a faster rate γ_a. A fraction f_q of the infected gets quarantined at a rate γ_q and leaves the quarantine state at a rate γ_{q1}. The average time that a quarantined individual spends in the infected and quarantined compartments is the same as the time a non-quarantined individual spends in the infective compartment. Therefore, $1/\gamma = 1/\gamma_q + 1/\gamma_{q1}$. The model is

$$\begin{aligned}
\frac{dS}{dt} &= -\frac{\beta}{N}SI - r(p_r S - V) + \nu R \\
\frac{dS_p}{dt} &= -\frac{\beta_p}{N}S_p I + \omega(1-f_v)V \\
\frac{dV}{dt} &= r(p_r S - V) - \omega(1-f_v)V \\
\frac{dI}{dt} &= \frac{\beta}{N}SI + \frac{\beta_p}{N}S_p I \\
&\quad - f_q\gamma_q I - f_a\gamma_a I - (1-f_q-f_a)\gamma I \\
\frac{dQ}{dt} &= f_q\gamma_q I - \gamma_{q1}Q \\
\frac{dR}{dt} &= f_a\gamma_a I + (1-f_q-f_a)\gamma I + \gamma_{q1}Q - \nu R.
\end{aligned} \tag{30}$$

The DFE is $I = 0, R = 0, Q = 0, V = 0, S = 0, S_p = N$. Since $Sp(0) = 0$, for the infective compartment to grow we need $R_0 = \frac{\beta}{(1-f_q-f_a)\gamma + f_q\gamma_q + f_a\gamma_a} > 1$. The delay versions and the model with exposed population have the same DFE and R_0.

3. Results

In this section, we present numerical results for the models given in the previous section. The numerical simulations, including the DDE were performed using the *julia* package DifferentialEqns.jl [52,53]. The parameter values used for most simulations are given in Table 1. If a different value is used for a particular simulation, it will be noted in the caption of the corresponding figure. The values of β, γ, ν and τ are based on those given [54] and CDC [1] corresponding to seasonal influenza. All the others are estimated. The value of β has a large variation depending on the conditions: rural, urban, suburban, high density work places, etc. The values of β and γ chosen for most simulations give $R_0 = 1.5$. Ref. [55] also gives similar parameter values for influenza.

Table 1. This table has the parameter values used in the simulations.

Symbol	Parameter Description	Value
β	Infection rate	0.5/d
γ	Recovery rate	1/3/d
ν	Loss of immunity rate	1/30/d
τ	Delay in time of infection	1.9 d
σ	1/(time of infection)	$1/\tau$1/d
N	Total population	1000
β_p	Infection rate for partial immunity	$\beta/2$
ν_1	Rate of loss or partial immunity	1/60/d
f_r	Fraction losing immunity at rate ν	0.75
β_1	Infection rate for f_r fraction	0.9β/d
β_2	Infection rate for $1 - f_r$ fraction	0.8β/d

Table 1. *Cont.*

Symbol	Parameter Description	Value
f_p	Fraction achieving total immunity	0.5
r	Vaccination rate	0.02/d
ω	Rate of loss of vaccination immunity	1/60/d
β_v	Infection rate of vaccinated with partial	0.8β
p_r	Fraction of susceptibles to be vaccinated	0.2/d
f_v	Fraction of vaccinated with permanent immunity	0.75

The initial conditions for all models are $S = 900$, $I = 100$ and all other populations equal to zero. The history for the delay models is $S = 1000$, and all other delayed variables equal to zero in the interval $[-\tau, 0]$. This corresponds to an epidemic starting at $t = 0$. All simulations, except where specified, have $R_0 = 1.5$.

3.1. SIS, SIR and SIRS Models

The SIS, SIR and SIRS models are widely used for epidemics with infection by direct contact. In the SIS model, infectives do not get immunity, in the SIR model they get total and permanent immunity and in the SIRS model they get total but temporal immunity.

Figure 1 presents the plots for the SIS model, in the left for the susceptible population and in the right for the infective population. Since $R_0 = \beta/\gamma > 1$ the solution tends to the endemic state. The infective plot for the model with the exposed population has a different limit since there are two infective classes, I and E, and only I is plotted. The solutions have the same form but the one for the ODE model reaches the equilibrium state the fastest. The solutions of the two delay system are all identical since the only difference is when a new infected stops being counted as a susceptible. The model with the exposed population has an average time in the exposed compartment equal to the delay but not the same delay for every individual. Real data is needed to determine which model is the best, since all the models, with no delay, with delay, and with an exposed population, have the same qualitative dynamics.

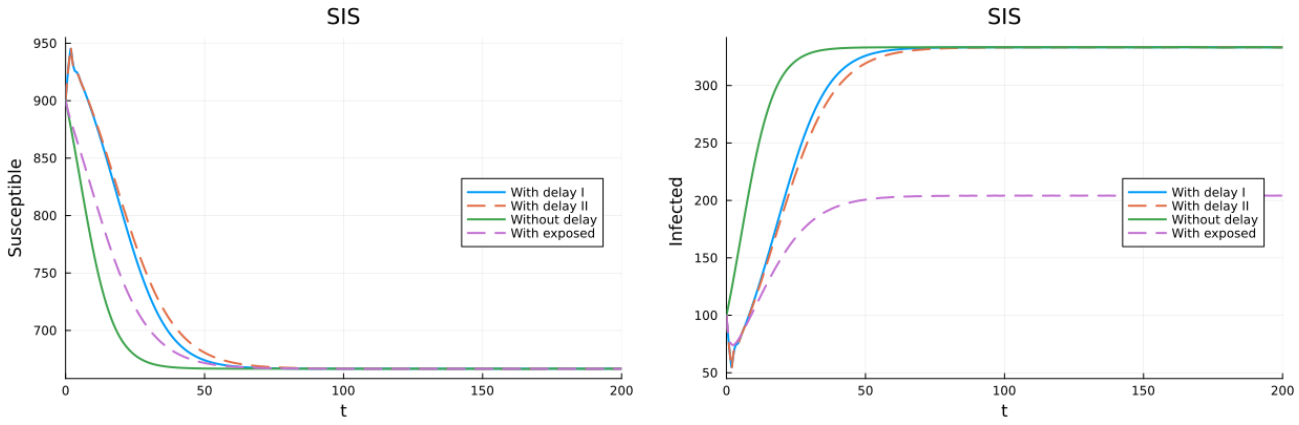


Figure 1. SIS model. (**left**) Susceptible population. (**right**) Infective population.

Figure 2 presents the plots for the SIR model, in the left for the susceptible population and for the infective in the right. Note that for $I = 0$, there are multiple solutions $S + R = N$. The simulated solutions depend on the initial values. The delay model I uses the history for both S and I, while all the other models only use $S(0)$.

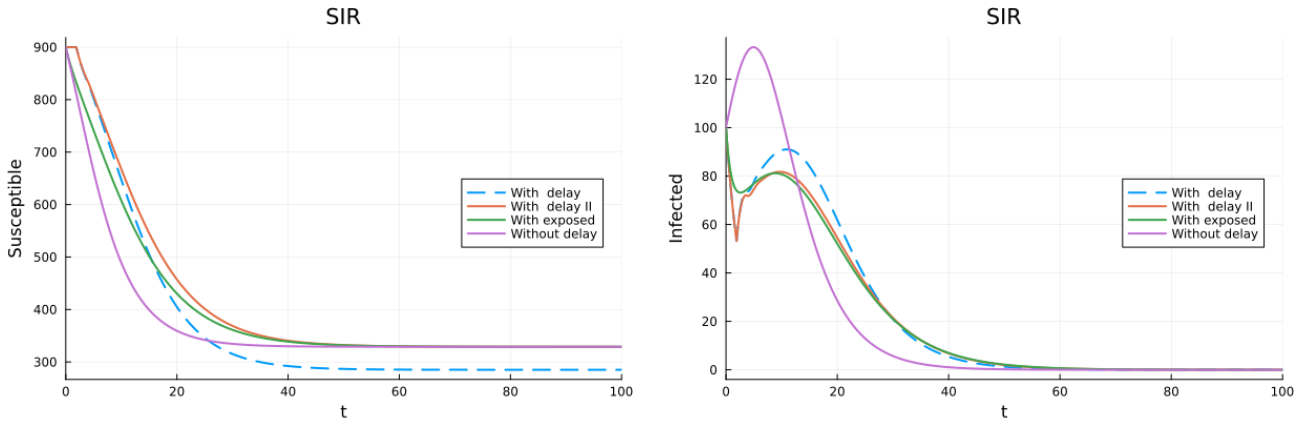


Figure 2. SIR model. (**left**) Susceptible population. (**right**) Infective population.

Figure 3 presents the plots for the susceptible (left) and infective (right) populations for the SIR model. This model has total but temporal immunity. As expected, the simulated solutions tend to the endemic equilibrium. The ODE model gives the highest infection peak, but all the models have the same qualitative dynamics.

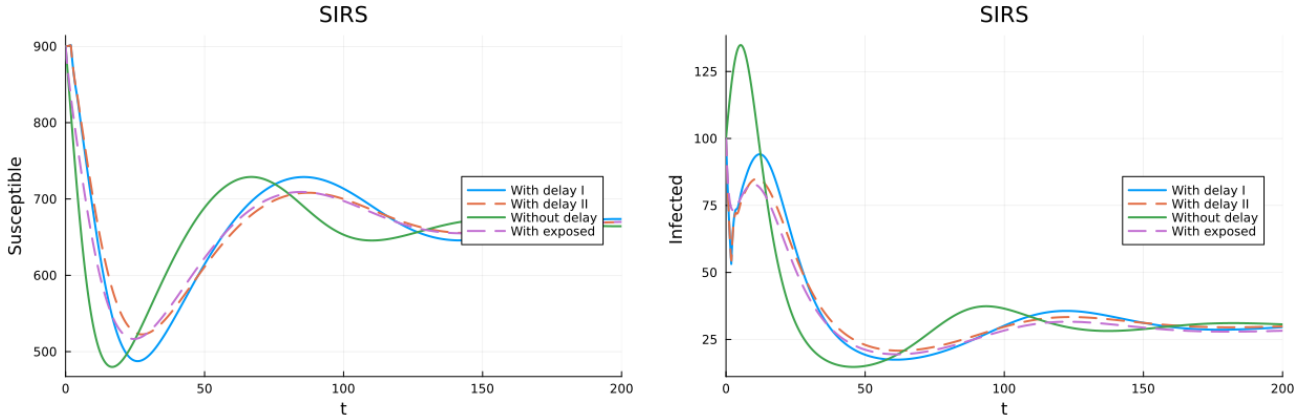


Figure 3. SIR model. (**left**) Susceptible population. (**right**) Infective population.

3.2. Models with Different Immunities

In this subsection, we consider models with different immunities: total, partial, permanent and temporal in different combinations. The first model (12) considers that all infectives after recovering get partial and permanent immunity. Figure 4 has the simulated solutions for the infective population for two different values of the infection rate for the population with partial immunity, β_p. On the left for $\beta_p = 0.25$ and the solution tends to $I = 0$ in spite of having $R_0 > 1$, and on the right for $\beta_p = 0.4$ and the solution tends to the endemic state. However, whether the epidemic ends or not depends on the values of $S_p\beta_p + S\beta - N\gamma$ which appears on the right-hand side of the differential equation for the infective population. The two delay models behave differently at the beginning since in the model with delay I, both S and I depend on the history, while for the model with delay II only I depends on the history.

The next model (16) considers that all infectives recover with partial and temporal immunity. Figure 5 has the plots for the infective population, on the left for $\beta_p = 0.25$, and I tends to a small but non zero value, and on the right for $\beta_p = 0.4$, and I tends to a much larger value, showing the effect of β.

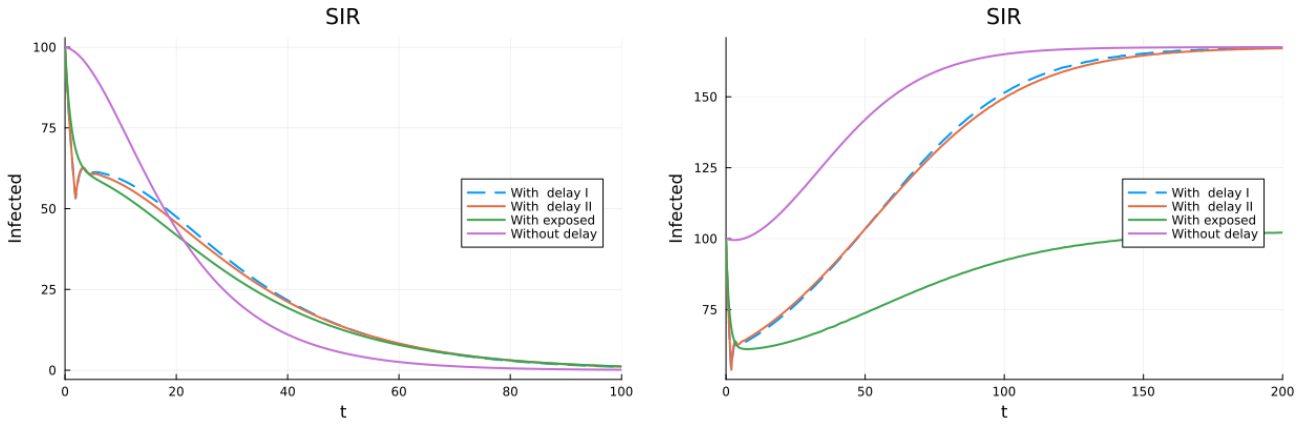


Figure 4. Model with partial and permanent immunity (**left**) $\beta_p = 0.25$. (**right**) $\beta_p = 0.4$.

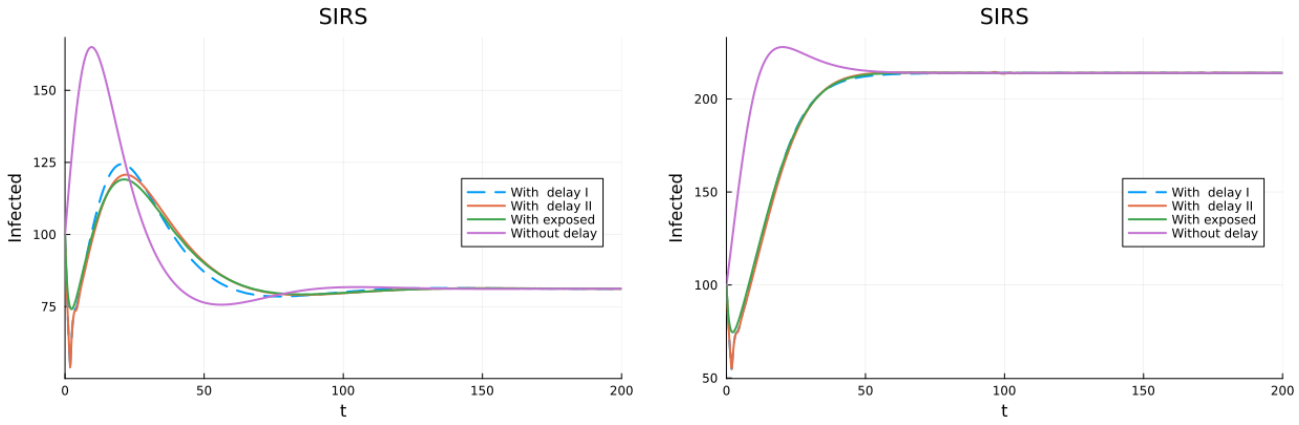


Figure 5. Model with partial and temporal immunity (**left**) $\beta_p = 0.25$. (**right**) $\beta_p = 0.4$.

Model (17) is for an infection with total but temporary immunity and then partial and permanent immunity. Figure 6 has the simulated plots for the infective population for two values of β_p. The figure on the left is for $\beta_p = 0.255$ and I tends to the DFE; the figure on the right is for $\beta_p = 0.4$ and I goes to an endemic equilibrium. For both simulations $R_0 > 1$ so initially the number of infectives grows. However, whether the epidemic ends or not depends on the values of $S_p\beta_p + S\beta - N\gamma$. All models have the same equilibria.

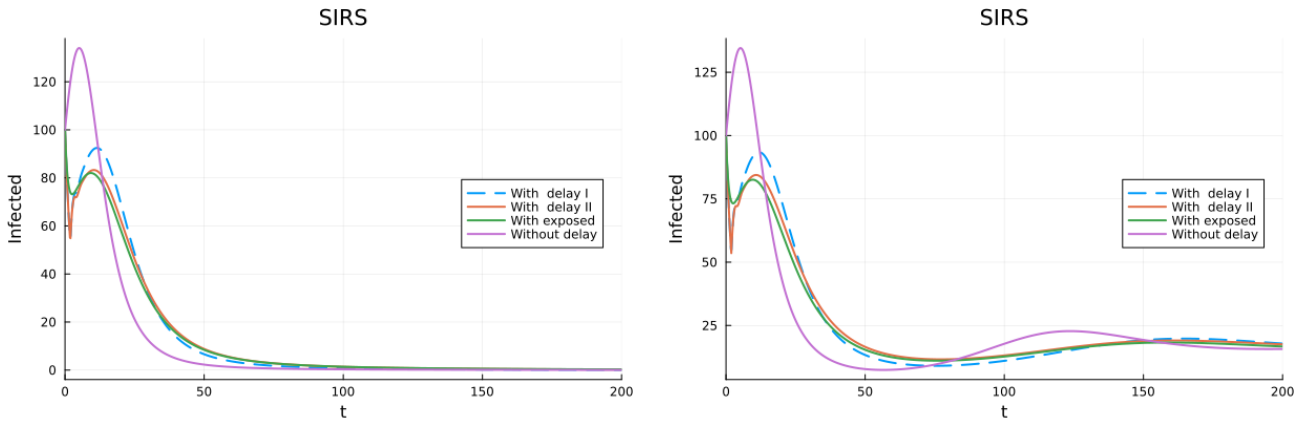


Figure 6. Model with total temporal and then partial permanent immunity (**left**) $\beta_p = 0.25$. (**right**) $\beta_p = 0.4$.

The next model (18) supposes that after being infected an individual gets total immunity which is temporal, and then partial immunity that is also temporal, finally losing all

immunity. Figure 7 shows the time solution of the infective population. On the left for $\beta_p = \beta/2$ and on the right for $\beta_p = 0.4$. In both cases the epidemic persists. The limit value for $\beta_p = 0.4$ is slightly larger. Since all individuals eventually lose immunity completely, the epidemic does not end.

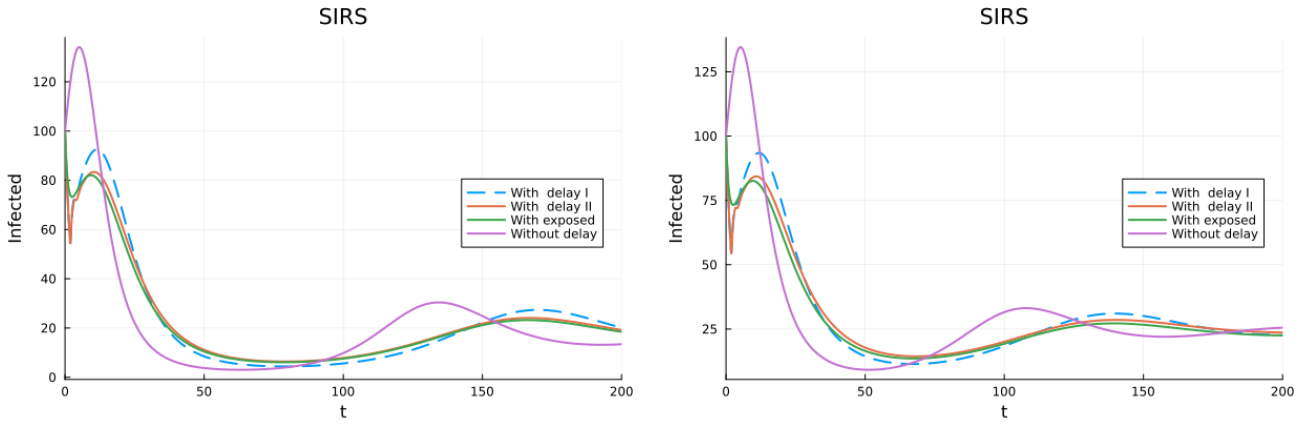


Figure 7. Model with total temporal and then partial temporal immunity (**left**) $\beta_p = 0.25$. (**right**) $\beta_p = 0.4$.

3.3. Models with Loss of Immunity and Two Groups

Different individuals have different immune system reactions to an infection. For example, all individuals may get total immunity to the infection after recovering but may lose it at different rates. For seasonal influenza some of the reasons are that children have been exposed to fewer variants of the virus, that older individuals have weaker immune systems, or that some individuals may have other risk factors such as smoking. For simplicity we consider just two classes of individuals, but the models can be extended to an arbitrary number.

The first model with two distinct populations (19) considers that the two groups get total and temporal immunity after infection (the usual SIR assumption), but lose the immunity at different rates ν and ν_1, with fractions f_r and $(1 - f_r)$, respectively. The simulations are in Figure 8. Both figures have $R_0 > 1$, the left one with $\beta = 0.5$ and the right one with $\beta = 0.666$. The limit value for the right figure is about twice that of the left showing the large influence that the infection rate has.

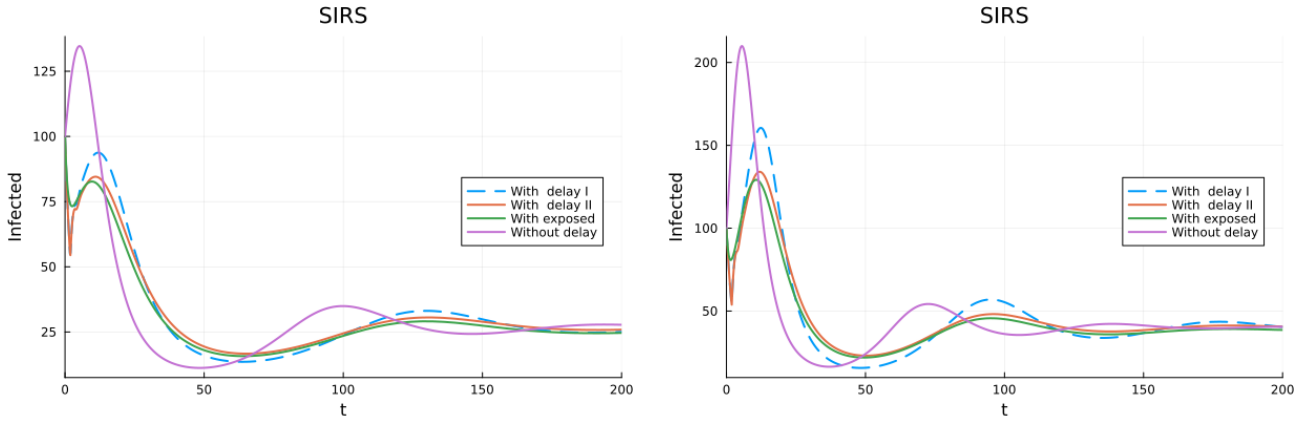


Figure 8. Model with two populations with total temporal immunity leaving the R compartment at different rates. (**left**) $\beta = 0.5$. (**right**) $\beta = 0.666$

The next model (23) assumes that one class gets total and permanent immunity, while the second gets partial and also permanent immunity. The fraction of individuals in both

classes is taken to be $f_p = 0.5$. For the class with partial immunity $\beta_p = \beta/2$. Since all individuals get permanent immunity it is expected that the epidemic will not persist. Figure 9 has two simulations for different values of β. The left one for $\beta = 0.5$ and the right one for $\beta = 1.5$. Even for the fairly large second value of β, that corresponds to $R_0 = 4.5$, the epidemic dies out.

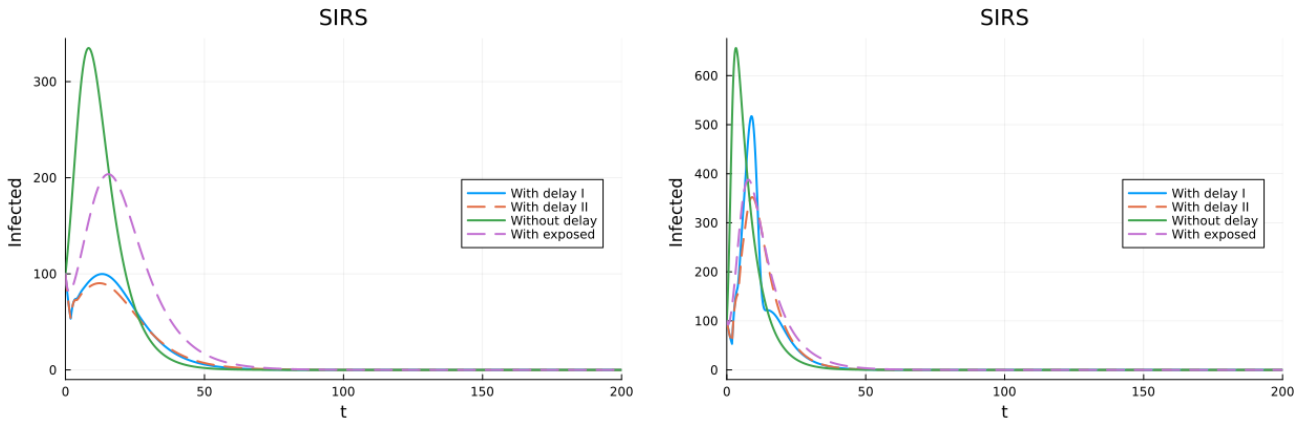


Figure 9. Simulations for the model with two classes, one with total and permanent immunity, and the other with partial and permanent immunity. (**left**) $\beta = 0.5$. (**right**) $\beta = 4.5$

Next we consider that half of the infective populations get total and temporal immunity, and half partial and also temporal immunity (24). Simulations were conducted for two values of β_p, the infection rate of the population with partial immunity, $\beta_p = \beta/2$ (left plot of Figure 10) and $\beta_p = \beta/8$, (right plot). In both cases the epidemic persists and the limit value of I has a relatively small change. The fact that all the individuals get immunity only for a finite time is the reason that the infection persists.

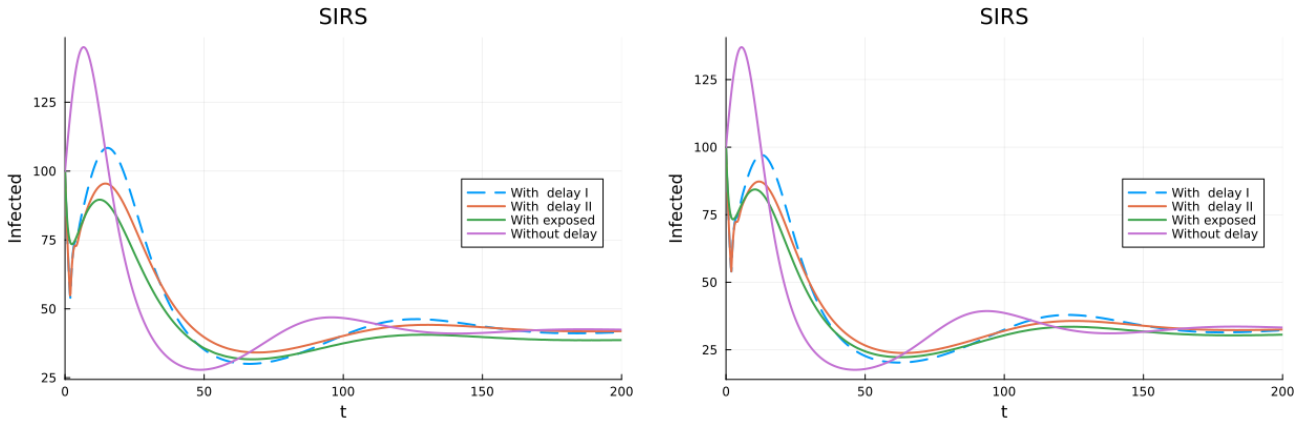


Figure 10. Simulations for the model with two classes, one with total and immunity, and the other with partial immunity, both temporal. (**left**) $\beta_p = \beta/2$. (**right**) $\beta_p = \beta/8$.

Model (25) assumes that all infectives recover with total temporal immunity and then get partial and permanent immunity. A fraction $f_r = 0.75$ has partial immunity with infection rate $\beta_1 = 0.9\beta$ and loss of permanent immunity rate $\nu = 1/30$ and the rest has infection rate $\beta_2 = 0.8\beta$ and immunity loss rate $\nu_1 = 1/60$ (left plot of Figure 11). The solution tends to the endemic equilibrium. A second simulation was conducted changing the infection rates to $\beta_1 = 0.45\beta$ and $\beta_2 = 0.4\beta$ (right half of Figure 11). Now the solution tends to the DFE. In the second simulation $\beta_1/\gamma < 1$ and $\beta_2/\gamma < 1$ so when most recovered individuals have partial permanent immunity, the differential equation for I shows that I decreases.

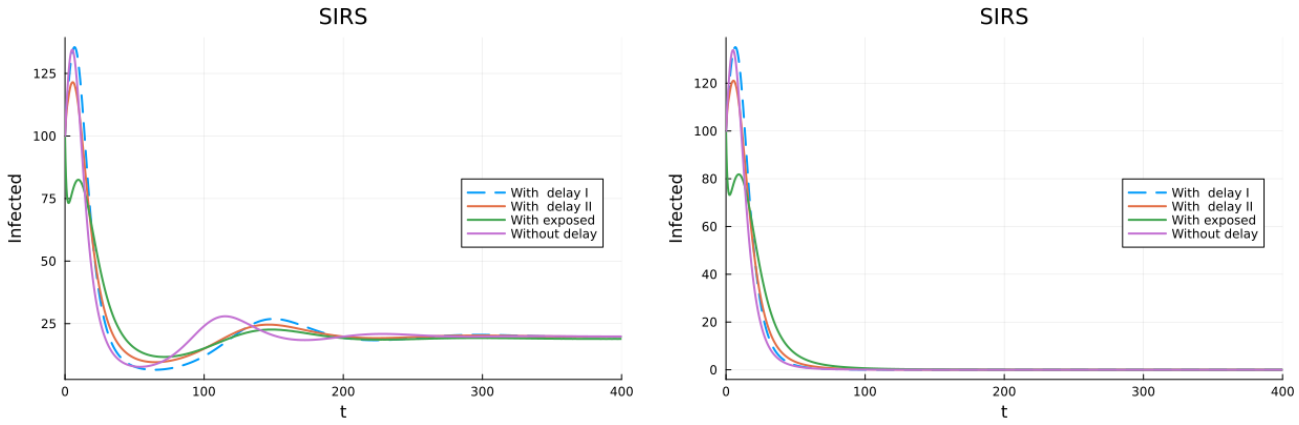


Figure 11. Model with two population classes. Both get total immunity (compartment R) and then lose it to get partial immunity at different rates with different infection rates. (**left**) $\beta_1 = 0.9\beta, \beta_2 = 0.8\beta$. (**right**) $\beta_1 = 0.45\beta, \beta_2 = 0.4\beta$.

3.4. Models with Vaccination and Immunity Loss

Vaccines can help control seasonal influenza. Since the influenza virus mutates from season to season, it is hard to predict what is the efficiency of the vaccine. It is also difficult to determine if the vaccine gives total or partial immunity, or if the effect wanes. A vaccinated person may not get influenza due to the vaccine or just because the person was not in close contact with an infective. Similarly, a person becoming infected only once may be due to having acquired total and permanent immunity, or not having close contacts with infectives. In this subsection, we present simulations conducted for different scenarios all with vaccination.

The first scenario represents individuals that get vaccinated at a rate rS, and achieve total immunity. The immunity from becoming infected with the disease is total and disappears at a rate ν, (26). In the left part of Figure 12, the effect of the vaccine wanes at a rate $\omega = 1/60$. In the right part, the effect of the vaccines does not wane, $\omega = 0$. In both cases the solution tends to the DFE. In the first case there is also a second endemic equilibrium.

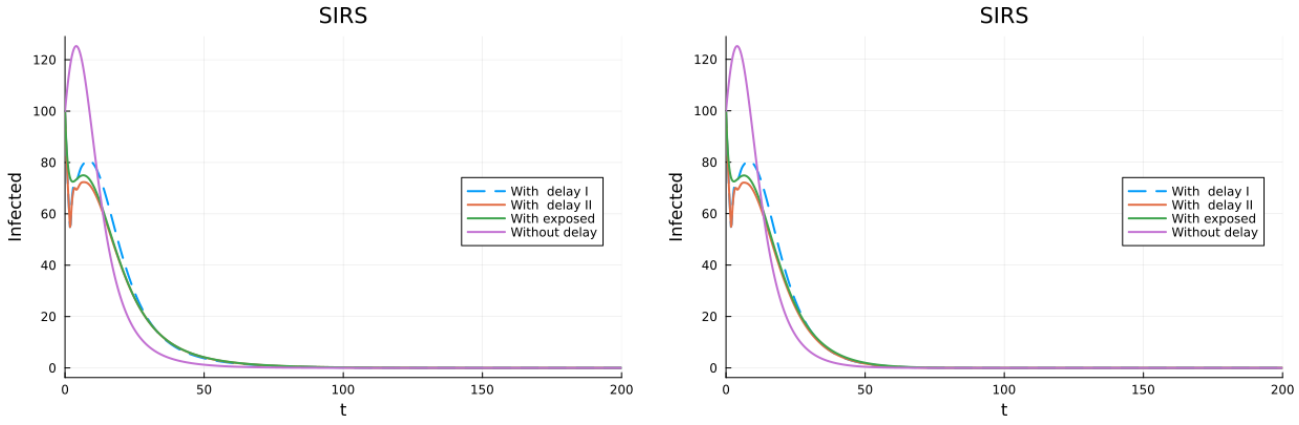


Figure 12. Model with vaccination that gives total immunity. (**left**) Immunity does not wane. (**right**) Immunity wanes.

The model in the next scenario is similar to the model given by (26), except that when the vaccine wanes the individual still gets partial permanent immunity with $\beta_v = 0.8\beta$. Figure 13 on the left shows the case with $\omega = 1/60$ and the simulated solution almost gets to the DFE but then it rebounds. On the right, the plot corresponds to $\omega = 0$ and

the solution trends to the DFE. The figure shows the importance that the waning of the vaccine has.

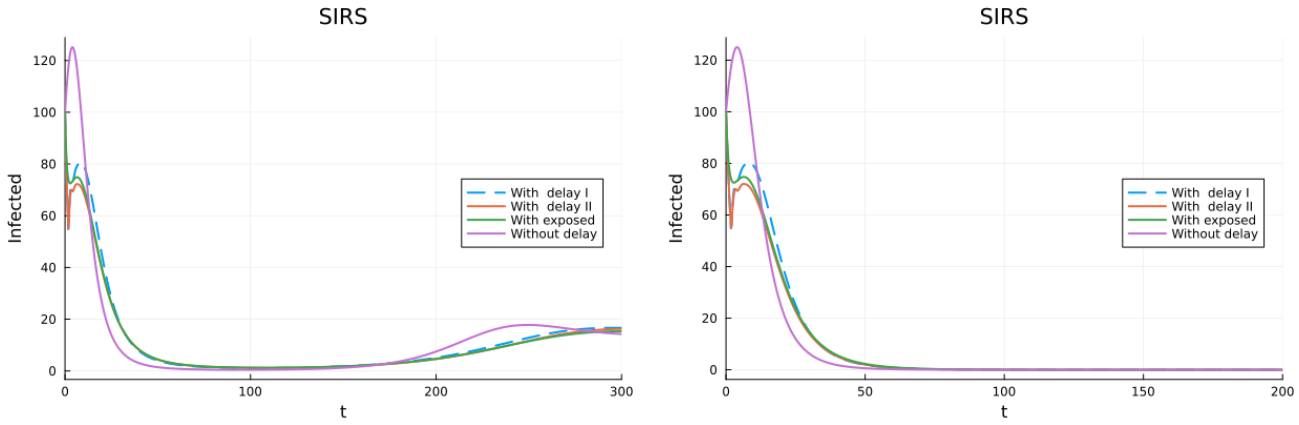


Figure 13. Model with vaccination. Infected individuals get total immunity that wanes to partial and permanent immunity. Vaccinated individuals get total immunity that may wane or not. (**left**) Vaccination immunity wanes to partial permanent. (**right**) Vaccination immunity does not wane.

The next simulations were performed with model (28). The vaccination rate is $r(p_r S - V)$. For this vaccination rate there is no DFE. The infected get to recovered that then get partial permanent immunity. Vaccinated get total immunity that, in one simulation wanes to partial permanent, and in the second one does not wane. We use $p_r = 0.4$, and the infection rate for the population with partial immunity $\beta_v = 0.8\beta$. The left plot of Figure 14 is for the simulation with vaccination waning, and the one of the right for no waning. In both plots the infective solution approaches close to zero and then rebounds.

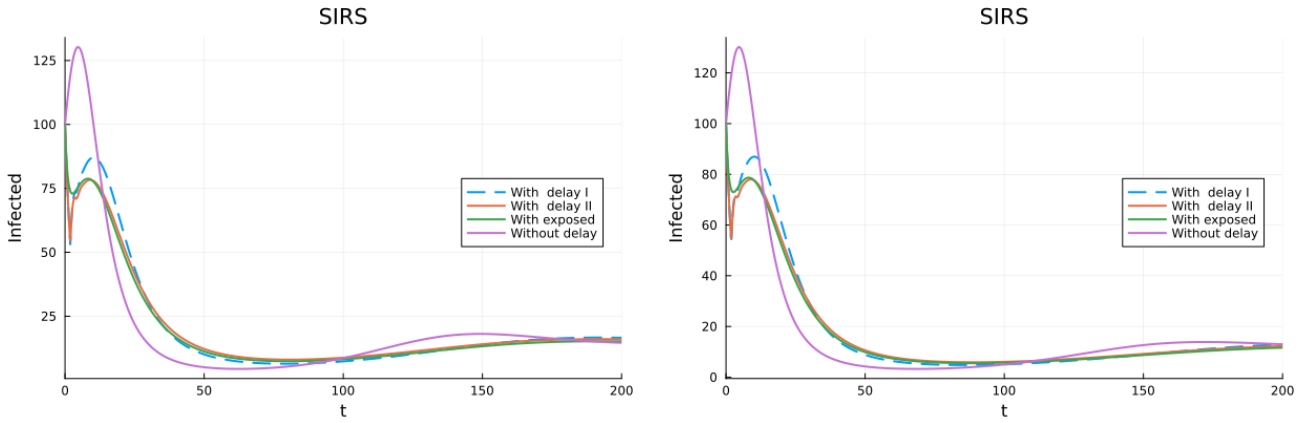


Figure 14. Model with recovered getting partial permanent immunity and vaccinated getting total immunity that in one case wanes and in the other does not. (**left**) Vaccination immunity wanes. (**right**) Vaccination immunity does not wane.

The next simulations are performed with the model given by (29). The vaccination rate is $r(p_r S - V)$ with $p_r = 0.6$ and a fraction $f_v = 0.75$ of the vaccinated population gets total and permanent immunity while a fraction $1 - f_r$ gets total immunity that wanes at a rate ω to partial and permanent immunity with infection rate $\beta_v = 0.8\beta$. The immunity from an infection is total and temporal. The calculated solution tends to an endemic equilibrium. A second simulation was performed with no waning of the effect of the vaccine $\omega = 0$. The solution tends to the DFE. The plots are in Figure 15. This example shows the importance of the waning or not of the vaccine.

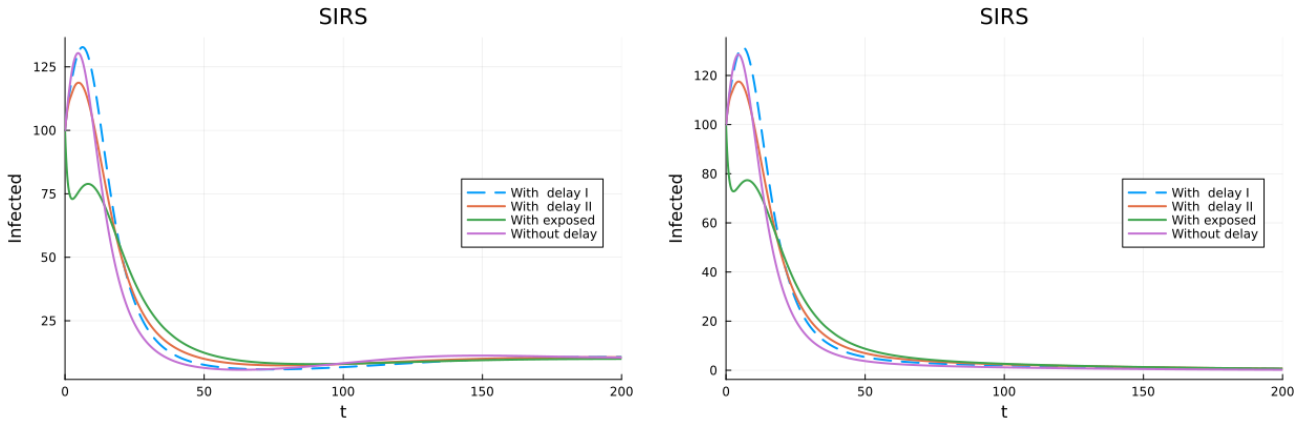


Figure 15. Model with two classes of vaccinated individuals, one with total and permanent immunity and the other with total immunity that wanes to partial and permanent immunity. (**left**) Waning vaccination immunity. (**right**) Permanent vaccination immunity.

The final model considers vaccination, the use of antivirals and isolation. The vaccination rate is $r(p_r S - V)$, a fraction $f_v = 0.75$ of the vaccinated keeps the total immunity forever and the rest loses the total immunity and then gets permanent partial immunity with infection rate $\beta_v = 0.8\beta$. The infected recover with total and temporal immunity. A fraction $f_a = 0.2$ is given antivirals and is removed from the infective class at a rate $\gamma_a = 2\gamma$. A fraction $f_q = 0.4$ is removed from the infectives and isolated at a rate $\gamma_q = 1$. Two simulations were performed. The first one with $\beta = 0.5$, that corresponds to $R_0 = 1.5$, is shown in the plot in the left of Figure 16 and has a solution that tends to the DFE. The second one, with $\beta = 1$ that gives $R_0 = 3$, is shown in the right plot and the solution tends to an endemic equilibrium. The figure shows that while R_0 predicts whether an epidemic starts or not, it does not indicate whether it will eventually die out or not.

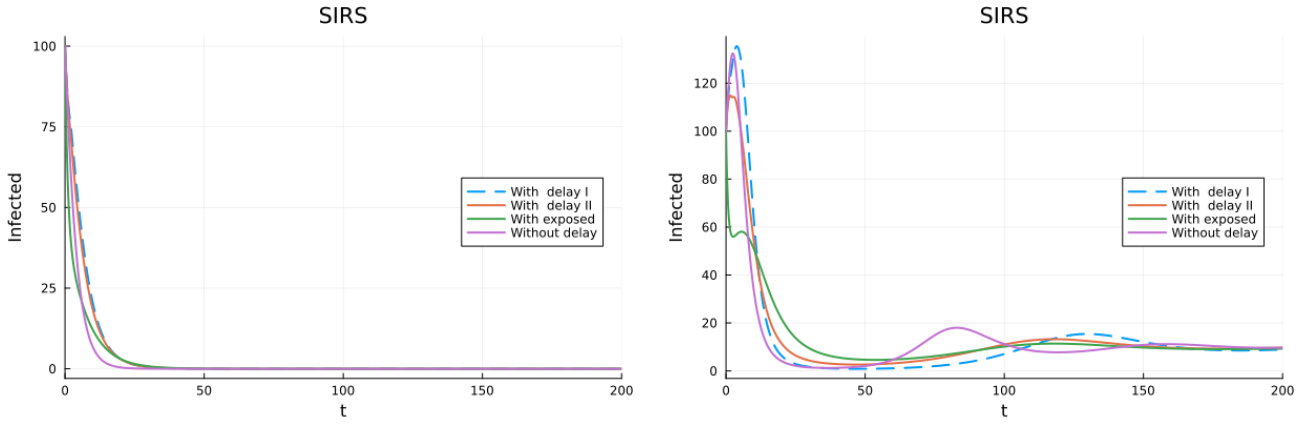


Figure 16. Model with vaccination, antivirals and isolation (30). (**left**) $\beta = 0.5$. (**right**) $\beta = 1$.

4. Discussion

SIR-based models have been used for may infectious diseases transmitted by direct contact. In this paper, we used models with no demographics for short-lived infections such as seasonal influenza. The first reason is that the infection has a short duration and that it is not very deadly. Therefore, the number of births and deaths during the infection period is small. The second reason is that it allows us to concentrate on the effects of losing immunity. The basic reproduction number R_0 is a very important index that gives a criterion for determining whether there will be an epidemic or not. It is usually calculated as we did in this paper by assuming that at the start of the infection there is one infective and $N - 1$ susceptibles with N the total population. For the majority of our examples it

is $R_0 = \beta S(0)/(\gamma N)$ and simplified to $R_0 = \beta/\gamma$. If there is also an initial time of the infection number of vaccinated individuals $V(0) > 0$, then $R_0 = \beta(N - V(0))/(\gamma N)$. And, if for example without vaccination $R_0 = 1.5$, by vaccinating more than $1/3$ of the total population then $R_0 < 1$ and there is not an epidemic. A similar estimate can be performed if initially there is also a number of individuals $S_p > 0$ with partial immunity and infection rate $\beta_p < \beta$.

The models presented start with the simple SIS, SIR and SIRS, that have no immunity, total and permanent immunity, and total and temporal immunity. They were first extended to include other situations with combinations of total and partial immunity, both permanent and temporal. For diseases like seasonal influenza it is very difficult to determine if an infected and recovered individual received total or partial immunity after the infection. Or if that immunity disappears or not. The individual may not get infected again due to total immunity or partial immunity or just because the individual was not in close contact with any infective. The different models permit running scenarios with different hypotheses and comparing the results. Two population groups with different immune reactions were considered next and different scenarios were run. Finally, vaccination was introduced with the different models having different assumptions and the final model being more complete having also antivirals and isolation to reduce the infection. For the various models that had DFEs simulations were performed where the epidemic died and where it did not. The effect of introducing a delay in the time of infection was also studied. The delay was introduced in two different ways using DDEs and also adding an exposed compartment. Apart from the delayed solutions taking more time to reach the equilibrium states, the behavior is very similar to that of the corresponding ODE model. That is, they produce solutions with the same equilibrium states and the same basic reproduction number. However, adding the delay makes the model more realistic and it is expected that the delayed models will fit real data better.

The parameters related to the time between contact and infection and to the duration of the infection are well-known for seasonal influenza. The infection rates can be estimated from the number of infectives. They vary from city to city and from population group to population group. As mentioned above, the parameters related to partial immunity and vaccination effects are hard to estimate and not reported. Numerical simulations using different models and comparing with actual epidemic data can help determine the more realistic hypotheses.

Since the two delay versions of each model have different characteristic equations, their solution behaviors can be different from each other and different from the behavior of the solutions of the non-delayed model. However, for the parameter values used in the simulations that did not happen and the qualitative dynamics were all similar. Determining whether an individual has total or partial immunity or has not come into close contact is not an easy task. We proposed different models with different types of immunity. Comparisons with real data are needed to establish which one is more realistic.

Future work will consider different initial conditions, the introduction of randomness and spatial dependence. The use of other compartments such as asymptomatics is another route.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

SIS	Susceptible, infective, susceptible
SIR	Susceptible, infective, recovered
SIRS	Susceptible, infective, recovered, susceptible
ODE	Ordinary differential equation
DDE	Delay differential equation
DFE	Disease-free equilibrium
R_0	Basic reproduction number

References

1. Centers for Disease Control. Influenza (Flu). 2024. Available online: https://www.cdc.gov/flu/about/keyfacts.htm#:~:text=Flu%20is%20a%20contagious%20respiratory,a%20flu%20vaccine%20each%20year (accessed on 24 April 2024).
2. Centers for Disease Control. Influenza. 2024. Available online: https://www.cdc.gov/nchs/fastats/flu.htm (accessed on 24 April 2024).
3. Ross, R.; Hudson, H.P. An application of the theory of probabilities to the study of a priori pathometry—Part III. *Proc. R. Soc. Lond. Ser. A Contain. Pap. Math. Phys. Character* **1917**, *93*, 225–240.
4. Kermack, W.O.; McKendrick, A.G. A contribution to the mathematical theory of epidemics. *Proc. R. Soc. Lond. Ser. A Contain. Pap. Math. Phys. Character* **1927**, *115*, 700–721.
5. Kuang, Y. *Delay Differential Equations: With Applications in Population Dynamics*; Academic Press: Cambridge, MA, USA, 1993.
6. Bellen, A.; Zennaro, M. *Numerical Methods for Delay Differential Equations*; Oxford University Press: Oxford, UK, 2013.
7. Ma, Z. *Dynamical Modeling and Analysis of Epidemics*; World Scientific: Singapore, 2009.
8. Rihan, F.A. *Delay Differential Equations and Applications to Biology*; Springer: Singapore, 2021.
9. Diekmann, O.; Heesterbeek, J.; Roberts, M.G. The construction of next-generation matrices for compartmental epidemic models. *J. R. Soc. Interface* **2010**, *7*, 873–885. [CrossRef] [PubMed]
10. Van den Driessche, P. Reproduction numbers of infectious disease models. *Infect. Dis. Model.* **2017**, *2*, 288–303. [CrossRef] [PubMed]
11. Edlund, S.; Kaufman, J.; Lessler, J.; Douglas, J.; Bromberg, M.; Kaufman, Z.; Bassal, R.; Chodick, G.; Marom, R.; Shalev, V.; et al. Comparing three basic models for seasonal influenza. *Epidemics* **2011**, *3*, 135–142. [CrossRef]
12. Van den Driessche, P.; Watmough, J. Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. *Math. Biosci.* **2002**, *180*, 29–48. [CrossRef] [PubMed]
13. Allen, L. *An Introduction to Mathematical Biology*; Pearson-Prentice Hall: Hoboken, NJ, USA, 2007.
14. Edelstein-Keshet, L. *Mathematical Models in Biology*; SIAM: Philadelphia, PA, USA, 2005.
15. Murray, J.D. *Mathematical Biology I. An Introduction*; Springer: Berlin/Heidelberg, Germany, 2002.
16. Fages, F.; Soliman, S. On robustness computation and optimization in BIOCHAM-4. In Proceedings of the International Conference on Computational Methods in Systems Biolog, Brno, Czech Republic, 12–14 September 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 292–299.
17. Cooke, K.L.; Yorke, J.A. Some equations modelling growth processes and gonorrhea epidemics. *Math. Biosci.* **1973**, *16*, 75–101. [CrossRef]
18. Jackson, M.; Chen-Charpentier, B.M. Modeling plant virus propagation with delays. *J. Comput. Appl. Math.* **2017**, *309*, 611–621. [CrossRef]
19. Khan, Q.J.A.; Krishnan, E.V. An Epidemic Model with a Time Delay in Transmission. *Appl. Math.* **2003**, *48*, 193–203. [CrossRef]
20. Al Basir, F.; Takeuchi, Y.; Ray, S.A. Dynamics of a delayed plant disease model with Beddington-DeAngelis disease transmission. *Math. Biosci. Eng.* **2021**, *18*, 583–599. [CrossRef]
21. Huang, G.; Ma, W.; Takeuchi, Y. Global analysis for delay virus dynamics model with Beddington–DeAngelis functional response. *Appl. Math. Lett.* **2011**, *24*, 1199–1203. [CrossRef]
22. Wei, H.M.; Li, X.Z.; Martcheva, M. An epidemic model of a vector-borne disease with direct transmission and time delay. *J. Math. Anal. Appl.* **2008**, *342*, 895–908. [CrossRef]
23. Chan, M.; Johansson, M.A. The incubation periods of dengue viruses. *PLoS ONE* **2012**, *7*, e50972. [CrossRef] [PubMed]
24. Jeger, M.; Madden, L.; Van Den Bosch, F. Plant virus epidemiology: Applications and prospects for mathematical modeling and analysis to improve understanding and disease control. *Plant Dis.* **2018**, *102*, 837–854. [CrossRef]
25. Liu, L. A delayed SIR model with general nonlinear incidence rate. *Adv. Differ. Equ.* **2015**, *2015*, 329. [CrossRef]
26. Hethcote, H.W. The Mathematics of Infectious Diseases. *SIAM Rev.* **2000**, *42*, 599–653. [CrossRef]
27. Wang, N.; Fu, Y.; Zhang, H.; Shi, H. An evaluation of mathematical models for the outbreak of COVID-19. *Precis. Clin. Med.* **2020**, *3*, 85–93. [CrossRef] [PubMed]
28. Bjørnstad, O.N.; Shea, K.; Krzywinski, M.; Altman, N. The SEIRS model for infectious disease dynamics. *Nat. Methods* **2020**, *17*, 557–559. [CrossRef]

29. Krammer, F. The human antibody response to influenza A virus infection and vaccination. *Nat. Rev. Immunol.* **2019**, *19*, 383–397. [CrossRef]
30. Patel, M.M.; York, I.A.; Monto, A.S.; Thompson, M.G.; Fry, A.M. Immune-mediated attenuation of influenza illness after infection: Opportunities and challenges. *Lancet Microbe* **2021**, *2*, e715–e725. [CrossRef]
31. Han, A.X.; de Jong, S.P.; Russell, C.A. Co-evolution of immunity and seasonal influenza viruses. *Nat. Rev. Microbiol.* **2023**, *21*, 805–817. [CrossRef] [PubMed]
32. Rodriguez-Maroto, G.; Atienza-Diez, I.; Ares, S.; Manrubia, S. Vaccination strategies in structured populations under partial immunity and reinfection. *J. Phys. A Math. Theor.* **2023**, *56*, 204003. [CrossRef]
33. Katriel, G. Epidemics with partial immunity to reinfection. *Math. Biosci.* **2010**, *228*, 153–159. [CrossRef] [PubMed]
34. Saha, P.; Mondal, B.; Ghosh, U. Dynamical behaviors of an epidemic model with partial immunity having nonlinear incidence and saturated treatment in deterministic and stochastic environments. *Chaos Solitons Fractals* **2023**, *174*, 113775. [CrossRef]
35. Thieme, H.R.; Yang, J. An endemic model with variable re-infection rate and applications to influenza. *Math. Biosci.* **2002**, *180*, 207–235. [CrossRef] [PubMed]
36. Carlsson, R.M.; Childs, L.M.; Feng, Z.; Glasser, J.W.; Heffernan, J.M.; Li, J.; Röst, G. Modeling the waning and boosting of immunity from infection or vaccination. *J. Theor. Biol.* **2020**, *497*, 110265. [CrossRef] [PubMed]
37. Bhattacharya, S.; Adler, F.R. A time since recovery model with varying rates of loss of immunity. *Bull. Math. Biol.* **2012**, *74*, 2810–2819. [CrossRef]
38. Viguerie, A.; Carletti, M.; Silvestri, G.; Veneziani, A. Mathematical Modeling of Periodic Outbreaks with Waning Immunity: A Possible Long-Term Description of COVID-19. *Mathematics* **2023**, *11*, 4918. [CrossRef]
39. Alexander, M.E.; Bowman, C.; Moghadas, S.M.; Summers, R.; Gumel, A.B.; Sahai, B.M. A vaccination model for transmission dynamics of influenza. *SIAM J. Appl. Dyn. Syst.* **2004**, *3*, 503–524. [CrossRef]
40. Sun, C.; Yang, W. Global results for an SIRS model with vaccination and isolation. *Nonlinear Anal. Real World Appl.* **2010**, *11*, 4223–4237. [CrossRef]
41. Sahu, G.P.; Dhar, J. Analysis of an SVEIS epidemic model with partial temporary immunity and saturation incidence rate. *Appl. Math. Model.* **2012**, *36*, 908–923. [CrossRef]
42. Moghadas, S.; Gumel, A. A mathematical study of a model for childhood diseases with non-permanent immunity. *J. Comput. Appl. Math.* **2003**, *157*, 347–363. [CrossRef]
43. Saad-Roy, C.M.; Morris, S.E.; Metcalf, C.J.E.; Mina, M.J.; Baker, R.E.; Farrar, J.; Holmes, E.C.; Pybus, O.G.; Graham, A.L.; Levin, S.A.; et al. Epidemiological and evolutionary considerations of SARS-CoV-2 vaccine dosing regimes. *Science* **2021**, *372*, 363–370. [CrossRef]
44. Goldberg, Y.; Mandel, M.; Bar-On, Y.M.; Bodenheimer, O.; Freedman, L.; Haas, E.J.; Milo, R.; Alroy-Preis, S.; Ash, N.; Huppert, A. Waning immunity after the BNT162b2 vaccine in Israel. *N. Engl. J. Med.* **2021**, *385*, e85. [CrossRef] [PubMed]
45. Turkyilmazoglu, M. An extended epidemic model with vaccination: Weak-immune SIRVI. *Phys. A Stat. Mech. Its Appl.* **2022**, *598*, 127429. [CrossRef] [PubMed]
46. Ho, S.H.; He, D.; Eftimie, R. Mathematical models of transmission dynamics and vaccine strategies in Hong Kong during the 2017–2018 winter influenza season. *J. Theor. Biol.* **2019**, *476*, 74–94. [CrossRef] [PubMed]
47. Reluga, T.C.; Medlock, J. Resistance mechanisms matter in SIR models. *Math. Biosci. Eng.* **2007**, *4*, 553–563. [PubMed]
48. Scherer, A.; McLean, A. Mathematical models of vaccination. *Br. Med. Bull.* **2002**, *62*, 187–199. [CrossRef] [PubMed]
49. Safan, M.; Kretzschmar, M.; Hadeler, K.P. Vaccination based control of infections in SIRS models with reinfection: Special reference to pertussis. *J. Math. Biol.* **2013**, *67*, 1083–1110. [CrossRef]
50. Krithika, B.; Tamilalagan, P. Unveiling the spread of epidemics involving partial immunity and reinfection: Insights from a time-delayed mathematical model. *Eur. Phys. J. Spec. Top.* **2023**, *232*, 2657–2673. [CrossRef]
51. Arif, F.; Majeed, Z.; Rahman, J.U.; Iqbal, N.; Kafle, J. Mathematical modeling and numerical simulation for the outbreak of COVID-19 involving loss of immunity and quarantined class. *Comput. Math. Methods Med.* **2022**, *2022*, 3816492. [CrossRef] [PubMed]
52. Rackauckas, C.; Nie, Q. DifferentialEquations.jl—A Performant and Feature-Rich Ecosystem for Solving Differential Equations in Julia. *J. Open Res. Softw.* **2017**, *5*. Available online: https://app.dimensions.aion2019/05/05 (accessed on 24 April 2024). [CrossRef]
53. Widmann, D.; Rackauckas, C. DelayDiffEq: Generating Delay Differential Equation Solvers via Recursive Embedding of Ordinary Differential Equation Solvers. *arXiv* **2022**, arXiv:2208.12879.
54. Colizza, V.; Barrat, A.; Barthelemy, M.; Valleron, A.J.; Vespignani, A. Modeling the worldwide spread of pandemic influenza: Baseline case and containment interventions. *PLoS Med.* **2007**, *4*, e13. [CrossRef]
55. Kanyiri, C.W.; Mark, K.; Luboobi, L. Mathematical analysis of influenza A dynamics in the emergence of drug resistance. *Comput. Math. Methods Med.* **2018**, *2018*, 2434560. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

AppliedMath

MDPI

Article

An Efficient Algorithm for Basic Elementary Matrix Functions with Specified Accuracy and Application

Huizeng Qin [1] and Youmin Lu [2,*]

[1] The School of Mathematics and Statistics, Shandong University of Technology, Zibo 255049, China; qinhz@sdut.edu.cn
[2] Department of Mathematical and Digital Sciences, Bloomsburg University, Bloomsburg, PA 17815, USA
* Correspondence: ylu@bloomu.edu

Abstract: If the matrix function $f(At)$ posses the properties of $f(At) = g(f((\frac{t}{k}A))$, then the recurrence formula $f_{i-1} = g(f_i), i = N, N-1, \cdots, 1, f(tA) = f_0$, can be established. Here, $f_N = f(A_N) = \sum_{j=0}^{m} a_j A_N^j, A_N = \frac{t}{k^N} A$. This provides an algorithm for computing the matrix function $f(At)$. By specifying the calculation accuracy p, a method is presented to determine m and N in a way that minimizes the time of the above algorithm, thus providing a fast algorithm for $f(At)$. It is important to note that m only depends on the calculation accuracy p and is independent of the matrix A and t. Therefore, $f(A_N)$ has a fixed calculation format that is easily computed. On the other hand, N depends not only on A, but also on t. This provides a means to select t such that N is equal to 0, a property of significance. In summary, the algorithm proposed in this article enables users to establish a desired level of accuracy and then utilize it to select the appropriate values for m and N to minimize computation time. This approach ensures that both accuracy and efficiency are addressed concurrently. We develop a general algorithm, then apply it to the exponential, trigonometric, and logarithmic matrix functions, and compare the performance with that of the internal system functions of Mathematica and Pade approximation. In the last section, an example is provided to illustrate the rapid computation of numerical solutions for linear differential equations.

Keywords: matrix function; efficient algorithms; high accuracy; linear differential equations

MSC: 68W25; 68U99

check for updates

Citation: Qin, H.; Lu, Y. An Efficient Algorithm for Basic Elementary Matrix Functions with Specified Accuracy and Application. *AppliedMath* **2024**, *4*, 690–708. https://doi.org/10.3390/appliedmath4020037

Received: 26 January 2024
Revised: 4 March 2024
Accepted: 19 April 2024
Published: 5 June 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Matrix exponent functions plays a crucial role in various fields, including ordinary differential equations and control theory. In particular, they are instrumental in analyzing general structural dynamic systems, represented by the initial value problem:

$$Mx'' + Gx' + Kx = r(t), x(0) = x_0, x'(0) = x_1, \tag{1}$$

where M, G, and K are n by n square matrices, and $r(t)$ denotes the external force applied. The introduction of dual variables of a Hamiltonian system,

$$p = Mx' + Gx/2, \tag{2}$$

results in the following system:

$$U' = HU + F(t), U(0) = U_0, \tag{3}$$

where

$$\begin{gathered} U=\begin{pmatrix} q \\ p \end{pmatrix}, H=\begin{pmatrix} -M^{-1}G/2 & M^{-1} \\ GM^{-1}G/4-K & -M^{-1}G/2 \end{pmatrix}, \\ F(t)=\begin{pmatrix} 0 \\ r(t) \end{pmatrix}, q=x, U_0=\begin{pmatrix} x_0 \\ Mx_1+Gx_0/2 \end{pmatrix}. \end{gathered} \tag{4}$$

The solution of (3) is provided as follows:

$$U(t)=e^{Ht}U_0+\int_0^t e^{H(t-s)}F(s)ds. \tag{5}$$

When a problem is oscillatory, the formal solution often involves both the sine and cosine of a matrix. Famous examples include the Schrődinger equation in quantum mechanics:

$$i\frac{d\psi}{dt}=H(t)\psi, \quad \psi(t_0)=\psi_0, \tag{6}$$

where $H(t)$ is a Hermitian operator and ψ is a complex wave function. If $H(t)=A$ is a real and constant matrix, the solution of (6) is as follows:

$$\psi(t)=e^{-iAt}\psi_0=(\cos(tA)-i\sin(tA))\psi_0. \tag{7}$$

There are other examples where the computation of the sine and cosine of a matrix can be of interest. Wave equations provided by the generic second order system are written as follows:

$$x''+Kx=F(t), x(0)=x_0, x'(0)=x_1, \tag{8}$$

where the solution is provided by the variation of the constant method:

$$x(t)=\cos(tA)x_0+\sin_1(tA)x_1+\int_0^t \sin_1((t-s)A)F(s)ds, \tag{9}$$

where $A=\sqrt{K}, \sin_1(tA)=\int_0^t \cos(sA)ds$. If A is invertible, then (9) can be written as follows:

$$x(t)=\cos(tA)x_0+\sin(tA)A^{-1}x_1+A^{-1}\int_0^t \sin((t-s)A)F(s)ds. \tag{10}$$

For (5), providing $U(k\tau)(k=1,2,\cdots)$ as accurately as possible is a concern for engineers and mathematicians, where τ is called the step. The specific method for this problem is as follows:

$$U_k=TU_{k-1}+\int_0^\tau e^{Hs}F(k\tau-s)ds, k=1,2,\cdots, T=e^{H\tau}. \tag{11}$$

Similarly, for (10), providing the algorithm $x(k\tau)(k=1,2,\cdots)$ as accurately as possible involves introducing vectors $x'=y$ and the following algorithm:

$$\begin{gathered} x_k=\cos(\tau A)x_{k-1}+\sin(\tau A)A^{-1}y_{k-1}+A^{-1}\int_0^\tau \sin(sA)F(k\tau-s)ds, \\ y_k=-A\sin(\tau A)x_{k-1}+\cos(\tau A)y_{k-1}+\int_0^\tau \cos(sA)F(k\tau-s)ds, \\ x_0=x(0), y_0=x'(0)=x_1. \end{gathered} \tag{12}$$

To improve the accuracy of $U(k\tau)$ and $x(k\tau)$, it is essential to develop an algorithm with high accuracy for computing $T=e^{H\tau}, \cos(\tau A)$ and $\sin(\tau A)$ and the integral containing these matrix functions. This paper improves the algorithms developed for the computation of e^A and $\cos A$ in the existing literature, see [1–15], and also discusses the computation of other trigonometric, hyperbolic, logarithmic, and related inverse triangular and inverse hyperbolic matrices.

2. The General Theorem of the Fast Algorithm for Matrix Functions

In [16], we addressed the challenge of rapidly calculating elementary functions with specified precision. The implementation method involves an analytic function $f(x)$ specifying property,

$$f(x) = g\left(f\left(\frac{x}{k}\right)\right), \text{ where } k > 1. \tag{13}$$

If C_0 is an approximation as close as possible to $f\left(\frac{x}{k^N}\right)$, with N as a positive integer, for example,

$$C_0 = \sum_{j=0}^{m} \frac{f^{(j)}(0)(\Delta x)^j}{j!}, \Delta x = \frac{x}{k^N}, m = 1, 2, \cdots \tag{14}$$

or C_0 being the Padé approximation of $f\left(\frac{x}{k^N}\right)$, then the following calculations are performed:

$$\text{for}(i = 0, i < N, i++)\ C_{i+1} = g(C_i). \tag{15}$$

Completing (14) and (15), $f(x) \approx C_N$ is obtained. This method is also applicable to the calculation of matrix functions, where the variable x is replaced by the n by n square matrix A. Now, we have the following:

$$C_0 = \sum_{j=0}^{m} \frac{f^{(j)}(0)(\Delta A)^j}{j!}, \Delta A = \frac{1}{k^N} A, m = 1, 2, \cdots \tag{16}$$

or C_0 is the Padé approximation of $f(\Delta A)$. Our objective is to determine m and N to minimize the computation time under the condition that the truncation error and computational precision are 10^{-p}. For the matrix function calculation, the coefficient calculation of (14) can be ignored, mainly focusing on the multiplication calculation of the matrix in (16) and (15). If the number of multiplications to complete $g(A)$ is K, then the total number of times to complete (16) and (15) is roughly

$$q(m, N) = (m + KN)M_n, \tag{17}$$

where M_n is the number of times to complete one multiplication of the n by n matrix. Based on the above discussion, we present a fast algorithm for the matrix function as follows.

Theorem 1. *Let m_p be an integer close to the minimum point of $h(m)$, where*

$$h(m) = m + \frac{pK \ln 10}{(m+1)\ln k} + \frac{K}{(m+1)\ln k} \ln\left|\frac{f^{(m+1)}(0)}{(m+1)!}\right| \tag{18}$$

and

$$N = \left[\frac{p \ln 10}{(m+1)\ln k} + \frac{1}{(m+1)\ln k} \ln\left|\frac{f^{(m+1)}(0)}{(m+1)!}\right| + \frac{\ln\|A\|}{\ln k}\right] + 1. \tag{19}$$

(1) $C_0 = \sum_{j=0}^{m_p} \frac{f^{(j)}(0)(\Delta A)^j}{j!}$, where $\Delta A = \frac{1}{k^N} A$, or C_0 is a Padé approximation of $f(\Delta A)$.
(2) for$(i = 0, i < N, i++)\ C_{i+1} = g(C_i)$ and $(A) \approx C_N$.

The above algorithm is a fast algorithm for the general matrix function.

Proof. With a truncation error of 10^{-p}, for (16) to be satisfied,

$$\left\|\frac{f^{(m+1)}(0)(\Delta A)^{m+1}}{(m+1)!}\right\| = \left|\frac{f^{(m+1)}(0)}{(m+1)!k^{(m+1)N}}\right| \|A^{m+1}\| \le \left|\frac{f^{(m+1)}(0)}{(m+1)!k^{(m+1)N}}\right| \|A\|^{m+1} < 10^{-p}, \tag{20}$$

where $\|A\|$ is a norm of A. So,

$$N > \frac{p \ln 10}{(m+1)\ln k} + \frac{1}{(m+1)\ln k} \ln\left|\frac{f^{(m+1)}(0)}{(m+1)!}\right| + \frac{\ln\|A\|}{\ln k}. \tag{21}$$

From (17), we obtain the following:

$$q(m,N) > g(m) = \left(m + \frac{pK\ln 10}{(m+1)\ln k} + \frac{K}{(m+1)\ln k} \ln \left| \frac{f^{(m+1)}(0)}{(m+1)!} \right| + \frac{K \ln \|A\|}{\ln k} \right) M_n. \tag{22}$$

Note that the minimum point of $g(m)$ has nothing to do with $\frac{K\ln\|A\|}{\ln k}$. By making $q(m,N)$ as small as possible, the algorithm obtained by (1) and (2) is a fast algorithm, thus completing the proof of the theorem. □

Remark 1. *In the specific calculation, $\|A\|$ can be replaced by the approximate eigenvalue with the maximum modulus in this paper.*

$$\left| \frac{f^{(m+1)}(0)}{(m+1)!k^{(m+1)N}} \right| \|A\|^{m+1} < 10^{-p}. \tag{23}$$

3. Fast Algorithm for e^A, $\cos A$, $\sin A$, $\sinh A$, and $\cosh A$

Suppose $\lambda = \|A\|$ is a norm for A. If we let $A = \lambda A_0$, then $\|A_0\| \leq 1$. From (19), we can obtain the following:

$$N_p = \left[\frac{p\ln 10}{(m+1)\ln k} + \frac{1}{(m+1)\ln k} \ln \left| \frac{f^{(m+1)}(0)}{(m+1)!} \right| + \frac{\ln \lambda}{K \ln k} \right] + 1. \tag{24}$$

In order to reduce the length of the article, the algorithms for several matrix functions are compiled into a table.

Theorem 2. *For $e^A, \cos A, \sin A, \cosh A$, and $\sinh A$, we devised the following fast algorithm using Theorem 1.*

Remark 2. *It is worth noting that m_p in Table 1 only depends on the calculation accuracy p and the matrix function, and is independent of matrix A.*

Table 1. The fast algorithm for $e^A, \cos A, \sin A, \cosh A$, and $\sinh A$.

	m_p, N_p	Algorithm
e^A	$m_p = \left[\sqrt{\frac{9p\ln 10}{8\ln 2}} \right], N_p = [N_e] + 1$	(1) $B = \frac{A}{2^{N_p}}, C = I + \sum_{j=1}^{m_p} \frac{B^j}{j!}$ (2) $for(i = 0, i < N, i++)\ C = C^2$ (3) $e^A = C$
$\left\{ \begin{array}{c} \sin A \\ \sinh A \end{array} \right\}$	$m_p = \left[\sqrt{\frac{p\ln 10}{2\ln 3}} \right], N_p = [N_{\sin}] + 1$	(1) $B = \frac{A^2}{3^{2N_p}}, C = \frac{A}{3^N} + \frac{A}{3^N} \sum_{j=1}^{m_p} \frac{(\mp B)^j}{(2j+1)!}$ (2) $for(i = 0, i < N, i++)\ C = 3C \mp 4C^3$ (3) $\left\{ \begin{array}{c} \sin A \\ \sinh A \end{array} \right\} = C$
$(I)\ \left\{ \begin{array}{c} \cos A \\ \cosh A \end{array} \right\}$	$m_p = \left[\sqrt{\frac{p\ln 10}{2\ln 3}} \right] + 1, N_p = [N_{\sin}]$	(1) $B = \frac{A^2}{3^{2N_p}}, C = I + \sum_{j=1}^{m_p} \frac{(\mp B)^j}{(2j)!}$ (2) $for(i = 0, i < N, i++)\ C = 4C^3 - 3C$ (3) $\left\{ \begin{array}{c} \cos A \\ \cosh A \end{array} \right\} = C$
$(II)\ \left\{ \begin{array}{c} \cos A \\ \cosh A \end{array} \right\}$	$m_p = \left[\sqrt{\frac{p\ln 10}{2\ln 3}} \right] + 1, N_p = [N_{\sin}]$	(1) $B = \frac{A^2}{3^{2N_p}}, C = I + \sum_{j=1}^{m_p} \frac{(\mp B)^j}{(2j)!}$ (2) $for(i = 0, i < N, i++)\ C = C^2 - I$ (3) $\left\{ \begin{array}{c} \cos A \\ \cosh A \end{array} \right\} = C$
$N_e = \frac{p\ln 10}{(m_p+1)\ln 2} + \frac{\ln\lambda}{\ln 2} - \frac{\ln(m_p+1)!}{(m_p+1)\ln 2}, N_{\cos} = \frac{p\ln 10}{(2m_p+1)\ln 2} + \frac{\ln\lambda}{\ln 2} - \frac{\ln(2m_p+1)!}{(2m_p+1)\ln 2},$ $N_{\sin} = \frac{p\ln 10}{(2m_p+2)\ln 3} + \frac{\ln\lambda}{\ln 3} - \frac{\ln(2m_p+2)!}{(2m_p+2)\ln 3}.$		

Now, we utilize the general fast algorithm to specifically demonstrate the correctness of the algorithms in Table 1. The algorithms in Table 1 are straightforward; they involve placing the specific functions f and g into the general algorithm, and the tricky part lies in deriving the formulas of m_p and N_p. We calculate the formulas for m_p and N_p for the function e^A for $12 \leq p \leq 36$ here. In fact, the graphing of the function of $h(m)$ supports our results for all values of p. The derivation of m_p and N_p for the other functions is very similar.

(1) For the function e^A, we have $g(u) = u^2, k = 2, K = 1$, and

$$h(m) = m + \frac{p \ln 10}{(m+1)\ln 2} - \frac{1}{(m+1)\ln 2} \ln(m+1)!. \tag{25}$$

Using Sterling's approximation formula, we obtain the following:

$$\ln((m+1)!) = \frac{1}{2}\ln(2\pi) + \frac{1}{2}\ln(m+1) + (m+1)\ln(m+1) - m - 1 + O(\frac{1}{m+1}).$$

Now, implementing it into (25), we obtain the following:

$$h(m) \approx m + \frac{p \ln 10}{(m+1)\ln 2} - \frac{\ln(2\pi)}{2(m+1)\ln 2} - \frac{\ln(m+1)}{2(m+1)\ln 2} + \frac{1}{\ln 2}.$$

Let

$$q(x) = x + \frac{p \ln 10}{(x+1)\ln 2} - \frac{\ln(2\pi)}{2(x+1)\ln 2} - \frac{\ln(x+1)}{2(x+1)\ln 2} + \frac{1}{\ln 2}.$$

Taking its derivative with respect to x, one obtains

$$q'(x) = 1 - \frac{p \ln 10}{(x+1)^2 \ln 2} + \frac{\ln(2\pi)}{2(x+1)^2 \ln 2} + \frac{\ln(x+1)}{2(x+1)^2 \ln 2} - \frac{1}{2(x+1)^2 \ln 2}.$$

To determine extreme values, we set it to 0 and multiply both sides by $(m+1)^2$:

$$(x+1)^2 - \frac{1}{\ln 2}(x+1) + \frac{\ln(x+1)}{2\ln 2} - \frac{2p\ln 10 - \ln(2\pi) + 1}{2\ln 2} = 0.$$

Using simple algebra, we obtain the following:

$$(x+1)^2 > \frac{(m+1) - \ln(m+1) + (m+1) - \ln(2\pi)}{2\ln 2} + \frac{2p\ln 10 + 1}{2\ln 2} > \frac{2p\ln 10 + 1}{2\ln 2}.$$

Using $p = 12$, we obtain a lower bound for x :

$$x > \sqrt{\frac{2p\ln 10 + 1}{2\ln 2}} - 1 = 6.37.$$

Thus, we have a lower bound for m: $m \geq 6$. We also can solve the quadratic equation to get

$$m + 1 = \frac{1}{\ln 2} + \sqrt{\frac{1}{4\ln^2 2} + \frac{8p\ln 10 - 4\ln(2\pi) + 4}{8\ln 2}}.$$

Now, using the smallest value for m obtained and $p = 36$, we can obtain our first upper bound:

$$m < -1 + \frac{1}{\ln 2} + \sqrt{\frac{1}{4\ln^2 2} - \frac{\ln 7}{2\ln 2} + \frac{276\ln 10 - 4\ln(2\pi) + 4}{8\ln 2}} \approx 11.31.$$

We can again use the values of m and $p = 12$ to update the lower bound of m as follows:

$$m > -1 + \frac{1}{\ln 2} + \sqrt{\frac{1}{4\ln^2 2} + -\frac{\ln 12}{2\ln 2} + \frac{96\ln 10 - 4\ln(2\pi) + 4}{8\ln 2}} \approx 6.61.$$

Now, our new lower bound for m is 7. With this small range of values for m, by inspecting the formula for the values of m, we can see that the value of m is very close to $\frac{p\ln 10}{\ln 2}$, and we can safely claim that m$= \left[\sqrt{\frac{9p\ln 10}{8\ln 2}}\right]$. Our following numerical calculation supports this claim completely.

Case (a): $p = 12$. Using the formula obtained by the quadratic formula and the upper bound $m = 11$, we obtain a new lower bound: $m \geq 6.63$. Using the formula obtained by the quadratic formula and lower bound $m = 7$, we can also obtain the upper bound: $m \leq 6.63$. Using our claim $\frac{9p\ln 10}{8\ln 2}$, we obtain $m = 6.70$. This clearly matches the lower and upper bound quite well.

Case (b): $p = 18$. Using the formula obtained by the quadratic formula and the upper bound $m = 11$, we obtain a new lower bound: $m \geq 8.05$. Using the formula obtained by the quadratic formula and lower bound $m = 7$, we can also obtain the upper bound: $m \leq 8.07$. Using our claim $\frac{9p\ln 10}{8\ln 2}$, we obtain $m = 8.20$. These values match very well for taking the closest integer.

Case (c): $p = 24$. Using the formula obtained by the quadratic formula and the upper bound $m = 11$, we obtain a new lower bound: $m \geq 9.27$. Using the formula obtained by the quadratic formula and lower bound $m = 7$, we can also obtain the upper bound: $m \leq 9.28$. Using our claim $\frac{9p\ln 10}{8\ln 2}$, we obtain $m = 9.47$.

Case (d): $p = 36$. Using the formula obtained by the quadratic formula and the upper bound $m = 11$, we obtain a new lower bound: $m \geq 11.29$. Using the formula obtained by the quadratic formula and lower bound $m = 7$, we can also obtain the upper bound: $m \leq 11.37$. Using our claim $\frac{9p\ln 10}{8\ln 2}$, we obtain $m = 11.60$.

Using $m_p = \sqrt{\frac{9p\ln 10}{8\ln 2}}$, we can obtain

$$N_p = \left[\frac{p\ln 10}{(m_p+1)\ln 2} - \frac{\ln(m_p+1)!}{(m_p+1)\ln 2} + \frac{\ln\lambda}{\ln 2}\right] + 1. \tag{26}$$

(2) For $\left\{\begin{array}{c}\sin A\\ \sinh A\end{array}\right\}$, we have $g(u) = 3u \mp 4u^3, k = 3, K = 2$ and

$$h(m) = m + \frac{2p\ln 10}{(2m+2)\ln 3} - \frac{2\ln(2m+2)!}{(2m+2)\ln k}. \tag{27}$$

In this case, for $p = 12, 16, 24$, and 32, the corresponding integer points for $h(m)$ to achieve its minimum value are $3, 3, 5$, and 6. Again, it can be estimated that

$$m_p = \sqrt{\frac{3p\ln 10}{5\ln 3}}, N_p = \left[\frac{p\ln 10}{(2m_p+2)\ln 3} - \frac{\ln(2m_p+2)!}{(2m_p+2)\ln 3} + \frac{\ln\lambda}{\ln 3}\right] + 1. \tag{28}$$

(3) For $\left\{\begin{array}{c}\cos A\\ \cosh A\end{array}\right\}$, there are two cases:

Case (I): $g(u) = 4u^3 - 3u, k = 3, K = 2$, and (29) hold. Compared to I, the values in the second part of $C = I + \sum_{j=1}^{m_p} \frac{(-B)^j}{(2j)!}$ are very small, so we may replace N with $N - 1$, and m_p with $m_p + 1$.

Case (II): $g(u) = u^2 \mp 1, k = 2, K = 1$ and

$$h(m) = m + \frac{p\ln 10}{(2m+1)\ln 2} - \frac{1}{(2m+1)\ln 2}\ln(2m+1)!. \tag{29}$$

Similarly, for $p = 12, 16, 24$, and 32, the corresponding integer points for $h(m)$ to achieve its minimum value are $4, 5, 6$, and 7. Consequently, we can determine the following:

$$m_p = \sqrt{\frac{3p\ln 10}{5\ln 2}}, N_p = \left[\frac{p\ln 10}{(2m_p+1)\ln 2} - \frac{\ln(2m_p+1)!}{(2m_p+1)\ln 2} + \frac{\ln\lambda}{\ln 2}\right] + 1. \tag{30}$$

The effectiveness of the algorithms is displayed in Table 1 by implementing numerical computations in Mathematica. N_p and m_p are replaced with $N_{p,\mu} = \mu N_p, m_{p,\mu}$ in Table 1, defined correspondingly by

$$\begin{aligned}
m_{p,\mu} &= \min_{\|B^j/j!\|<10^{-p}} j, B = \left\|\frac{A}{2^{N_{p,\mu}}}\right\| \text{ for } e^A,\\
m_{p,\mu} &= \min_{\|B^{j+1/2}/(2j+1)!)\|<10^{-p}} j, B = \left\|\frac{A^2}{3^{2N_{p,\mu}}}\right\| \text{ for } \sin A,\\
\text{(I) } m_{p,\mu} &= \min_{\|B^j/(2j)!\|<10^{-p}} j, B = \left\|\frac{A^2}{3^{2N_{p,\mu}}}\right\| \text{ for } \cos A,\\
\text{(II) } m_{p,\mu} &= \min_{\|B^j/(2j)!\|<10^{-p}} j, B = \left\|\frac{A^2}{2^{2N_{p,\mu}}}\right\| \text{ for } \cos A,
\end{aligned} \tag{31}$$

where one has $N_{p,1} = N_p$ and $m_p = m_{p,1}$.

For comparison, let A take the following form:

$$A = UDU^{-1}, \tag{32}$$

where U is a randomly generated n by n matrix, $D_a = Dial(d_1, d_2, \cdots, d_n)$ is a randomly generated n by n diagonal matrix with $d_i \in (a, b), i = 1, 2, \cdots, n$, so $f(A) = Uf(D)U^{-1}$, $f(D) = Dial(f(d_1), f(d_2), \cdots, f(d_n))$. In all the examples in this paper, we agree on the relative error as follows:

$$err(g)(A) = \frac{\|f(A) - Uf(D)U^{-1}\|_\infty}{\|Uf(D)U^{-1}\|}. \tag{33}$$

Let $A_i (i = 1, 2, \cdots, M)$ be M matrices produced by (32). From Figure 1, it is evident that the calculation accuracy is essentially the same for $\mu = 1, \frac{2}{3}, \frac{4}{3}$, but $\mu = 1$ provides the highest efficiency. The double-angle formula is notably more efficient and accurate than the triple-angle formula for $\cos A$. Henceforth, we will exclude the case that utilizes the triple-angle formula for $\cos A$.

Now let us juxtapose the algorithms presented in Table 1 with the internal functions and commonly used algorithms found in mathematical software. For ease of reference, let $err(g)$ denote the calculation error of matrix function $f(A_i)(i = 1, 2, \cdots, M)$, where a blank g implies the utilization of the algorithms in Tables 1–3 of this article. If $g = Sym, Pade, \cdots$, it signifies calculation according to the internal function, Padé approximation, $\cdots$, respectively. When $p = 16$, $m_p = 7, 4, 3$ for $e^A, \cos A(\cosh A), \sin A(\sinh A)$, respectively. We denote the Padé approximation of $\sin x$ as $r_{66}(x)$ and the Padé approximation of e^x and $\cos x$ as $r_{88}(x)$. In Table 1, the Padé approximation of $e^A, \cos A(\cosh A), \sin A(\sinh A)$ is as follows:

$$\begin{aligned}
r_{88}(x) &= \frac{1+\frac{x}{2}+\frac{x^2}{60}+\frac{x^3}{60}+\frac{x^4}{624}+\frac{x^5}{9360}+\frac{x^6}{205920}+\frac{x^7}{7207200}+\frac{x^8}{518918400}}{1-\frac{x}{2}+\frac{x^2}{60}-\frac{x^3}{60}+\frac{x^4}{624}-\frac{x^5}{9360}+\frac{x^6}{205920}-\frac{x^7}{7207200}+\frac{x^8}{518918400}} \text{ for } e^x,\\
r_{88}(x) &= \frac{1\mp\frac{260735x^2}{545628}+\frac{4375409x^4}{141863280}\mp\frac{7696415x^6}{1310867072}+\frac{807373373x^8}{2359470072960}}{1\pm\frac{12079x^2}{545628}+\frac{34709x^4}{141863280}\pm\frac{109247x^6}{65540835360}+\frac{11321x^8}{1814976979200}} \text{ for } \left\{\begin{array}{c}\cos x\\ \cosh x\end{array}\right\},\\
r_{66}(x) &= \frac{x\mp\frac{7x^3}{60}+\frac{x^5}{600}}{1\pm\frac{x^2}{20}+\frac{x^4}{600}\pm\frac{x^6}{14400}} \text{ for } \left\{\begin{array}{c}\sin x\\ \sinh x\end{array}\right\}.
\end{aligned} \tag{34}$$

For the algorithms in Table 1, the $err(g)(A)$ of random matrix from order 1 to 800 is as follows.

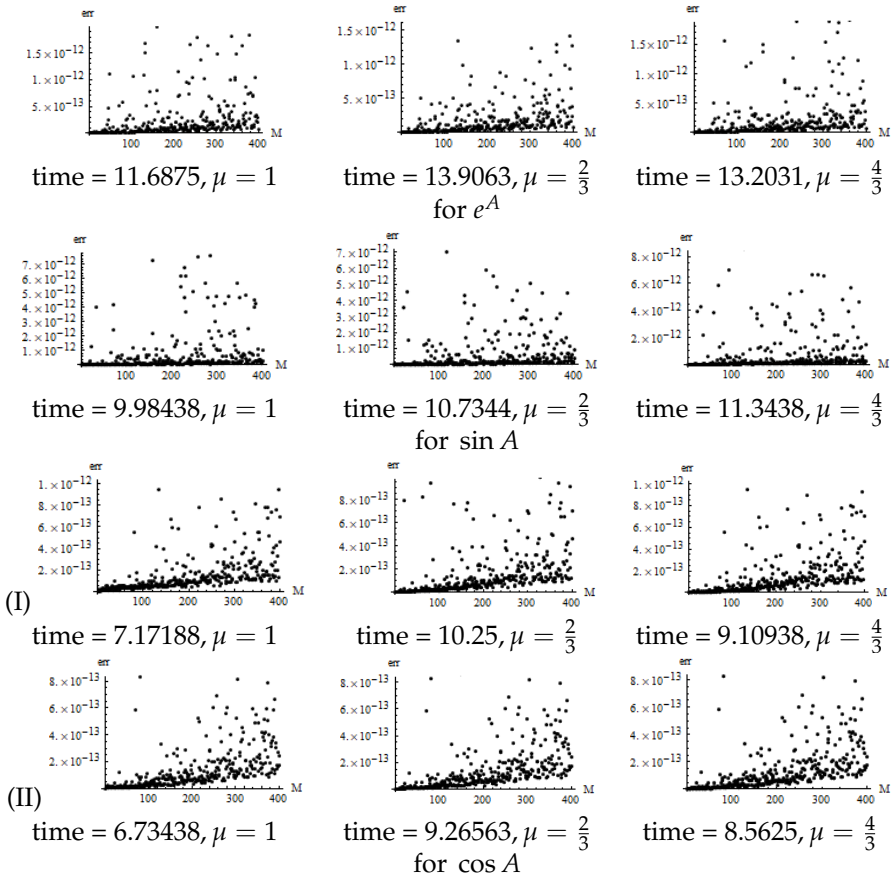


Figure 1. Error comparison of the algorithms in Table 1 for $\mu = 1, \frac{2}{3}, \frac{4}{3}$.

Table 2. The fast algorithms for $\tan A$, $\sec A$, $\tanh A$, and $sechA$.

	m_p, N	**Algorithm**
$\left\{ \begin{array}{c} \tan A \\ \tanh A \end{array} \right\}$	$m_p = \left[\sqrt{\frac{2p\ln 10}{\ln 3}}\right], N = N_1 + 1$	$B = \frac{A^2}{3^{2N}}$, $C = \frac{2A}{3^N}\sum_{j=0}^{m}(\pm 1)^j \sigma(j+1)\left(4^{j+1}-1\right)B^j$ $for(i=0, i<N, i++)$ $\quad C = C(3I \mp C^2)(I \mp 3C^2)^{-1}$ $\left\{ \begin{array}{c} \tan A \\ \tanh A \end{array} \right\} \approx C$
$\left\{ \begin{array}{c} \sec A \\ \sec hA \end{array} \right\}$	$m_p = \left[\sqrt{\frac{2p\ln 10}{\ln 3}}\right], N = N_2 + 1$	$B = \frac{A^2}{3^{2N}}, I + \sum_{j=1}^{m}(\mp 1)^j \widehat{E}_j B^j$ $for(i=0, i<N, i++)$ $\quad C = C^3(4I - 3C^2)^{-1}$ $\left\{ \begin{array}{c} \sec A \\ \sec hA \end{array} \right\} \approx C$
$N_1 = \left[\frac{p\ln 10 + \ln\left(2\sigma(m_p+2)\right)}{(2m_p+3)\ln 3} + \frac{\ln\lambda}{\ln 3}\right], N_2 = \left[\frac{p\ln 10 + \ln \widehat{E}_{m_p+1}}{(2m_p+3)\ln 3} + \frac{\ln\lambda}{\ln 3}\right]$.		

It is evident from Figure 2 that, for e^A, $\cos A$, and $\sin A$, the accuracy obtained by the three methods is essentially the same. However, when the matrix order is relatively high, the Taylor expansion method presented in Table 1 outperforms the others. In particular, for cos and sin A, the calculation time of the internal functions is approximately eight times longer than that of the algorithms in Table 1. Numerical calculations demonstrate that while the calculation speed of $\cos A$ and $\sin A$ using the internal functions $Re(e^{iA}) = \cos A$

and $Im(e^{iA}) = \sin A$ is faster than that of the internal functions $\cos A$ and $\sin A$, their calculation accuracy is relatively poor.

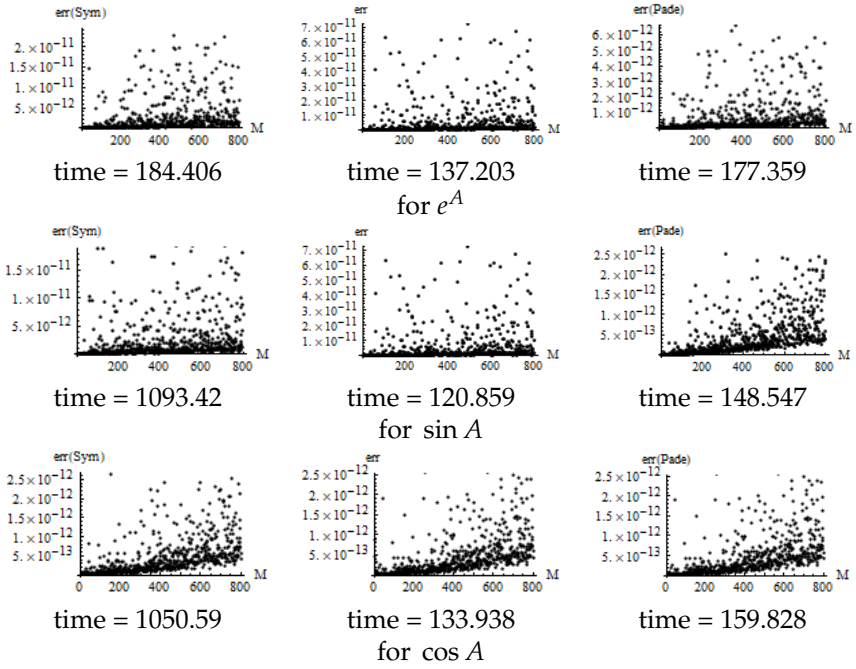


Figure 2. Error of e^A, $\cos A$, and $\sin A$ for three algorithms.

Table 3. The fast algorithms for $\cot A$, $\coth A$, $\csc A$, and $\mathrm{csch} A$.

	m_p, N	Algorithm
$\left\{ \begin{array}{c} \cot A \\ \coth A \end{array} \right\}$	$m_p = \left[\sqrt{\frac{7p\ln 10}{5\ln 3}}\right], N = N_1 + 1$	$B = \frac{A^2}{3^{2N}}, C = I - 2\sum_{j=1}^{m}(\pm 1)^j \sigma(j) B^j$ $for(i = 0, i < N, i++)$ $(C = 3(C(3B \mp C^2)(B \mp 3C^2)^{-1}; B = 9B)$ $\left\{ \begin{array}{c} \cot A \\ \coth A \end{array} \right\} \approx C$
$\left\{ \begin{array}{c} \csc A \\ \mathrm{csch} A \end{array} \right\}$	$m_p = \left[\sqrt{\frac{5p\ln 10}{2\ln 3}}\right], N = N_2 + 1$	$B = \frac{A^2}{3^{2N}}, I + 2\sum_{j=1}^{m+1}(\pm 1)^j \sigma(j)\left(1 - \frac{2}{4^j}\right) B^j$ $for(i = 0, i < N, i++)$ $(C = 3C^3(3C^2 \mp 4B)^{-1}; B = 9B)$ $\left\{ \begin{array}{c} \csc A \\ \mathrm{csch} A \end{array} \right\} \approx C$
$N_1 = \left[\frac{p\ln 10 + \ln(2\sigma(m_p+2))}{(2m_p+3)\ln 3} + \frac{\ln\lambda}{\ln 3}\right], N_2 = \left[\frac{p\ln 10 + \ln \widehat{E}_{m_p+1}}{(2m_p+3)\ln 3} + \frac{\ln\lambda}{\ln 3}\right]$		

For $\cosh A$ and $\sinh A$, the internal functions can be calculated by utilizing the following equation:

$$\cosh A = \frac{1}{2}(e^A + e^{-A}), \sinh A = \frac{1}{2}(e^A - e^{-A}). \tag{35}$$

Our experiment provides the following results for $err(g)(A)$ of random matrices with orders from 1 to 800.

It is evident from Figure 3 that using (35) not only doubles the calculation time but also makes the calculation error substantial. Furthermore, by observing Figures 2 and 3, it becomes apparent that the algorithms in Table 1 outperform the Padé approximation algorithm.

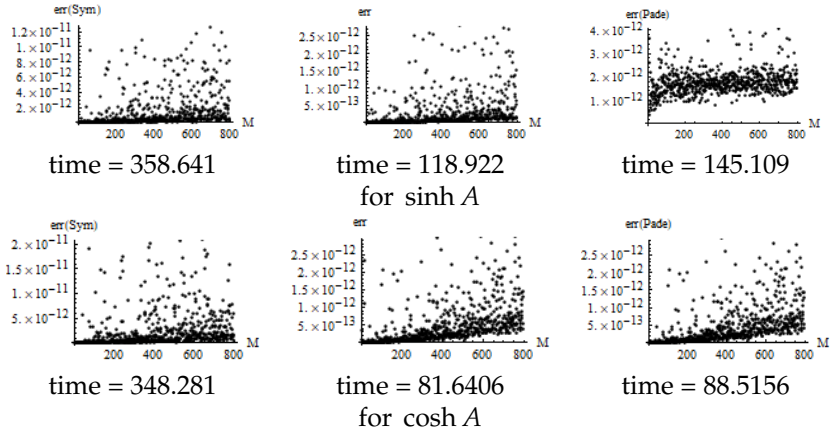


Figure 3. Error of sinh A and cosh A for three algorithms.

4. Fast Algorithm for tan A, tanh A, sec A, and $sechA$

Now let us contemplate the swift computation of tan A, tanh A, sec A, and $sechA$. Although these matrix functions can be computed using the following approach, we aim to directly compute them, as demonstrated in the previous section.

$$\begin{aligned} &\tan A = \sin A(\cos A)^{-1}, \sec A = (\cos A)^{-1}, \\ &\tanh A = \sinh A(\cosh A)^{-1} = \left(e^{A} - e^{-A}\right)\left(e^{A} + e^{-A}\right)^{-1}, \\ &sechA = (\cosh A)^{-1} = 2\left(e^{A} + e^{-A}\right)^{-1}. \end{aligned} \tag{36}$$

While the Padé approximation significantly contributes to matrix function computation, the numerical results in the previous section indicate that, based on the algorithm presented in this paper, neither the calculation accuracy nor the calculation speed is predominant. Therefore, this section does not adopt it.

Theorem 3. *For* tan A, sec A, tanh A, *and* $sechA$, *we have the following fast algorithm based on Theorem 1.*

In Table 2, for $\sigma(j)$ and $\widehat{E}_j$, we have the following recursive formula [1]:

$$\begin{aligned} &\sigma(2n) = \tfrac{4}{4n+1}\sum_{l=1}^{n-1}\sigma(l)\sigma(2n-l) + \tfrac{2}{4n+1}\sigma^{2}(n), n = 1, 2, \cdots, \\ &\sigma(2n-1) = \tfrac{4}{4n-1}\sum_{l=1}^{n-1}\sigma(l)\sigma(2n-1-l), n = 2, 3, \cdots \\ &\widehat{E}_n = -\sum_{k=0}^{n-1}\tfrac{\widehat{E}_k}{(2n-2k)!}, \widehat{E}_0 = 1, \sigma(1) = \tfrac{1}{6}. \end{aligned} \tag{37}$$

Using the algorithms in Table 2, our experiments generated the following data for the $err(g)(A)$ of random matrices from order 1 to 200.

Here, *Sym* indicates that matrix functions tan A, tanh A, sec A, and sech$A$, expressed by $\sin A(\cos A)^{-1}$, $e^{A} - e^{-A}\left(e^{A} + e^{-A}\right)^{-1}$, $(\cos A)^{-1}$ and $2\left(e^{A} + e^{-A}\right)^{-1}$, respectively, are computed using the internal functions. As observed in Figure 4, the algorithm's computation speed in Table 2 surpasses that of the internal function, particularly for tan A and sec A, where the speed is increased by 8–11 times compared to the internal function, while maintaining consistent accuracy is consistent. When the order n of the matrix is relatively large, for tanh A and sec hA, the calculation speed is twice that of the internal function, yet the precision remains the same as the internal function.

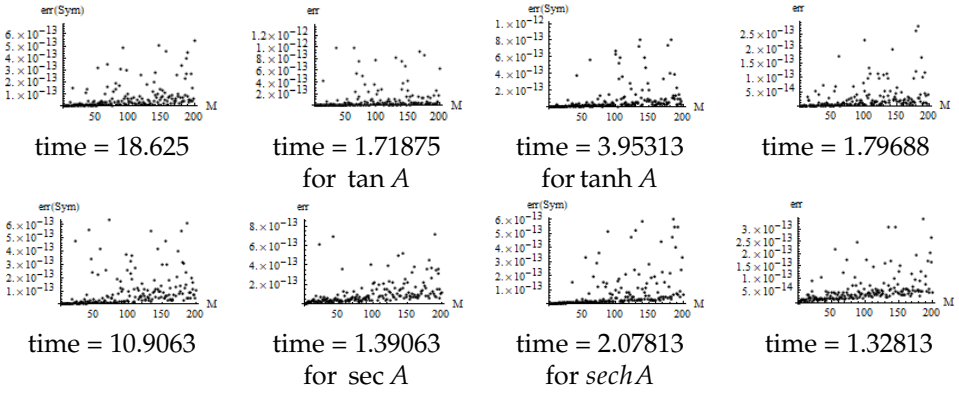


Figure 4. Error of the tan A, sec A, tanh A, and $sechA$ for two algorithms.

Let us consider implementing our algorithms for calculating the functions cot A, cot A, csc A, and csch$A$. By

$$\begin{aligned}
&\cot(At) \approx A^{-1} - 2\sum_{j=0}^{m} \sigma(j+1)A^{2j+1};\ (o < |A| < \pi, z \neq 0),\\
&\coth(At) \approx A^{-1} + 2\sum_{j=0}^{m} (-1)^j\sigma(j+1)A^{2j+1} \quad (0 < |A| < \pi),\\
&\operatorname{csc}(At) \approx A^{-1} + 2\sum_{j=0}^{m} \sigma(j+1)\left(1 - \tfrac{2}{4^{j+1}}\right)A^{2j+1} \quad (0 < |A| < \pi),\\
&\operatorname{csch}(At) \approx A^{-1} - 2\sum_{j=0}^{m} (-1)^j\sigma(j+1)\left(1 - \tfrac{2}{4^{j+1}}\right)A^{2j+1} \quad (0 < |A| < \pi),
\end{aligned} \tag{38}$$

we see that when $\|A\|$ is small, the norm of the inverse of A will be large, so the calculation using (38) will produce large errors. Therefore, we must utilize the following method:

$$\begin{aligned}
&\cot A = A^{-1}\cot_1 A, \coth A = A^{-1}\coth_1 A,\\
&\operatorname{csc} A = A^{-1}\operatorname{csc}_1 A, \operatorname{csch} A = A^{-1}\operatorname{csch}_1 A,
\end{aligned} \tag{39}$$

where

$$\begin{aligned}
&\cot_1 A \approx I - 2\sum_{j=1}^{m+1} \sigma(j)A^{2j};\ (|z| < \pi, z \neq 0),\\
&\coth_1 A \approx I - 2\sum_{j=1}^{m+1} (-1)^j\sigma(j)A^{2j} \quad (|z| < \pi, z \neq 0),\\
&\operatorname{csc}_1 A \approx I + 2\sum_{j=1}^{m+1} \sigma(j)\left(1 - \tfrac{2}{4^j}\right)A^{2j} \quad (|z| < \pi, z \neq 0),\\
&\operatorname{csch}_1 A \approx I + 2\sum_{j=1}^{m+1} (-1)^j\sigma(j)\left(1 - \tfrac{2}{4^j}\right)A^{2j} \quad (|z| < \pi, z \neq 0).
\end{aligned} \tag{40}$$

Theorem 4. *Using Theorem 1, we have the following algorithm for* cot A, coth A, csc A, *and* sec hA.

Implementation of the algorithms in Table 3 provides the following results for the matrix functions in (39).

Here, *Sym* signifies that matrix functions cot A, coth A, csc A, and sec hA are calculated using the internal functions with the regular formulas $\cos A(\sin A)^{-1}$, $\left(e^A + e^{-A}\right)\left(e^A - e^{-A}\right)^{-1}$, $(\sin A)^{-1}$, and $2\left(e^A - e^{-A}\right)^{-1}$, respectively. The results in Figure 5 are essentially consistent with the corresponding results in Figure 2.

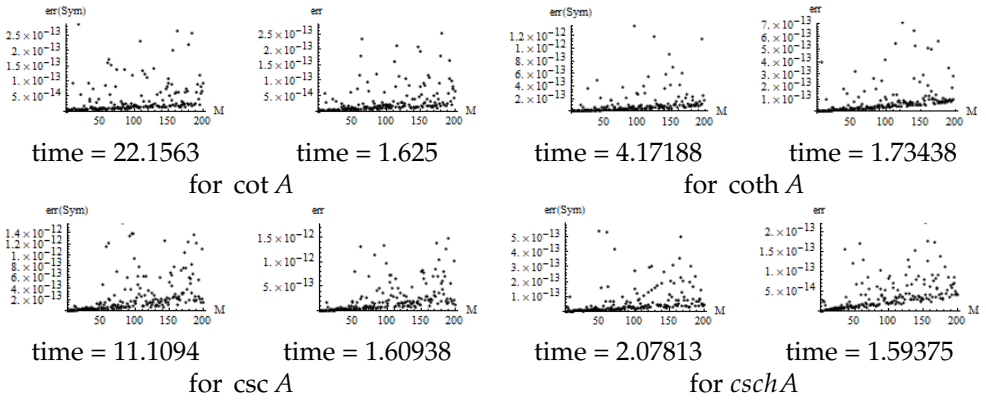


Figure 5. Error of cot A, cot A, csc A, and *cschA for two algorithms*.

5. Fast Algorithm for Logarithmic Matrix Function ln A and Some Related Functions

People consider various algorithms [17–22] for logarithmic matrix function ln A. One of the most effective methods we noticed was the following algorithm. The standard way of dealing with this problem is to use the square root operator repeatedly to bring A near the identity:

$$\ln A = 2^k \ln A_k, A_k = A^{1/2^k}. \tag{41}$$

As k increases, $A^{1/2^k} \to I$, so for sufficiently large k, we can apply a direct method to $A^{1/2^k}$. This procedure for the logarithm was introduced by Kenney and Laub [19]. The specific method is to choose k such that

$$\left\| A^{1/2^k} - I \right\| < \alpha < 1, \tag{42}$$

and usually, k is selected from 8. We slightly optimize the method as follows:

$$\ln A = \ln \lambda \cdot I + 2^k \ln A_k, A_k = A_0^{1/2^k}, A_0 = \frac{1}{\|A\|} A, \tag{43}$$

and

$$\left\| A_0^{1/2^k} - I \right\| < \alpha < 1, \tag{44}$$

where k is usually selected from 4.

For ln x, the algorithm that converges faster is

$$\ln x = \ln \frac{1-y}{1+y} \approx -2 \sum_{j=1}^{m} \frac{y^{2j+1}}{2j+1}, \tag{45}$$

where $y = \frac{1-x}{1+x}$. For ln A_k, we have

$$\ln A_k = \ln(I - B) - \ln(I + B) \approx -2 \sum_{j=1}^{m} \frac{B^{2j}}{2j+1}, \tag{46}$$

where $B = (I - A_k)(I + A_k)^{-1}$. Applying the Euler–Abel transformation [21], we obtain the following accelerated series:

$$\ln A_k \approx \tfrac{1}{3} BF^2 (I - 12 \sum_{j=1}^{m} \frac{B^{2j}}{(2j+3)(4j^2-1)}) - F, \tag{47}$$

where $F = 2B(I - B^2)^{-2}$. We can use the following Padé approximation of ln A_k [20].

$$\begin{aligned} \ln A_k &\approx (B + 2B^2 + \tfrac{47}{36} B^3 + \tfrac{11}{36} B^4 + \tfrac{137}{7560} B^5) F, \\ F &= (I + \tfrac{5}{2} B + \tfrac{20}{9} B^2 + \tfrac{5}{6} B^3 + \tfrac{5}{42} B^4 + \tfrac{1}{252} B^5)^{-1}, \end{aligned} \tag{48}$$

where $B = A_k - I$. We also provide some numerical results for the matrix function $\ln A$.

Based on (41)–(44) we also provide some numerical results of $\ln A$ using (46)–(48). It can be seen from Figure 6 that through optimization (43), the calculation accuracy of (46)–(48) is significantly improved, and the calculation speed of (46)–(48) is more than twice as fast as the inner function. The calculation speed advantage of the Padé approximation is obvious.

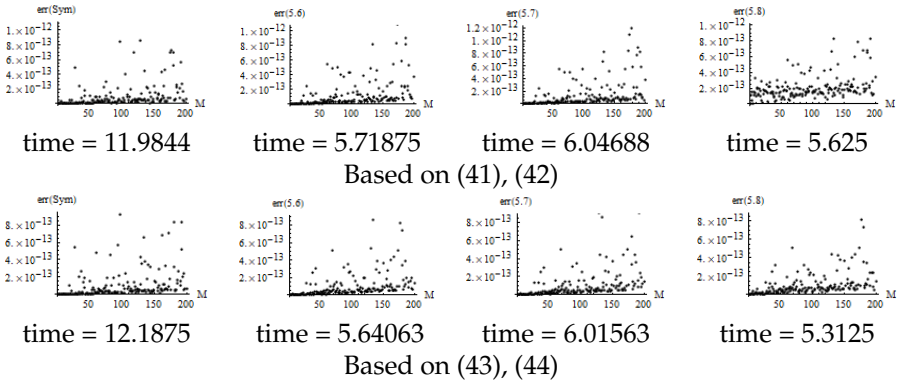


Figure 6. Error of $\ln A$ for the several algorithms.

We used the following identities to implement our general algorithm for the computations of the inverse trigonometric and hyperbolic functions. To control the length of the paper, we will not display the results here.

$$\begin{aligned} &\arcsin x = 2\arcsin \tfrac{x}{\sqrt{2(1+\sqrt{1-x^2})}}, \\ &\arctan x = 2\arctan \tfrac{x}{1+\sqrt{1+x^2}}, \\ &\arcsin x = x + \sum_{l=1}^{\infty} \tfrac{(2l-1)x^{2l+1}}{(2l)!(2l+1)}, \arctan x = \sum_{l=0}^{\infty} \tfrac{x^{2l+1}}{(2l+1)}, \end{aligned} \tag{49}$$

$$\begin{aligned} &\mathrm{arcsinh} x = 2\mathrm{arcsinh} \tfrac{x}{\sqrt{2(1+\sqrt{1+x^2})}}, \mathrm{arctanh} x = 2\mathrm{arctanh} \tfrac{x}{1+\sqrt{1-x^2}}, \\ &\mathrm{arcsinh} x = x + \sum_{l=1}^{\infty} \tfrac{(-1)^l(2l-1)x^{2l+1}}{(2l)!(2l+1)}, arctanhx = \sum_{l=0}^{\infty} \tfrac{(-1)^l x^{2l+1}}{(2l+1)}. \end{aligned} \tag{50}$$

For $\arcsin A, \arctan A, \mathrm{arcsinh} A,$ and $\mathrm{arctanh} A,$ under certain conditions, we can use the following identity.

$$\begin{aligned} &\arcsin A = -i\ln\left(\sqrt{I-A^2} + iA\right), \mathrm{arcsinh} A = \ln\left(\sqrt{I+A^2} + A\right), \\ &\arctan A = \tfrac{i}{2}\ln\left((iI+A)(iI-A)^{-1}\right), \\ &\mathrm{arctanh} A = \tfrac{1}{2}\ln\left((I+A)(I-A)^{-1}\right). \end{aligned} \tag{51}$$

6. Improved Calculation Formula for Initial Value Problems (3) and (8)

Now, let us consider the calculation of the integral in (11), i.e.,

$$\widehat{F}_k = \int_0^{\tau} e^{Hs} F_H(k\tau - s)ds. \tag{52}$$

If numerical integration is used to calculate this integral, it is quite time-consuming, and the calculation accuracy cannot be guaranteed. We adopt the following method here. For $nn = 1, 2, \cdots$, there is the following relationship:

$$\begin{aligned} \widehat{F}_k &= \sum_{j=0}^{nn-1} \int_{\frac{j\tau}{nn}}^{\frac{j\tau}{nn}} e^{Hs} F_H(k\tau - s)ds = \sum_{j=0}^{nn-1} T_{nn,j} \int_0^{\frac{\tau}{nn}} e^{Hs} F_H(k\tau - \tfrac{j\tau}{nn} - s)ds \\ &\approx \sum_{j=0}^{nn-1} T_{nn,j} \sum_{l=0}^{m} \tfrac{H^l}{l!} \int_0^{\frac{\tau}{nn}} s^l F_H(k\tau - \tfrac{j\tau}{nn} - s)ds, \end{aligned} \tag{53}$$

where

$$T_{nn,0} = I, T_{nn} = e^{H\frac{\tau}{nn}}, T_{nn,j} = e^{H\frac{j\tau}{nn}} = T_{nn}T_{nn,j-1}, j = 1, 2, \cdots, nn-1. \tag{54}$$

T_{nn} can be calculated by the following method:

$$\begin{array}{l} T_{nn} = I, zz = I; \\ \text{for}(i = 1, i \le m, i++)\ (zz = \frac{\frac{\tau}{nn}}{i} H \times zz; T_{nn} = T_{nn} + zz). \end{array} \tag{55}$$

So, (53) can be written as follows:

$$\widehat{F}_k \approx \sum_{j=0}^{nn-1} T_{nn,j} H_j, H_j = \sum_{l=0}^{m} \frac{H^l}{l!} \int_0^{\frac{\tau}{nn}} s^l F_H(k\tau - \frac{j\tau}{nn} - s)ds, \tag{56}$$

so we can convert the integral of the product of matrix exponential function and vector function into the integral of pure vector function.

The improved calculation formula for (3) is as follows:

$$\begin{array}{l} (55), (56), \\ U_k = TU_{k-1} + \sum_{j=0}^{nn-1} T_{nn,j} H_j, k = 1, 2, \cdots, \end{array} \tag{57}$$

where $T = e^{H\tau}$ is calculated by the algorithm in Table 1, and $T_{nn,j}$ *and* H_j are calculated by (55) and (56).

For

$$\int_0^{\frac{\tau}{nn}} s^l F_H(k\tau - \frac{j\tau}{nn} - s)ds, \tag{58}$$

the following two algorithms can be used.

(1) Integrable function method If $F(t) = (f_1(t), f_2(t), \cdots, f_v(t), \cdots, f_n(t))^T$, $f_v(t)$ is the algebraic sum of functions of the following types:

$$t^\alpha, t^l e^{-\omega t}, t^l \sin \omega t, t^l \cos \omega t, \tag{59}$$

then (58) can be expressed in analytical form, so H_j can also be expressed in analytical form. In this way, the calculation of (57) can be completed.

(2) Gauss three-point integral formula

Note that, under 16-bit calculation accuracy,

$$\frac{\tau}{nn} < \frac{1}{2^{n_{16}}\lambda_H} \approx \frac{0.0376434}{|\lambda_H|} < 0.0376434, \tag{60}$$

and the algebraic accuracy of the Gauss three-point integral formula is 5. Using the Gauss three-point integral formula,

$$\frac{1}{2}\left(\frac{\tau}{nn}\right)^{l+1} \sum_{i=1}^{3} w_i \left(\frac{1+b_i}{2}\right)^l F_H(k\tau - \frac{(j+1/2)\tau}{nn} - \frac{b_i\tau}{2nn}), \tag{61}$$

where $w_1 = w_3 = \frac{5}{9}, w_2 = \frac{8}{9}, b_1 = -\frac{\sqrt{15}}{5}, b_2 = 0, b_3 = \frac{\sqrt{15}}{5}$, to approximate (58) with high accuracy. So, (57) can be replaced by the following algorithm.

Algorithm 1. *Numerical solution algorithm of initial value problem (3): for* $nn = 1, 2, \cdots$

$$U_k = TU_{k-1} + \frac{\tau}{2nn} \sum_{j=0}^{nn-1} T_{nn,j} \sum_{i=1}^{3} w_i e^{H\frac{(1+b_i)\tau}{2nn}} F_H\left(k\tau - \frac{(2j+1+b_i)\tau}{2nn}\right), k = 1, 2, \cdots, \tag{62}$$

where

$$e^{H\frac{(1+b_i)\tau}{2nn}} \approx \frac{w_i\tau}{2nn}\left(I + \sum_{l=1}^{m} \frac{1}{l!}\left(\frac{(1+b_i)\tau}{2nn}\right)^l H^l\right). \tag{63}$$

In particular, if we let

$$nn = \begin{cases} [2^{n_p}\lambda_H\tau]+1 & 2^{n_p}\lambda_H\tau > 1 \\ 1 & 2^{n_p}\lambda_H\tau \le 1 \end{cases}, \tag{64}$$

then $N_p = 0, m = m_p = \left[\sqrt{\frac{9p\ln 10}{8\ln 2}}\right]$,

$$W_{i,H} \approx \frac{w_i\tau}{2nn}\left(I + \sum_{l=1}^{m_p}\frac{1}{l!}\left(\frac{(1+b_i)\tau}{2nn}\right)^l H^l\right). \tag{65}$$

At this time, the theoretical calculation accuracy of (62) is p.

Remark 3. *Note that,* $W_{i,H}(i = 1, 2, 3)$ *in (63) and (65) only needs to be calculated once, and is independent of* $F(t)$*, which is not only efficient but also easy to program.*

Now let us consider the numerical solution of the initial value problem (8). If it is converted to the initial value problem (3), then

$$\begin{pmatrix} x(t) \\ x'(t) \end{pmatrix} = U, H = \begin{pmatrix} 0 & I_{3\times 3} \\ -K & 0 \end{pmatrix}, F_H(t) = \begin{pmatrix} 0 \\ F(t) \end{pmatrix}, U_0 = \begin{pmatrix} x_0 \\ x_1 \end{pmatrix}. \tag{66}$$

By

$$W_{i,H} = e^{\tau H} \approx \begin{pmatrix} \cos_p(\tau A) & \tau\sin_{1,p}(\tau A) \\ -K\tau\sin_{1,p}(\tau A) & \cos_p(\tau A) \end{pmatrix}, A = \sqrt{K}, \tag{67}$$

where

$$\sin_1 A = \sum_{j=0}^{\infty}\frac{(-1)^j}{(2j+1)!}A^{2j} = \sum_{j=0}^{\infty}\frac{(-1)^j}{(2j+1)!}K^j, \tag{68}$$

and (68), we obtain the following algorithm.

Algorithm 2.

$$\begin{array}{l} \Delta t = \tau/3^N; T_\alpha = I; zz = I; A_2 = (A\Delta t)\times(A\Delta t); \\ for(i = 1, i \le m, i++)\,(zz = -\frac{A_2}{(2i)(2i+1)}\times zz; T_\alpha = T_\alpha + zz), \end{array} \tag{69}$$

and

$$\begin{array}{l} for(i = 0, i < N, i++) \\ (zz = T_\alpha \times T_\alpha, zz = zz \times A_2, A_2 = 9A_2, T_\alpha = T_\alpha - \frac{4}{3}T_\alpha \times zz)\,, \end{array} \tag{70}$$

where m_p *and* N_p *are determined using (31).*

Algorithm 3. *Numerical solution algorithm of initial value problem (8):*

$$\begin{array}{l} x_k = (\cos(\tau A)x_{k-1} + \tau\sin_1(\tau A)y_{k-1}) \\ \quad + \frac{\tau}{2nn}\sum\limits_{j=0}^{nn-1}\sum_{i=1}^{3}\frac{w_i(2j+1+b_i)\tau}{2nn}\sin_1\left(\frac{(2j+1+b_i)\tau}{2nn}A\right)F(k\tau - \frac{(2j+1+b_i)\tau}{2nn}) \\ y_k = -A\sin(\tau A)x_{k-1} + \cos(\tau A)y_{k-1} \\ \quad + \frac{\tau}{2nn}\sum\limits_{j=0}^{nn-1}\sum_{i=1}^{3}w_i\cos\left(\frac{(2j+1+b_i)\tau}{2nn}A\right)F(k\tau - \frac{(2j+1+b_i)\tau}{2nn}) \end{array} \tag{71}$$

for $nn = 1, 2, \cdots$, where $A = \sqrt{K}$. In particular, if nn is determined using (64), the theoretical accuracy of this algorithm is p and

$$\begin{aligned} x_k &= \cos(\tau A)x_{k-1} + \tau \sin_1(\tau A)y_{k-1} + \tfrac{\tau}{2nn}\sum_{j=0}^{nn-1}\sum_{i=1}^{3} w_i TW_{j,i}^{\sin} F(k\tau - \tfrac{(2j+1+b_i)\tau}{2nn}), \\ y_k &= \cos(\tau A)y_{k-1} - A\sin(\tau A)x_{k-1} + \tfrac{\tau}{2nn}\sum_{j=0}^{nn-1}\sum_{i=1}^{3} w_i TW_{j,i}^{\cos} F(k\tau - \tfrac{(2j+1+b_i)\tau}{2nn}), \\ x_0 &= x(0), y_0 = x'(0), \end{aligned} \tag{72}$$

where

$$\begin{aligned} TW_{j,i}^{\sin} &= \tfrac{(1+b_i)\tau}{2nn}\cos\Big(\tfrac{j\tau}{nn}A\Big)\sin_{1,m_p}\Big(\tfrac{(1+b_i)\tau}{2nn}A\Big) \\ &\quad + \tfrac{j\tau}{nn}\sin_1\Big(\tfrac{j\tau}{nn}A\Big)\cos_{m_p}\Big(\tfrac{(1+b_i)\tau}{2nn}A\Big), \\ TW_{j,i}^{\cos} &= \cos\Big(\tfrac{j\tau}{nn}A\Big)\cos_{m_p}\Big(\tfrac{(1+b_i)\tau}{2nn}A\Big) \\ &\quad - \sin\Big(\tfrac{j\tau}{nn}A\Big)\sin_{m_p}\Big(\tfrac{(1+b_i)\tau}{2nn}A\Big), \end{aligned} \tag{73}$$

and

$$\begin{aligned} \sin_{m_p}(sA) &= A\sum_{l=0}^{m_p}\tfrac{(-1)^l s^{2l+1}K^l}{(2l+1)!}, \\ \sin_{1,m_p}(sA) &= \sum_{l=0}^{m_p}\tfrac{(-1)^l s^{2l}K^l}{(2l+1)!}, \\ \cos_{m_p}(sA) &= I + \sum_{l=1}^{m_p+1}\tfrac{(-1)^l s^{2l}K^l}{(2l)!}. \end{aligned} \tag{74}$$

If A is reversible, then

$$\begin{aligned} x_k &= \cos(\tau A)x_{k-1} + A^{-1}\sin(\tau A)y_{k-1} \\ &\quad + \tfrac{\tau}{2nn}A^{-1}\sum_{j=0}^{nn-1}\textstyle\sum_{i=1}^{3} w_i \sin\Big(\tfrac{(2j+1+b_i)\tau}{2nn}A\Big)F(k\tau - \tfrac{(2j+1+b_i)\tau}{2nn}) \\ y_k &= -A\sin(\tau A)x_{k-1} + \cos(\tau A)y_{k-1} \\ &\quad + \tfrac{\tau}{2nn}\sum_{j=0}^{nn-1}\textstyle\sum_{i=1}^{3} w_i \cos\Big(\tfrac{(2j+1+b_i)\tau}{2nn}A\Big)F(k\tau - \tfrac{(2j+1+b_i)\tau}{2nn}). \end{aligned} \tag{75}$$

Therefore, (75) can be written in the form of complex function as follows:

$$z_k = e^{-i\tau A}z_{k-1} + \tfrac{i\tau}{2nn}A^{-1}\sum_{j=0}^{nn-1} e^{-\frac{j\tau i}{nn}A}\textstyle\sum_{l=1}^{3} w_l e^{-\frac{(1+b_l)\tau i}{2nn}A}F\Big(k\tau - \tfrac{(2j+1+b_l)\tau}{2nn}\Big), \tag{76}$$

or

$$Z_k = e^{-i\tau A}Z_{k-1} + \tfrac{\tau}{2nn}\sum_{j=0}^{nn-1} e^{-\frac{j\tau i}{nn}A}\textstyle\sum_{l=1}^{3} w_l e^{-\frac{(1+b_l)\tau i}{2nn}A}F\Big(k\tau - \tfrac{(2j+1+b_l)\tau}{2nn}\Big), \tag{77}$$

where $i = \sqrt{-1}$,

$$\begin{aligned} z_k &= x_k + iA^{-1}y_k, x_k = Rez_k, y_k = AImz_k, \\ Z_k &= y_k - iAx_k, y_k = ReZ_k, x_k = -A^{-1}ImZ_k. \end{aligned} \tag{78}$$

We first provide an examples to illustrate the effectiveness of Algorithms 1. Although we skip examples of Algorithm 2, we then include an example for the effectiveness of Algorithm 3.

Example 1. *In (8), let*

$$K = \begin{pmatrix} 1 & -1 & \\ -1 & \frac{5}{2} & \\ & 1 & 4 \end{pmatrix}, F(t) = \begin{pmatrix} \sin t \\ \frac{1}{2}\sin\ t \\ \cos\frac{\sqrt{2}t}{2} - \cos\Big(\sqrt{3}t\Big) + \sin t \end{pmatrix}, \tag{79}$$
$$x_0 = \begin{pmatrix} \frac{5}{2} & 0 & -2 \end{pmatrix}^T, x_1 = \begin{pmatrix} 1 & 1 & 2 \end{pmatrix}^T.$$

The analytic solution of the initial value problem (8) is

$$x(t) = \begin{pmatrix} 2\cos\frac{\sqrt{2}t}{2} + \frac{1}{2}\cos\left(\sqrt{3}t\right) + \sin t \\ \cos\frac{\sqrt{2}t}{2} - \cos\left(\sqrt{3}t\right) + \sin t \\ \sin(2t) - 2\cos(2t) \end{pmatrix}. \tag{80}$$

Let $err_{i,k} = \left\|x(k\tau) - x_{i,k}\right\|_\infty i = 1,2$, where $x_{1,k}, x_{2,k}(k = 0,1,2,\cdots,N)$ are the numerical solutions of the initial value problem (8) determined by Algorithm 1 and Algorithm 3, respectively. $err_{sym,k} = \left\|x(k\tau) - x_{sym}(k\tau)\right\|_\infty$, where $x_{sym}(t)$ is the numerical solution obtained by calling the function *NDSolve* in Mathematica. The numerical results of $err_{1,k}, err_{2,k}, err_{sym,k}[0,T]$ $(k = 0,1,2,\cdots,N\tau = T)$ are as follows.

It can be seen from Figure 7 that the calculation accuracy of Algorithm 1 is higher than that of Algorithm 3 but the calculation speed of Algorithm 3 is three times that of Algorithm 1. However, the error of numerical results of Mathematica internal functions is far greater than the error of Algorithms 1 and 3. In particular, the collapse is calculated at $t = 667.787$. After this point, the error increases exponentially with time t. However, Algorithms 1 and 3 are quite stable. For example, the calculation error in $[0, 8000]$ is still very small.

In Algorithms 1 and 3, let $nn = \mu nn_A, \mu = 0.5, 0.25, 0.125$, where nn_A is calculated according to (64). The numerical results are as follows.

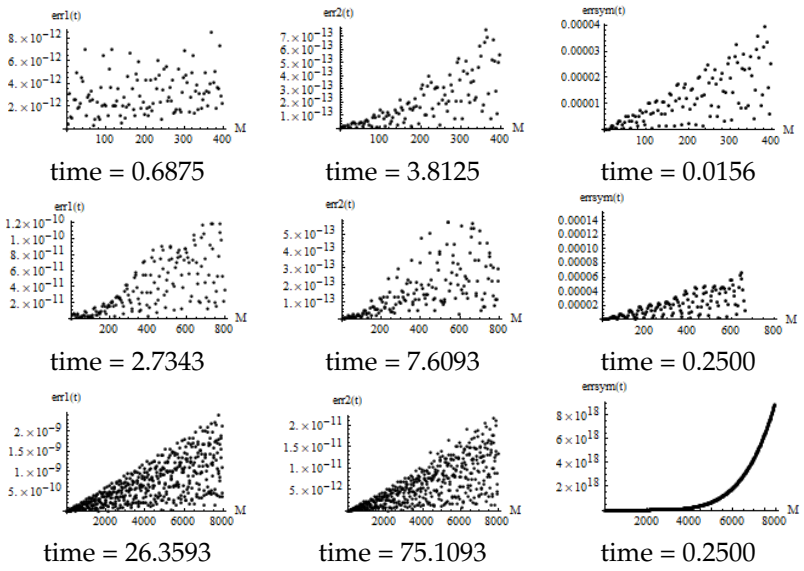


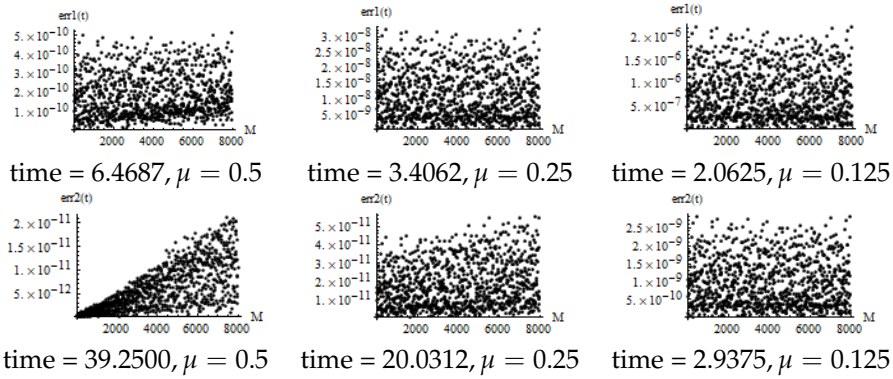
Figure 7. Error of numerical solution (8) for several algorithms (*nm* is calculated according to (64)).

Compared with Figures 7 and 8, when $\mu = 0.5$, the calculation accuracy is not reduced by much, and the calculation speed is almost doubled. For $\mu = 0.125$ Algorithms 1 and 3, the calculation accuracy is still high, and the calculation speed is 10 times higher than that in Figure 8.

Example 2. *In (8), let* $K = (k_{i,j})_{n\times n}$ *be a tridiagonal matrix.* $F(t) = (f_1(t), f_2(t), \cdots, f_n(t))^T$, $f_i(t) = A_i cos(\omega_i t + \varphi_i)$ $(i = 1,2,\cdots,n)$, $x_0 = (x_{0,1}, x_{0,2}, \cdots, x_{0,n})^T$, $x_1 = (x_{1,1}, x_{1,2}, \cdots, x_{1,n})^T$, *and they are randomly generated and satisfy the following conditions:*

$$\begin{gathered} |k_{i,i} - 9| \le 1, \left|k_{i,i+1} - \tfrac{3}{2}\right| \le \tfrac{1}{2}, \left|k_{i+1,i} - \tfrac{3}{2}\right| \le \tfrac{1}{2}, \\ \left|A_i - \tfrac{3}{2}\right| \le \tfrac{1}{2}, \left|\omega_i - \tfrac{5}{2}\right| \le \tfrac{1}{2}, 0 \le \varphi_i \le \tfrac{\pi}{2}, \\ |x_{0,i}| \le 1, |x_{1,i}| \le 1 (i = 1,2,\cdots,n). \end{gathered}$$

For $n = 50, 100, 200,$ and 400, the following numerical results are provided by Algorithms 1 and 3.

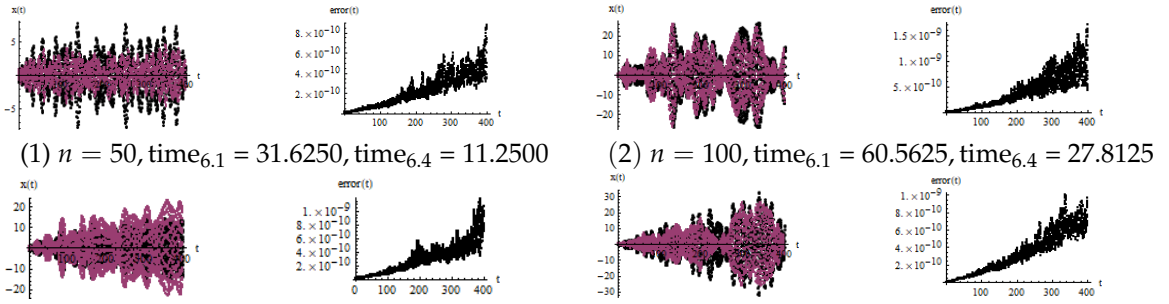


time = 6.4687, $\mu = 0.5$ time = 3.4062, $\mu = 0.25$ time = 2.0625, $\mu = 0.125$

time = 39.2500, $\mu = 0.5$ time = 20.0312, $\mu = 0.25$ time = 2.9375, $\mu = 0.125$

Figure 8. Error of numerical solution (8) for several algorithms.

In Figure 9, (1)–(4), the left is the first two components of the numerical solution, and the right is the difference between the numerical solutions obtained by the two algorithms. It is noted that sometimes ω_i $(i = 1, 2, \cdots, n)$ is close to the frequency of the solution, and flutter or resonance may occur. However, the values obtained by the two algorithms are still very close.



(1) $n = 50$, $\text{time}_{6.1} = 31.6250$, $\text{time}_{6.4} = 11.2500$ (2) $n = 100$, $\text{time}_{6.1} = 60.5625$, $\text{time}_{6.4} = 27.8125$

(3) $n = 200$, $\text{time}_{6.1} = 137.5625$, $\text{time}_{6.4} = 57.6875$ (4) $n = 100$, $\text{time}_{6.1} = 281.3125$, $\text{time}_{6.4} = 151.2187$

Figure 9. Numerical solutions and errors of two algorithms.

7. Conclusions

In this paper, we established algorithms for the computation of matrix exponential functions, matrix trigonometric functions, and matrix hyperbolic functions. Our numerical results show that these algorithms are faster with higher accuracy compared to existing algorithms and the system algorithms in Mathematica. In the series expansion of matrix function, the number of expanded items m_p is only related to the requirement of calculation accuracy, and has nothing to do with the matrix itself. However, the number of times in the iterative process N_p is not only related to the matrix itself but also closely related to t. When $|3^{n_p}\lambda_A t| \leq 1 (|2^{n_p}\lambda_A t| \leq 1)$, $N_p = 0$, the calculation of the matrix function becomes an algebraic operation, so the integral in (9)–(12) becomes an easy integral to compute. In brief, our algorithm allows users to set a required level of accuracy and then use it to choose appropriate values for parameters m and N to minimize computation time. In this manner, both accuracy and efficiency are addressed simultaneously.

Author Contributions: Conceptulation, H.Q. and Y.L.; mehtodology, H.Q. and Y.L.; software, H.Q.; validation, H.Q. and Y.L.; formal analysis, H.Q. and Y.L.; investigation, H.Q. and Y.L.; visualization, H.Q.; supervision, H.Q.; project administration, Y.L. All authors have read and agreed to the publication version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ruiz, P.; Sastre, J.; Ibáñeza, J.; Defez, E. High performance computing of the matrix exponential. *J. Comput. Applied Math.* **2015**, *291*, 370–379. [CrossRef]
2. Molert, C.; Loans, C.V. Nineteen dubious ways to compute the exponential of a matrix. *SIAM Rev. Soc. Ind. Applied Math.* **1978**, *20*, 801–836.
3. Higham, N.J. *Functions of Matrices: Theory and Computation*; SIAM: Philadelphia, PA, USA, 2008.
4. Al-Mohy, A.H.; Higham, N.J. Computing the Action of the Matrix Exponential, with an Application to Exponential Integrators. *SIAM J. Sci. Comput.* **2011**, *33*, 488–511. [CrossRef]
5. Higham, N.J. The scaling and squaring method for the matrix exponential resisted. *SIAM J. Matrix Anal. Appl.* **2005**, *26*, 1179–1193. [CrossRef]
6. Bader, P.; Blanes, S.; Casas, F. Computing the Matrix Exponential with an Optimized Taylor Polynomial approximation. *Mathematics* **2019**, *7*, 1174. [CrossRef]
7. Seydaoglu, K.M.; Bader, P.; Blanes, S.; Casas, F. Computing the matrix sine and cosine simultaneously with a reduced number of products. *arXiv* **2020**, arXiv:2010.00465v1.
8. Hargreaves, G.I.; Higham, N.J. Efficient algorithms for the matrix cosine and sine. *Numer. Algorithms* **2005**, *40*, 383–400. [CrossRef]
9. Alonso, P.; Ibáñez, J.; Sastre, J.; Peinado, J.; Defez, E. Efficient and accurate algorithms for computing matrix trigonometric functions. *J. Comput. Appl. Math.* **2017**, *309*, 325–332. [CrossRef]
10. Alonso, P.; Peinado, J.; Ibáñez, J.; Sastre, J.; Defez, E. Computing Matrix Trigonometric Functions with GPUs through Matlab. *J. Supercomput.* **2019**, *75*, 1227–1240. [CrossRef]
11. Al-Mohy, A.H.; Higham, N.J.; Relton, S.D. New Algorithms for Computing the Matrix Sine and Cosine Separately or Simultaneously. *SIAM J. Sci. Comput.* **2015**, *37*, A456–A487. [CrossRef]
12. Serbin, S.M.; Blalock, S.A. An algorithm for computing the matrix cosine. *SIAM J. Sci. Statist. Comput.* **1980**, *1*, 198–204. [CrossRef]
13. Cooper, B.L.; Bingulac, S. Computational improvement in the calculation of the natural log of a square matrix. *Electron. Lett.* **1990**, *26*, 861–862. [CrossRef]
14. Wu, F.; Zhu, L.; Shi, Q. Efficient Computational method for matrix function in dynamic problems. *Acta Mech. Sin.* **2023**, *39*, 522451. [CrossRef]
15. Cardoso, J.R.; Sadeghi, A. A technique for improving the computation of functions of triangular matrices. *Int. J. Comput. Math.* **2022**, *99*, 2449–2465. [CrossRef]
16. Jiang, Y.; He, C.; Qin, H. An accelerated algorithm for high precision and fast calculation of basic elementary functions. *Math. Pract. Theory* **2017**, *47*, 238–246.
17. Fasi, M.; Higham, N.J. Multiprecision algorithm for Computing matrix logarithm. *SIAM J. Matrixanal. Appl.* **2018**, *39*, 472–491. [CrossRef]
18. Ibáñez, J.; Sastre, J.; Ruiz, P.; Alonso, J.M.; Defez, E. An Improved Taylor Algorithm for Computing the Matrix Logarithm. *Mathematics* **2021**, *9*, 2018. [CrossRef]
19. Cheng, S.H.; Higham, N.J.; Kenney, C.S.; Laub, A.J. Approximating the Logarithm of a Matrix to Specified Accuracy. *SIAM J. Matrix Anal. Appl.* **2001**, *22*, 1112–1125. [CrossRef]
20. Kenney, C.; Laub, A.J. Condition estimates for matrix functions. *SIAM J. Matrix Anal. Appl.* **1989**, *10*, 191–209. [CrossRef]
21. Kenney, C.; Laub, A.J. Padé error estimates for the logarithm of a matrix. *Internat. J. Control.* **1989**, *50*, 707–730. [CrossRef]
22. Milovanović, G.V.; Rančić, S.L. Some improvements in calculation of the natural log of a square matrix. In Proceedings of the Third International Conference on Functional Analysis and Approximation Theory, Aquafredda di Maratea, Potenza, Italy, 23–28 September 1996; Issue 52, Volume II.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

AppliedMath

MDPI

Article

Some Comments about the p-Generalized Negative Binomial (NBp) Model

Daniel A. Griffith

School of Economic, Political, and Policy Science, University of Texas at Dallas, Richardson, TX 75080, USA; dagriffith@utdallas.edu

Abstract: This paper describes various selected properties and features of negative binomial (NB) random variables, with special reference to NB2 (i.e., p = 2), and some generalizations to NBp (i.e., p ≥ 2), specifications. It presents new results (e.g., the NBp moment-generating function) with regard to the relationship between a sample mean and its accompanying variance, as well as spatial statistical/econometric numerical and empirical examples, whose parameter estimators are maximum likelihood or method of moment ones. Finally, it highlights the Moran eigenvector spatial filtering methodology within the context of generalized linear modeling, demonstrating it in terms of spatial negative binomial regression. Its overall conclusion is a bolstering of important findings the literature already reports with a newly recognized empirical example of an NB3 phenomenon.

Keywords: eigenvector spatial filter; moment-generating function; negative binomial; population density; Puerto Rico

check for updates

Citation: Griffith, D.A. Some Comments about the p-Generalized Negative Binomial (NBp) Model. *AppliedMath* **2024**, *4*, 731–742. https://doi.org/10.3390/appliedmath4020039

Academic Editor: Tommi Sottinen

Received: 24 March 2024
Revised: 3 June 2024
Accepted: 6 June 2024
Published: 11 June 2024

Copyright: © 2024 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Many contemporary random variables (RVs) measure counts (with a zero to infinity support), implying that they are Poisson variates, but without equi-dispersion (i.e., a violation of the $\mu = \sigma^2$ property, most often such that $\mu < \sigma^2$), further implying that their best distributional description is a Poisson variable with a gamma-distributed mean (i.e., a gamma–Poisson mixture), which translates nicely into a negative binomial (NB) RV. However, a standard NB does not always successfully account for this mean–variance discrepancy in such circumstances. Cameron and Trivedi [1] introduce a generalized version of the NB probability model now known as the NBp model, where p is a positive integer, whose goal is to better capture under/over-dispersion; this variate differs from the negative binomial process (NBP) model (e.g. [2]). The literature contains little discussion about this former generalized model, with the exception of Cameron and Trivedi's initial treatment of it and papers by Greene [3] and by Di et al. [4]. Of note is that an R package (version ≥ 3.00) exists for implementing this model [4]. The purpose of this paper is to address this gap in the literature by (1) clarifying/correcting several misleading technical details in part of Greene's paper; (2) presenting new results; (3) illuminating selected small and large sample properties on NB models; and (4) presenting new empirical spatial statistics/econometrics instances involving spatial autocorrelation (SA).

A parametric mixture conceptualization—analogous to a Bayesian conjugate prior formulation—is the context of this discussion: the NB specification describes a count's variable Y that is Poisson distributed, with its mean μ being gamma distributed rather than constant. Accordingly, the gamma RV has parameters $\alpha\mu^{2-p}$ and μ^{p-1}/α, where μ denotes its mean and $1/\alpha$ denotes its dispersion parameter, yielding a global NBp model with parameters $r = \alpha\mu^{2-p}$ and probability $p = 1/(1 + \mu^{p-1}/\alpha)$ for some non-negative integer $p > 0$; $p = 1, 2$ are standard NB specifications, with the former being the Poisson RV and the latter being the popular NB2 variate.

2. Materials and Methods: Clarifying Greene's Presentation

Confusing equations appearing in Greene [1] include (2-13) and (2-14), particularly when one contrasts them with his appendix equation (A-2) with regard to the p exponent. In addition, this latter equation is missing a subscript i on the λ variable appearing in the numerator of its third term on its right-hand side. Discussion pertaining to equations (2-4) and (2-5) would be clearer if the posited gamma distribution is stated in its standard symbology as $\Gamma(\theta, 1/\theta)$. Greene's equations (2-13)—in which he incorrectly states $q = 1/1(1 + \theta)$ rather than $q = 1/(1 + 1/\theta)$—and (2-14)—in which he states $s = \lambda/(\lambda + \theta\lambda^{2-p})$ rather than $1/(1 + \mu^{p-2}/\alpha)$—yield, for independent and identically distributed (iid) NBps, the following probability mass function:

$$\frac{\Gamma(y + \alpha\mu^{3-p})}{\Gamma(y+1)\Gamma(\alpha\mu^{3-p})}\left(\frac{1}{1+\mu^{p-2}/\alpha}\right)^{\alpha\mu^{3-p}}\left(\frac{\mu^{p-2}/\alpha}{1+\mu^{p-2}/\alpha}\right)^{y}, y = 0, 1, 2, \ldots, \quad (1)$$

where, adopting more typical literature-appearing notation (particularly for gamma–Poisson specifications) in this paper, μ (denoting the mean) is equivalent to λ and α is the same as θ in Greene's paper, the subscripts i are not used in this version because it represents a global equation, and the ratio of gamma functions term also can be rewritten employing combinatorial binomial coefficients. The difference is that Greene's equation (2-14) has an exponent of (2–p) rather than (p–2). Equation (1) renders the revised results of

$$E(Y) = \mu \text{ and } E[(Y - \mu)^2] = \mu(1 + \mu^{p-1}/\alpha). \quad (2)$$

Greene's equation (2-14) specification renders

$$\frac{\Gamma(y + \alpha\mu^{2-p})}{\Gamma(y+1)\Gamma(\alpha\mu^{2-p})}\left(\frac{1}{1+\alpha\mu^{1-p}}\right)^{\alpha\mu^{2-p}}\left(\frac{\alpha\mu^{1-p}}{1+\alpha\mu^{1-p}}\right)^{y}, y = 0, 1, 2, \ldots,$$

which is not his equation (2-15), even for the specific standard case of p = 2. His probability mass function should be Equation (1), rather than

$$\frac{\Gamma(y + \alpha\mu^{2-p})}{\Gamma(y+1)\Gamma(\alpha\mu^{2-p})}\left(\frac{1}{1+\alpha\mu^{1-p}}\right)^{\alpha\mu^{2-p}}\left(\frac{\alpha\mu^{1-p}}{1+\alpha\mu^{1-p}}\right)^{y}, y = 0, 1, 2, \ldots,$$

which is neither the result for a Poisson parametric mixture probability defined by $Prob(y) = e^{-\mu}\mu^{y}/(y!)$ with $\mu \sim \Gamma(\alpha\mu^{2-p}, \mu^{p-1}/\alpha)$, nor the outcome from Greene's replacing α with $\alpha\mu^{2-p}$ in this mixture, which renders $E(Y) = \mu$ and $E[(Y - \mu)^2] = \mu(1 + \mu^{p-3}/\alpha)$.

3. Results: Articulating NBp Variates

Both mathematical statistics analytical results and simulation experimental output help establish an understanding of a generalization of NB1 and NB2 to NBp. This section furnishes some of each.

3.1. The NBp Moment-Generating Function and Asymptotic Normality

Particularly because the method of moments is a useful estimation technique for parameter α [5], knowing the moments of the NBp specification is desirable. The absence of closed-form maximum likelihood estimators (MLEs, e.g., see [6–10]) makes this ensuing invention even more appealing.

3.1.1. The NBp Moment-Generating Function (mgf)

The NBp mgf is

$$\left[\frac{1}{1 + (\mu^{p-1}/\alpha)(1 - e^{t})}\right]^{\alpha\mu^{2-p}},$$

with the first moment about the origin being

$$\frac{\partial}{\partial t}\left[\frac{\alpha\ \mu^{2-p}}{\alpha\ \mu^{2-p}+\mu\ \left(1-e^{t}\right)}\right]^{\alpha\ \mu^{2-p}}\Bigg|_{t=0}=\mu,$$

and, hence, the moments about the mean mgf being

$$e^{-\mu t}\left[\frac{\alpha\ \mu^{2-p}}{\alpha\ \mu^{2-p}+\mu\ \left(1-e^{t}\right)}\right]^{\alpha\ \mu^{2-p}}. \tag{3}$$

The second moment about the mean is $\mu(1+\mu^{p-1}/\alpha)$.

3.1.2. Asymptotic Normality and an NBp RV

Matching moments is one way to establish equivalency between distributions [11]. Skewness and kurtosis involve the third and fourth moments about the mean, which are derivable from Equation (3), such that for an NBp RV

$$\text{skewness}=\frac{1+2\ \mu^{p-1}/\alpha}{\sqrt{\mu}\sqrt{1+\mu^{p-1}/\alpha}},\ \text{and kurtosis}=3+\frac{6\ \mu^{p-2}}{\alpha}+\frac{1}{\mu\ \left(1+\mu^{p-1}/\alpha\right)}.$$

The Poisson skewness term, $1/\sqrt{\mu}$, is visible in this preceding first expression. Here, skewness does not necessarily converge on zero, and kurtosis does not necessarily converge on 3, unless the over-dispersion parameter, $1/\alpha$, as well as p exhibit certain behaviors and/or properties. In other words, asymptotical normality for the NBp distribution is not obvious. Bagui and Mehra [12] offer ways to prove this asymptotic feature, one of which involves its mgf.

Meanwhile, a classical theorem states that if random samples are drawn from the same normal distribution (i.e., iid), then their sample mean and variance statistics are independent [because s^2 is a function of $(y_i-\bar{y})$; Figure 1a]. Figure 1 portrays scatterplots of relationships between sample s^2 and s^2 from simulation experiment results for 10,000 replications of samples of size n = 100 (a well-known appropriate size to ensure that the classical Central Limit Theorem is at work). For this experiment, the linear correlation between $\bar{y}$ and s^2 for a standard normal RV (i.e., iid) is –0.014 (Figure 1a), whereas that for an NBp RV with p = 2, μ = 1000, and α = 1000 (i.e., iid) is 0.027 (Figure 1b). An extension of this conventional, well-known theorem is as follows:

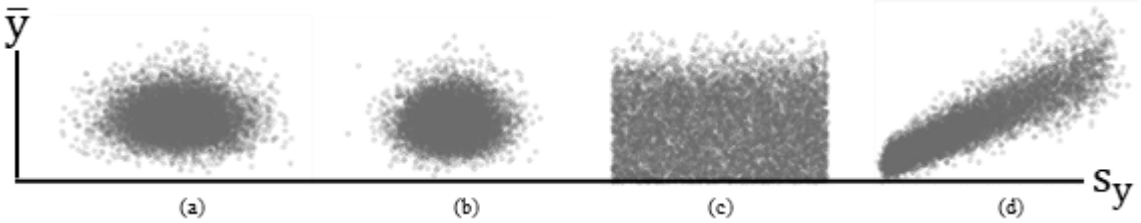


Figure 1. Selected simulated (10,000 replications, n = 100) relationships between $\bar{y}$ and s^2. Left (**a**): iid normal RVs. Left middle (**b**): iid NB2 RVs. Right middle (**c**): i~id normal RVs. Right (**d**): i~id NBp RVs.

Theorem 1. *The linear correlation between $\bar{y}$ and s^2 for a mixture of independent and non-identically distributed (i~id) normal RVs is zero.*

Proof of Theorem 1. For a normal RV, μ can be any real value, and σ^2 can be any positive real number. Because these two sets of values can be paired in any of the infinite number of possible ways that exist, each value of μ is paired with relatively large, intermediate, and

small values of σ^2, and each value of σ^2 is paired with relatively large, intermediate, and small values of μ. Consequently, the linear correlation between these two variates is zero. □

This result (see Figure 1c; r = 0.001) is in contrast to that for an NBp RV for which μ appears in its variance formula, introducing linear correlation between the two (Figure 1d; r = 0.905). The formal postulation extension here is as follows:

Theorem 2. *The linear correlation between $\bar{y}$ and s^2 for a mixture of i~id NBp RVs is $\frac{\sqrt{3(2p+1)}}{p+2} > 0$, $p \geq 1$.*

Proof of Theorem 2. The linear product moment correlation coefficient is

$$\rho_c = \frac{\left\{\int_0^c \left[\mu - \left(\int_0^c \mu \, d\mu\right)/c\right] \times \left[\mu\left(1+\mu^{p-1}/\alpha\right) - \left(\int_0^c \mu\left(1+\mu^{p-1}/\alpha\right) d\mu\right)/c\right] d\mu\right\}}{\sqrt{\int_0^c \left[\mu - \left(\int_0^c \mu \, d\mu\right)/c\right]^2 d\mu} \times \sqrt{\int_0^c \left[\mu\left(1+\mu^{p-1}/\alpha\right) - \left(\int_0^c \mu\left(1+\mu^{p-1}/\alpha\right) d\mu\right)/c\right]^2 d\mu}}.$$

$$\lim_{c\to\infty} \rho_c = \frac{\sqrt{3(2p+1)}}{p+2} \text{ because } p > 0. \square$$

This new theorem indicates that if p = 1 (i.e., a Poisson RV), then $\rho = 1$, and for the standard NB2 case of p = 2, $\rho \approx 0.968$. The parameter p needs to be 20 before this linear correlation is approximately 0.5. The dispersion parameter α does not affect this linear correlation. This new theorem indicates that for the aforementioned NBp, the sample mean and variance, $\bar{y}$ and s^2, are correlated (Figure 1d). In other words, the linear correlation between $\bar{y}$ and s^2 for a mixture of normal distributions is zero, whereas for a mixture of NBp distributions, it is asymptotically zero as $p \to \infty$. Although this linear correlation remains zero when normal distributions are pooled, it can be as large as 1 for pooled NBp distributions.

Of note is that Theorem 1 holds even for a normal distribution restricted in such a way that $\sigma^2 = \mu(1 + \mu^{p-1}/\alpha)$, with suitable restrictions on both p and μ (e.g., p must be even and the μ domain interval [–1, 0] must be removed from the range of σ^2). Linear correlation occurs for this particular normal distribution case by, for example, imposing a constraint such as $\mu > 0$, which then also allows p to be odd and linear correlation results to be described by Theorem 2.

3.2. MLE and Method of Moments Estimation (MME) of the Dispersion Parameter α

Various estimation techniques exist, including MLE, MME (e.g., see [6,13]), and least squares (see [14–16]). Frequently, these techniques yield exactly the same estimators. The estimation of the dispersion parameter $1/\alpha$ of an NBp often is by MLE and/or MME. A μ closer to zero and a small sample size can create a situation in which an MLE and/or MME is inaccurate and/or unstable. MME is the simpler of the two techniques for estimating $1/\alpha$ [5]. An MLE is unable to produce a solution that is an estimator in closed form for this NBp parameter (see Appendix A).

3.2.1. Estimation of Parameter Exponent p

Rather than setting p = 2, Equation (1) alludes to p as a parameter, albeit one restricted to the positive integers (i.e., discrete, not continuous). As a third parameter, its MME estimation requires, for example, adding the third moment about the mean, m_3, to the set of estimation equations. Accordingly,

$$\widehat{\mu} = \sum_{i=1}^{n} y_i/n = \bar{y},$$

$$\widehat{\alpha} = \widehat{\mu}\frac{\widehat{\mu}^{p}}{\widehat{\mu} + \sum_{i=1}^{n}(y_i - \widehat{\mu})^2/(n-1)}, \text{and}$$
$$\frac{1 + 2\,\widehat{\mu}^{p-1}/\widehat{\alpha}}{\sqrt{\widehat{\mu}}\sqrt{1 + \widehat{\mu}^{p-1}/\widehat{\alpha}}} = \frac{\sum_{i=1}^{n}(y_i - \widehat{\mu})^3}{n-1}.$$

Solving this triplet of equations for p yields an equation that does not contain p, nor does the one incorporating the fourth moment. Thus, MME cannot render an estimate of p. Meanwhile, an MLE, which involves the difference calculus here because p is an integer, fails to produce a closed-form solution.

The implication of this outcome is that generalizing the standard NB2 variance term from $\mu(1 + \mu/\alpha)$ to the generalized NBp term $\mu(1 + \mu^{p-1}/\alpha)$ is similar to quasi-likelihood estimation with regard to parameter p. Denote $\frac{1}{\widehat{\alpha}}$ with $\widehat{\alpha}^{-1}$; if $\widehat{\alpha}^{-1}$ seems too large, then a k^{th} multiple of $\widehat{\mu}$ (i.e., $\widehat{\mu}^{k}$), where k is a positive integer, can be factored from it, effectively setting p = 2 + k. This factorization is nonconstant across the values of μ_i, with subscript i denoting an individual observation, when the mean varies from observation to observation (this is the situation Greene [1] presents).

3.2.2. A Simple n = 3 Numerical Example

Consider a sample of size n = 3, with the number of counts per observation, Y, being as follows: $y_i = \{2, 5, 8\}$. Then

$$\bar{y} = (2 + 5 + 8)/3 = 5 \text{ and } s^2 = 18/(3-1) = 9 = 5\left[1 + 5/\left(25/4\right)\right].$$

The MLEs are

$$\widehat{\mu} = 5 \text{ and } \widehat{\alpha}^{-1} \approx 0.0476.$$

The MMEs are

$$\widehat{\mu} = 5 \text{ and } \widehat{\alpha}^{-1} = 0.1600.$$

Table 1 summarizes results from a Monte Carlo simulation experiment with 10,000 replications (to invoke the Law of Large Numbers), for NB2[5, 5(1 + 5/0.0476)] and NB2[5, 5(1 + 5/0.1600)]. Figure 2 portrays the collective MLE and MME estimates for these simulation results.

As expected, both estimation techniques yield a sound estimate of the population mean, μ. The techniques furnish different estimates of the dispersion parameter, although both are within the 95% confidence interval of the other's estimate. The MLEs are zero more often than MMEs are. Most of these zero MMEs actually are negative estimates that have been replaced by zero via truncation (these zero values were employed when calculating the summary statistics appearing in Table 1). These values should be considered infeasible because α is the shape parameter of a gamma RV, and as such must be positive. Figure 2 furnishes a visual box-and-whisker plot—the upper and lower edges of a box denote the first and third quartiles, the internal box line denotes the median, and the whisker tips denote 1.5 times the inter-quartile range (i.e., the third minus the first quartile values)—comparison between the two estimators (i.e., the MLEs and MMEs), highlighting that the MLEs have more outliers (denoted by asterisks). Excessive zeroes and frequency of outlier estimates for MLEs, and negative estimates for MMEs, are among the reasons these estimators are considered inaccurate and/or unstable. One general implication of these results is that neither estimator is particularly good for small samples, but that MME may well be preferable. This small sample case illustrates that an MLE and MME can yield different estimates (also see [17]).

Table 1. Summary Monte Carlo simulation experiment results for NB2[5, 5(1 + 5/0.0476)] and NB2[5, 5(1 + 5/0.1600)].

Parameter	Estimate	Standard Deviation	Range	% 0 s	% < 0
using MLEs as the population parameters					
μ	4.984	1.410	0.7–11.7		
$1/\alpha_{MLE}$	0.073	0.231	0.0–7.2	71.1	
$1/\alpha_{MME}$	0.113	0.213	0.0–2.7	5.4	52.2
using MMEs as the population parameters					
μ	5.003	1.714	0.3–14.3		
$1/\alpha_{MLE}$	0.137	0.315	0.0–7.7	58.4	
$1/\alpha_{MME}$	0.190	0.295	0.0–2.8	1.3	40.6

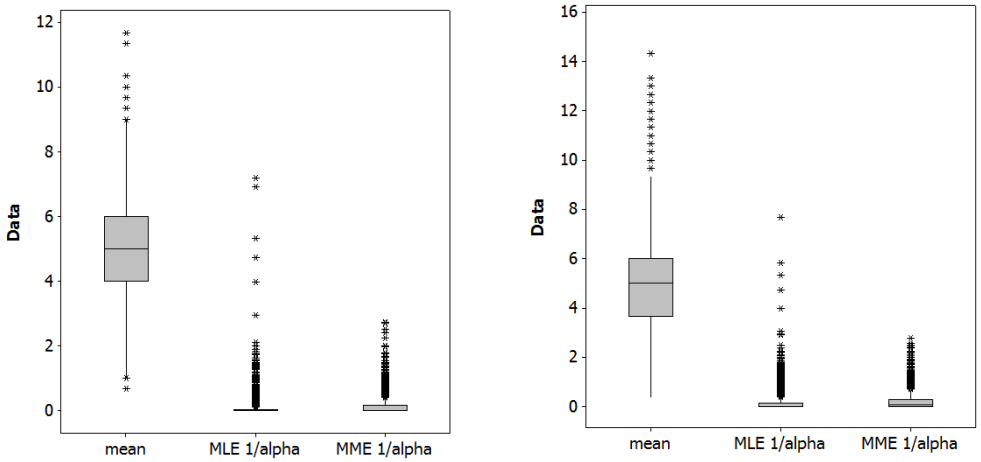


Figure 2. Monte Carlo simulation experiment boxplots (as usual, asterisks denote outliers). Left (**a**): NB2[5, 5(1 + 5/0.0476)]. Right (**b**): NB2[5, 5(1 + 5/0.1600)].

3.2.3. Moran Eigenvector Spatial Filtering: A Brief Overview

The next section presents a simple example of interest, especially to criminologists, ecologists, entrepreneurs and marketers, environmental scientists, epidemiologists, geospatial analysts, sociologists, urban planners and urbanists, and urban and regional economists, among others, namely, population density, perhaps one of the most commonly utilized covariates in the social and behavioral sciences; it serves as a preface to Section 4. Georeferenced data require spatial statistics/econometrics methodology in order to properly account for latent SA in those data. SA is the correlation among observations that arises because of their relative proximity in geographic space, because either they are affected by some common factor or they directly interact with each other over geographic space. It is analogous to, but more complex than, serial correlation in time series data (e.g., see [18]).

Moran eigenvector spatial filtering (MESF) is a relatively novel methodology to handle SA contained in data [19–21]. MESF uses a set of synthetic proxy variables, which are extracted as eigenvectors from a doubly centered n-by-n spatial weights matrix (SWM), say **C**, that often is binary (0–1) and that ties n geographic objects together in space (indicating which are pairwise directly correlated by an entry of one), and then adds these vectors as control variables to a regression model specification; SWMs are analogous to undirected planar graph adjacency matrices. These control variables identify and isolate stochastic spatial dependencies among a set of georeferenced observations, filtering these dependencies out of a model's residuals and adding them to the model's mean response, thus allowing regression model building to proceed with georeferenced observations that mimic being independent.

The Moran Coefficient (MC; [22]) SA index, one of several such indices that exist—e.g., the Geary Ratio (GR; [23])—furnishes the basis for MESF; its numerator contains the doubly centered SWM exploited by MESF. This index can be written in matrix form as follows:

$$(n/\mathbf{1}^T\,\mathbf{C1})\mathbf{Y}^T(\mathbf{I}-\mathbf{11}^T/n)\mathbf{C}(\mathbf{I}-\mathbf{11}^T/n)\mathbf{Y}/\mathbf{Y}^T(\mathbf{I}-\mathbf{11}^T/n)\mathbf{Y}, \quad (4)$$

where n is the number of, for example, polygons in a geographic information system shapefile; **Y** is an n-by-1 vector of attribute values; **I** is an n-by-n identity matrix; **1** is an n-by-1 vector of ones; superscript T denotes matrix transpose and, here, **C** is a binary 0–1 SWM for which 1 denotes that the row and column areal units (i.e., locations, e.g., shapefile polygons) share a non-zero length common boundary, and 0 denotes that they do not (i.e., the rook adjacency definition, phraseology based upon chess game pieces and their moves). The eigenfunctions (i.e., the paired eigenvalues and eigenvectors) are extracted from the following matrix, which is doubly centered because matrix **C** is pre- and post-multiplied by the projection matrix $(\mathbf{I}-\mathbf{11}^T/n)$, appearing in the numerator of expression (4) as follows:

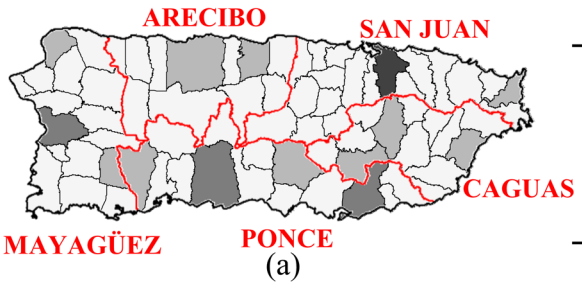
$$(\mathbf{I}-\mathbf{11}^T/n)\mathbf{C}(\mathbf{I}-\mathbf{11}^T/n). \quad (5)$$

When multiplied by $(n/\mathbf{1}^T\mathbf{C1})$, an eigenvalue of matrix (5) is converted to the MC measuring the SA in its associated eigenvector map pattern of real number elements [24,25]. The sign of an eigenvalue indicates the nature of SA represented by its corresponding eigenvector, whereas its magnitude indicates the degree of SA. The extreme eigenvalues of matrix expression (4) determine the limits of MC, which are not necessarily ± 1 [26].

Extracting the eigenfunctions from adjusted SWMs constitutes a spectral decomposition of those matrices. The extracted eigenvectors with associated eigenvalues relatively far from zero, and hence representing other-than-negligible SA, may be viewed as portraying global, regional, or local components of SA because of the particular map patterns they depict when visualized [27,28]. In other words, SA manifests itself in terms of similar (positive SA; PSA) or dissimilar (negative SA; NSA) values of variable Y clustering on a map: hot spots (i.e., similar relatively high values clustering), cold spots (i.e., similar relatively low values clustering), or contrasts (i.e., relatively dissimilar values clustering).

3.2.4. An Empirical 2010 Puerto Rico Population Density Toy Illustration

Puerto Rico is partitioned into five agricultural administrative regions, as a scheme applied to it for a century or more, each focusing on one of the island's major urban areas (Figure 3). The 1899 geographic distribution of population density for this spatial resolution displays excessive extra-Poisson variation (deviance = 1723, which is far greater than its expected value of 1) and moderate PSA. The table in Figure 3b reports the adjusted SWM eigenvector representing this PSA [MC = 0.221, with its maximum (MC_{max}) being 1.094, and GR = 0.474]. This scenario acknowledges a well-recognized source of over-dispersion, namely correlation among observations (i.e., dependent data, autocorrelation), which is frequent in spatial statistics; for example, SA accounts for roughly half of any detected over-dispersion (e.g., extra-Poisson variation).



(a)

region	SJ	A	M	P	C	eigenvector
SJ: San Juan	0	1	0	1	1	0.37175
A: Arecibo	1	0	1	1	0	−0.37175
M: Mayagüez	0	1	0	1	0	−0.60150
P: Ponce	1	1	1	0	1	0.00000
C: Caguas	1	0	0	1	0	0.60150

(b)

Figure 3. The 1899 Puerto Rico agricultural administrative regions (AAR; denoted by red borders). Black lines denote municipio boundaries. Tertile choropleth map: black, gray, and white, respectively, denote relatively high, intermediate, and relatively low population density. Left (**a**): the municipality-AAR choropleth map. Right (**b**): the AAR spatial weights matrix and its relevant eigenvector.

An NB2 description of this geographic distribution of population density yields the global MLE of $\widehat{\mu} = e^{5.6361} \approx 280.4$ people per square mile, with a dispersion parameter MLE of $\widehat{\alpha}^{-1} \approx 0.0010$ (i.e., nearly a Poisson RV). The eigenvector spatial filter (ESF) specification, a linear combination of a subset of the eigenvectors extracted from the adjusted SWM, describes a spatially varying mean population density by region [i.e., 283.6 (SJ), 278.6 (A), 295.7 (M), 271.6 (P), and 272.4 (C)] about a global mean of 280.3 and reduces the over-dispersion parameter to 0.0006 (even more closely resembling a Poisson RV).

The global MME of $\widehat{\mu}$ also is $e^{5.6361} \approx 280.4$. However, the dispersion parameter MME of $\widehat{\alpha}^{-1}$ is –0.0001, which would result in its being rounded to 0. Finally, the spatially varying mean population density by region is as follows: 280.1 (SJ), 280.6 (A), 280.8 (M), 280.4 (P), and 280.0 (C). These outcomes seem inferior to their preceding MLE counterparts.

4. Discussion: Two MESF Empirical Examples

The preceding examples in Sections 3.2.2 and 3.2.4 are only for exemplification purposes. This section presents two empirical examples whose results furnish informative substantive implications.

4.1. An Empirical 2010 Puerto Rico Urban Population Density Case Study

Figure 4a portrays the 2010 geographic distribution of urban population density (people per square mile) across the main island of Puerto Rico. The map pattern characterizing this geographic distribution has an MC of 0.487 and a GR of 0.541, implying moderate PSA. From the MLEs for an NB2 model specification, the global density is $e^{7.0549} \approx 1158.5$, with a dispersion parameter of 0.6707 (i.e., sizeable extra-Poisson variation). Historically, the population has been concentrated on the coastal lowlands of the island, the preferred locations of the initial Spanish settlers because of convenience and accessibility, furnishing a rationale for including mean elevation as a covariate in an NB2 regression. Primarily because of the way urban areas expand (e.g., the establishment of infrastructure at the rural–urban fringe), this variable should contain PSA. In this case, executing NB2 stepwise regression estimation results in the selection of five PSA and no NSA eigenvectors from the candidate set extracted from the adjusted SWM for which $MC_j/MC_{max} \geq 0.25$, with subscript j denoting the jth eigenvalue. Table 2 summarizes the MLE results. It reveals that SA accounts for roughly 60% of the over-dispersion in these data, and the single covariate coupled with the constructed ESF accounts for more than three-fourths of the spatial variation in urban population density across the main island of Puerto Rico. With a final $\widehat{\alpha}^{-1} < 0.2$, a dramatic shrinkage from nearly 0.7, p = 2 seems adequate for the NB model specification.

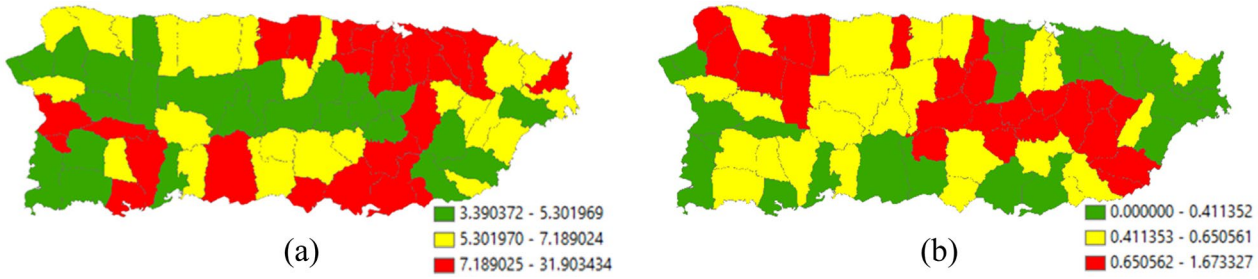


Figure 4. Municipio geographic resolution 2010 population density (i.e., people per square mile) across the main island of Puerto Rico; green, yellow, and red, respectively, denote relatively low, intermediate, and high density values. Left (**a**): urban population (extra-Poisson variation is 41,183 >> 1). Right (**b**): rural population (extra-Poisson variation is 1427 >> 1).

Table 2. Summary of model estimation results.

Model Specification	NB2 Urban Population Density		Rural Population Density			
			NB2		NB3	
	$\widehat{\alpha}^{-1}$	Pseudo–R^2	$\widehat{\alpha}^{-1}$	Pseudo–R^2	$\widehat{\alpha}^{-1}$	Pseudo–R^2
Intercept only	0.6707	0	1.8760	0	1.8285	0
Single covariate	0.4848	0.18	1.3964	0.48	1.1847	0.48
Covariate + ESF	0.2021	0.79	1.0399	0.52	1.0065	0.65

NOTE: pseudo—R^2 (analogous to Efron's) denotes the R^2 from a bivariate linear regression of the observed on the NBp-predicted population density (e.g. [29–32]).

4.2. An Empirical 2010 Puerto Rico Rural Population Density Illustration

Figure 4b portrays the 2010 geographic distribution of rural population density across the main island of Puerto Rico. The map pattern characterizing this geographic distribution has an MC of 0.496 and a GR of 0.511, implying moderate PSA. From the MLEs for an NB2 model specification, the global density is $e^{4.1298} \approx 62.2$, with a dispersion parameter of 1.8760. Rural and urban land are mutually exclusive, furnishing a rationale for including the density of rural land as a covariate. Because rural land tends to concentrate outside of urban areas, suggesting it contains PSA, and primarily because the expansion of urban land means the contraction of rural land, particularly along the rural–urban fringe, with this spatial competition suggesting it contains NSA, this variable should contain a PSA-NSA mixture. In this case, the stepwise NB2 estimation results in the selection of five eigenvectors, four with PSA and one with NSA (i.e., a mixture). Table 2 summarizes the MLE results. It reveals that SA accounts for roughly 26% of the over-dispersion in these data, and the single covariate coupled with the constructed ESF accounts for more than half of the spatial variation in rural population density across the main island of Puerto Rico. With a final $\widehat{\alpha}^{-1} > 1$, p = 2 may seem inadequate for the NB model specification. Setting p = 3—with estimation executed via SAS PROC NLMIXED (see Appendix B), in which the NB3 probability density log-likelihood function was written expressing the modified NB as a general distribution [see https://stats.oarc.ucla.edu/sas/faq/how-can-i-compute-negative-binomial-models-with-random-intercepts-and-slopes-using-nlmixed/ (last accessed on 20 March 2024)], alters $\widehat{\alpha}^{-1}$, which remains greater than one, while increasing the pseudo-R^2 values when an ESF is present; these latter values decrease when p = 4, implying that p = 3 is the adequate value.

4.3. An Empirical 2010 Puerto Rico Rural Population Den

The two population density geographic distributions have markedly different island-wide averages and extra-Poisson variation. Although they have very similar SA index values, their ESFs indicate that they have very different latent map patterns (e.g., one is characterized by PSA, whereas the other is characterized by a mixture of PSA and

NSA). A population classified as urban is not rural, and vice versa, which implies a negative relationship between these two geographic distributions. The bivariate Poisson RV correlation between them is –0.62 (see [33]). Finally, urban population density seems to be well described by an NB2 model, whereas rural population density may benefit from being described by an NB3 model.

5. Concluding Comments and Implications

This paper contributes to the academic literature in a number of different ways. First, it clarifies/corrects several inaccurate technical details in Greene's earlier paper. Second, it presents new results that include the mgf for the NBp model, two new theorems, and original empirical spatial statistical/econometric applications. Third, it exemplifies selected small and large sample properties on NBp models. Finally, yet again, it demonstrates the power of MESF.

One prominent implication is that quantitative analysts, and especially applied spatial statisticians, studying Poisson and/or NB variables need to expand their toolbox to include NBp specifications for which p > 2. Another is the need for a wider appreciation and adoption of MESF-type methodology for correlated data, including for traditional dependent datasets, time series, and contemporary social network projects. A third repercussion of this paper is a recognition that the NBp random variable needs a Wikipedia page paralleling the existing one for the NB (https://en.wikipedia.org/wiki/Negative_binomial_distribution; last accessed on 15 May 2024).

Perhaps the most serious NBp limitation is its lack of closed-form MLEs. Although this variable drawback is not new to statistics, it does constitute an impediment to how the utilization of a variable flourishes. This certainly is a theme that should motivate subsequent research about the NBp. Another of its limitations is its newness, which means establishing strengths and weaknesses of employing it vis à vis alternative RVs is yet to be determined.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data employed in this paper are readily available from the United States Census Bureau (i.e., the Puerto Rico data) or are obtainable (with sampling errors) by executing a suitable pseudo-random number generator (this paper used SAS implementations).

Acknowledgments: The author is an Ashbel Smith Chaired Professor of Geospatial Information Sciences.

Conflicts of Interest: The author declares no conflicts of interest.

Appendix A. Estimation of iid NBp Parameters

MLE

$$\mathbf{L}(\alpha,\mu|\mathbf{p}) = \prod_{\mathbf{i}=1}^{\mathbf{n}} \binom{\mathbf{y_i} + \alpha\mu^{2-\mathbf{p}} - 1}{\alpha\mu^{2-\mathbf{p}} - 1} \left(\frac{1}{1+\mu^{\mathbf{p}-1}/\alpha}\right)^{\alpha\mu^{2-\mathbf{p}}} \left(1 - \frac{1}{1+\mu^{\mathbf{p}-1}/\alpha}\right)^{\mathbf{y_i}}$$

$$\mathbf{LN}\left[\mathbf{L}(\alpha,\mu|\mathbf{p})\right] = \sum_{\mathbf{i}=1}^{\mathbf{n}} \mathbf{LN}\left[\Gamma\left(\mathbf{y_i} + \alpha\mu^{2-\mathbf{p}}\right)\right] - \sum_{\mathbf{i}=1}^{\mathbf{n}} \mathbf{LN}\left[\Gamma(\mathbf{y_i}+1)\right] - \mathbf{nLN}\left[\Gamma\left(\alpha\mu^{2-\mathbf{p}}\right)\right] +$$
$$\left(\mathbf{n}\mu^{2-\mathbf{p}}/\alpha\right)\mathbf{LN}\left(\frac{1}{1+\mu^{\mathbf{p}-1}/\alpha}\right) + \sum_{\mathbf{i}=1}^{\mathbf{n}} \mathbf{y_i}\mathbf{LN}\left(\frac{\mu^{\mathbf{p}-1}/\alpha}{1+\mu^{\mathbf{p}-1}/\alpha}\right)$$

Both $\partial\mathbf{LN}\left[\mathbf{L}(\alpha,\mu|\mathbf{p})\right]/\partial\mu$ and $\partial\mathbf{LN}\left[\mathbf{L}(\alpha,\mu|\mathbf{p})\right]/\partial\alpha$ produce the polygamma function (e.g., see [34]), involving the derivative of the logarithm of the gamma function, which prevents

algebraically formulating closed-form solutions. In addition, the relationship among α, μ, and p determines whether or not an MLE exists.

MME (using the unbiased variance estimator)

$$\widehat{\mu} = \frac{\sum_{i=1}^{n} y_i}{n}$$

$$\mu\,(1+\mu/\alpha) = \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1} \Rightarrow \widehat{\alpha} = \frac{\left(\sum_{i=1}^{n} y_i/n\right)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2/(n-1) - \sum_{i=1}^{n} y_i/n}$$

Appendix B. SAS PROC NLMIXED Code

The following SAS code tricks its procedure into estimating a BN3 specification by not including a random effects term. The following input dataset (**STEP3**) contains the needed adjusted SWM eigenvectors, the response (i.e., population counts) and offset (i.e., log-area) variables, and the covariate (i.e., the logarithm of rural land area percentage; to avoid undefined terms for zero rural land, zero was replaced with one, which is trivial because the smallest non-zero measure is 9703):

```
PROC NLMIXED DATA = STEP3;
PARMS B0 = 0 B1 = 0 B2 = 0 B3 = 0 B4 = 0 B5 = 0 B6 = 0 A = 1.4;
P = 3;
*XB = B0 + OFFSET;
*XB = B0 + B1*PCTRL + OFFSET;
XB = B0 + B1*PCTRL + B2*COL10 + B3*COL12 + B4*COL15 + B5*COL20
+B6*COL44 + OFFSET;
MU = EXP(XB);
M = 1/A;
LL = LGAMMA(Y + M) — LGAMMA(Y + 1) - LGAMMA(M)
+Y*LOG(A*MU**(P - 1)) - (Y + M)*LOG(1 + A*MU**(P - 1));
MODEL Y ~ GENERAL(LL);
PREDICT MU OUT = MU;
RUN;
```

XB differentiates between a constant mean, a bivariate regression with the rural area percentage covariate, and an MESF regression that includes this preceding covariate. **P = 3** allows the estimation of NB3; setting **P** to 2 estimates NB2, and setting it to 1 estimates NB1 (i.e., a Poisson). Finally, the temporary SAS file **WORK.MU** stores predicted counts for post-processing. The NB1 and NB2 options support comparative output checks with other software modules.

References

1. Cameron, C.; Trivedi, P. *Regression Analysis of Count Data*; Cambridge University Press: New York, NY, USA, 1998.
2. Barreto-Souza, W.; Ombao, H. The negative binomial process: A tractable model with composite likelihood-based inference. *Scand. J. Stat.* **2021**, *49*, 568–592. [CrossRef]
3. Greene, W. Functional forms for the negative binomial model for count data. *Econ. Lett.* **2008**, *99*, 585–590. [CrossRef]
4. Di, Y.; Schafer, D.W.; Cumbie, J.S.; Chang, J.H. The NBP Negative Binomial Model for Assessing Differential Gene Expression from RNA-Seq. *Stat. Appl. Genet. Mol. Biol.* **2011**, *10*, 1–28. [CrossRef]
5. Zhang, Y.; Ye, Z.; Lord, D. Estimating dispersion parameter of negative binomial distribution for analysis of crash data: Bootstrapped maximum likelihood method. *Transp. Res. Rec. J. Transp. Res. Board* **2019**, *2019*, 15–21. [CrossRef]
6. Pearson, K. Method of Moments and Method of Maximum Likelihood. *Biometrika* **1936**, *28*, 34–59. [CrossRef]
7. Hendry, D.; Nielsen, B. *Econometric Modeling: A Likelihood Approach*; Princeton University Press: Princeton, NJ, USA, 2007.
8. Chambers, R.L.; Steel, D.G.; Wang, S.; Welsh, A. *Maximum Likelihood Estimation for Sample Surveys*; CRC Press: Boca Raton, FL, USA, 2012.
9. Rossi, R. *Mathematical Statistics: An Introduction to Likelihood Based Inference*; Wiley: New York, NY, USA, 2018.
10. Ward, M.; Ahlquist, J. *Maximum Likelihood for Social Science: Strategies for Analysis*; Cambridge University Press: New York, NY, USA, 2018.

11. Hoeffding, W. The Large-Sample Power of Tests Based on Permutations of Observations. *Ann. Math. Stat.* **1952**, *23*, 169–192. [CrossRef]
12. Bagui, S.; Mehra, K. On the convergence of negative binomial distribution. *Am. J. Math. Stat.* **2019**, *9*, 44–50.
13. Bowman, K.; Shenton, L. *Estimator: Method of moments, Encyclopedia of Statistical Sciences, vol. 3 (D'Agostino Test of Normality to Eye Estimate)*, 2nd ed.; Kotz, S., Read, C., Balakrishnan, N., Vidakovic, B., Johnson, N., Eds.; Wiley: New York, NY, USA, 2006; pp. 2092–2098.
14. van de Geer, S. A New Approach to Least-Squares Estimation, with Applications. *Ann. Stat.* **1987**, *15*, 587–602. [CrossRef]
15. van de Geer, S. Least squares estimation. In *Encyclopedia of Statistics in Behavioral Science*; Everitt, B., Howell, D., Eds.; Wiley: Chichester, UK, 2005; Volume 2, pp. 1041–1045. Available online: https://people.math.ethz.ch/~geer/bsa199_o.pdf (accessed on 1 March 2024).
16. Wolberg, J. *Data Analysis Using the Method of Least Squares: Extracting the Most Information from Experiments*; Springer: Berlin, Germany, 2005.
17. Delicado, P.; Goria, M.N. A small sample comparison of maximum likelihood, moments and L-moments methods for the asymmetric exponential power distribution. *Comput. Stat. Data Anal.* **2008**, *52*, 1661–1673. [CrossRef]
18. Griffith, D.A. A Family of Correlated Observations: From Independent to Strongly Interrelated Ones. *Stats* **2020**, *3*, 166–184. [CrossRef]
19. Griffith, D. Spatial filtering. In *Encyclopedia of Geographic Information Science*; Kemp, K., Ed.; SAGE: Thousand Oaks, CA, USA, 2008; pp. 413–415.
20. Griffith, D.; Getis, A. Spatial filtering. In *Encyclopedia of GIS*, 2nd ed.; Shekhar, S., Xiong, H., Zhou, X., Eds.; Springer: Cham, Switzerland, 2016; pp. 1–14. [CrossRef]
21. Griffith, D.; Chun, Y. Spatial Autocorrelation and Spatial Filtering. In *Handbook of Regional Science*, 2nd ed.; Fischer, M., Nijkamp, P., Eds.; Springer-Verlag: Berlin, Germany, 2022; pp. 1863–1892, (revised, update to 2014 version).
22. Moran, P. The interpretation of statistical maps. *J. R. Stat. Soc. B* **1948**, *10*, 243–251. [CrossRef]
23. Geary, R.C. The Contiguity Ratio and Statistical Mapping. *Incorp. Stat.* **1954**, *5*, 115–127+129–146. [CrossRef]
24. Tiefelsdorf, M.; Boots, B. The exact distribution of Moran's I. *Environ. Plan. A* **1995**, *27*, 985–999. [CrossRef]
25. Griffith, D.A. Spatial autocorrelation and eigenfunctions of the geographic weights matrix accompanying geo-referenced data. *Can. Geogr. Can.* **1996**, *40*, 351–367. [CrossRef]
26. de Jong, P.; Sprenger, C.; van Veen, F. On extreme values of Moran's I and Geary's c. *Geogr. Anal.* **1984**, *16*, 17–24. [CrossRef]
27. Borcard, D.; Legendre, P. All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. *Ecol. Model.* **2002**, *153*, 51–68. [CrossRef]
28. Borcard, D.; Legendre, P.; Avois-Jacquet, C.; Tuomisto, H. Dissecting the spatial structure of ecological data at multiple scales. *Ecology* **2004**, *85*, 1826–1832. [CrossRef]
29. Cameron, A.C.; Windmeijer, F.A.G. An R-squared measure of goodness of fit for some common nonlinear regression models. *J. Econ.* **1997**, *77*, 329–342. [CrossRef]
30. Christensen, R. *General Prediction Theory and the Role of R2, Unpublished Manuscript*; Department of Mathematics and Statistics, University of New Mexico: Albuquerque, NM, USA, 2007; Available online: https://www.math.unm.edu/~fletcher/JPG/rsq.pdf (accessed on 1 March 2024).
31. Hoetker, G. The use of logit and probit models in strategic management research: Critical issues. *Strat. Manag. J.* **2007**, *28*, 331–343. [CrossRef]
32. Cakmakyapan, S.; Demirhan, H. A Monte Carlo-based pseudo-coefficient of determination for generalized linear models with binary outcome. *J. Appl. Stat.* **2016**, *44*, 2458–2482. [CrossRef]
33. Kawamura, K. The structure of bivariate Poisson distribution. *Kodai Math. J.* **1973**, *25*, 246–256. [CrossRef]
34. Mortici, C. Very accurate estimates of the polygamma functions. *Asymptot. Anal.* **2010**, *68*, 125–134. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

AppliedMath

MDPI

Article

Relationship between Inverse Langevin Function and r_0-r_1-Lambert W Function

Roy M. Howard

School of Electrical Engineering, Computing and Mathematical Sciences, Curtin University, Perth, WA 6845, Australia; r.howard@curtin.edu.au

Abstract: The relationship between the inverse Langevin function and the proposed r_0-r_1-Lambert W function is defined. The derived relationship leads to new approximations for the inverse Langevin function with lower relative error bounds than comparable published approximations. High accuracy approximations, based on Schröder's root approximations of the first kind, are detailed. Several applications are detailed.

Keywords: inverse Langevin function; Lambert W function; Schröder approximation

1. Introduction

The Langevin function arises in diverse contexts including the classical model of paramagnetism [1], and the ideal freely jointed chain model, e.g., [2]. It is defined, e.g., [3], according to:

$$\begin{aligned} y = L(x) &= \begin{cases} \coth(x) - \frac{1}{x}, & x \in (0, \infty) \\ 0, & x = 0 \end{cases} \\ L(-x) &= -L(x) \end{aligned} \tag{1}$$

Whilst the Langevin function is well defined, its inverse, denoted L^{-1}, does not have a known analytical form and significant research efforts have led to many approximations with the papers, for example, and in chronological order, by Itskov (2012) [4], Nguessong (2014) [5], Darabi (2015) [6], Jedynak (2015) [7], Kröger (2015) [8], Marchi (2015) [9], Rickaby (2015) [10], Petrosyan (2017) [11], Jedynak (2018) [12], Marchi (2019) [13], and Howard (2020) [14], providing useful approximations. Graphs of the Langevin function and the inverse Langevin function are shown in Figure 1.

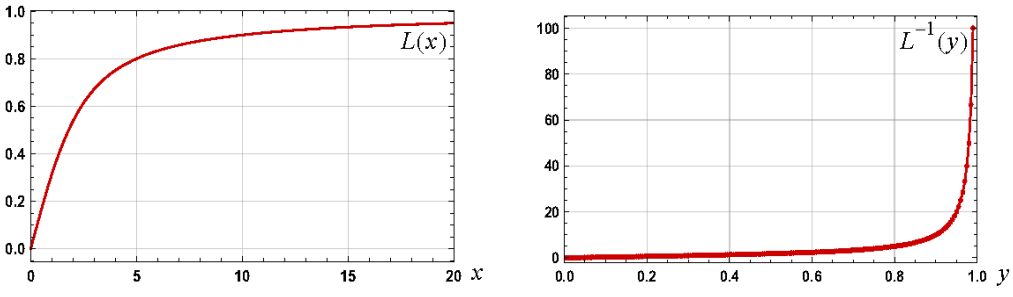


Figure 1. Graphs of the Langevin and inverse Langevin functions.

Mező and Keady [15] detail the relationship between the inverse Langevin function and a generalization of the Lambert W function. Whilst such a relationship is important, it does not directly facilitate finding approximations for the inverse Langevin function as the generalized Lambert W function does not have an explicit analytical form. The contribution of this paper is to detail a relationship between the inverse Langevin function and the r_0-r_1-Lambert W function which is a generalization of the r-Lambert W function

check for updates

Citation: Howard, R.M. Relationship between Inverse Langevin Function and r_0-r_1-Lambert W Function. *AppliedMath* **2024**, *4*, 743–762. https://doi.org/10.3390/appliedmath4020040

Academic Editor: Valery Karachik

Received: 25 March 2024
Revised: 18 May 2024
Accepted: 21 May 2024
Published: 14 June 2024

Copyright: © 2024 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

that has been considered by Mező [16], Mező and Baricz [17], and Jamilla [18]. The derived relationship leads to new approximations for the inverse Langevin function with potentially lower relative error bounds than comparable approximations. Improved approximations, based on Schröder's root approximations of the first kind, e.g., [19–21], are detailed. Several applications of the results are also detailed.

Section 2 details the definitions that underpin the paper. Section 3 details the relationship between the inverse Langevin function and the proposed r_0-r_1-Lambert W function. Section 4 details approximations for the r_0-r_1-Lambert W function and, thus, the inverse Langevin function. Section 5 details the establishment of approximations of higher accuracy through the use of Schröder's root approximations of the first kind. Several applications of the results are detailed in Section 6. A summary of results, and conclusions, are detailed in Section 7.

2. Definitions

By utilizing the exponential function-based definition for the hyperbolic cotangent function, the following alternative definition for the Langevin function is valid:

$$y = L(x) = \frac{x-1+(1+x)e^{-2x}}{x(1-e^{-2x})}, \quad x > 0, y \in (0,1). \tag{2}$$

It then follows that the relationship between x and $y = L(x)$ is equivalently defined by either of the equations:

$$e^{-2x} = \frac{1-x(1-y)}{1+x(1-y)} \tag{3}$$

$$f(x,y) = 1-(1-y)x-[1+(1+y)x]e^{-2x}, \quad f(x,y) = 0, \tag{4}$$

For $x > 0$, $y \in (0,1)$. The graph of the function f is shown in Figure 2. This function underpins a link between the inverse Langevin function and a generalization of the Lambert W function. The following definitions are fundamental to establishing the required link.

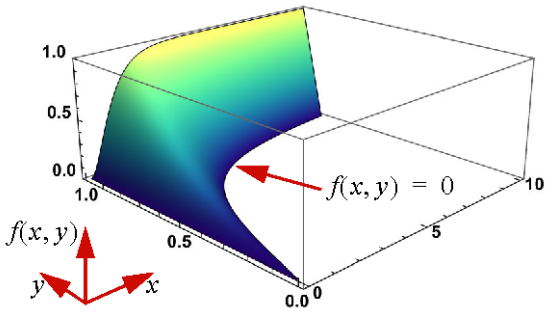


Figure 2. Graph of $f(x, y)$.

Definition 1. *Lambert W Function.*

The Lambert W function, denoted W, is the inverse of h, i.e., $h^{-1}(z) = W(z)$, where h is defined according to:

$$z = h(w) = we^w. \tag{5}$$

For the general case, $w \in \boldsymbol{C}$, the Lambert W function is multivalued, e.g., [22]. For the real case, $w \in (-\infty, \infty)$, the principal branch of the Lambert W function, denoted W_0, is associated with the interval $w \in [-1, \infty)$ whilst the negative one branch, denoted W_{-1}, is associated with the interval $w \in (-\infty, -1)$.

Definition 2. *Generalized Lambert W Function.*

The generalized Lambert *W* function, e.g., [15], is denoted

$$W\left[\begin{matrix}\alpha_1,\cdots,\alpha_n\\ \beta_1,\cdots,\beta_n\end{matrix}\middle| z\right] \tag{6}$$

and is the inverse of the function h which is defined according to

$$z = h(w) = e^w.\frac{(w-\alpha_1)\cdots(w-\alpha_n)}{(w-\beta_1)\cdots(w-\beta_n)} \tag{7}$$

For the general case $w, \alpha_1, \cdots, \alpha_n, \beta_1, \cdots, \beta_n \in \boldsymbol{C}$ but often the real case is of interest.

Definition 3. *r-Lambert W Function.*

The *r*-Lambert *W* function, e.g., [16–18], denoted h_r^{-1} or W_r, is the inverse of h_r defined according to

$$h_r(w) = rw + we^w. \tag{8}$$

For the general case $w, r \in \boldsymbol{C}$. A generalization of this function is the r_0-r_1-Lambert *W* function.

Definition 4. *r_0-r_1-Lambert W Function.*

The r_0-r_1-Lambert *W* function, denoted h_{r_0,r_1}^{-1} or W_{r_0,r_1}, is the inverse of h_{r_0,r_1} which is defined according to:

$$h_{r_0,r_1}(w) = r_0 + r_1 w + we^w. \tag{9}$$

For the general case $w, r_0, r_1 \in \boldsymbol{C}$. The real case is considered in this paper.

2.1. Link: Inverse Langevin Function and Generalized Lambert W Function

Manipulation of (3) (see [15] Section 2.2.1) leads to the relationship

$$L^{-1}(y) = -2W\left[\begin{matrix}\frac{2}{y+1}\\ \frac{2}{y-1}\end{matrix}\middle| \frac{y-1}{y+1}\right] \tag{10}$$

which directly links the inverse Langevin function to the general Lambert *W* function. Whilst such a link is important, it does not directly facilitate finding approximations for the inverse Langevin function, as an explicit analytical approximation for the Lambert *W* function is unknown and research to develop approximations is ongoing, e.g., [23,24]. The relationship between the inverse Langevin function and the r_0-r_1-Lambert *W* function, which yields new approximations, is detailed below after the next subsection, which lists established approximations for the inverse Langevin function.

2.2. Published Approximations

Representative approximations for the inverse Langevin function include:

$$L_{0,1}^{-1}(y) = \frac{3y}{1-y}.\left[1 - \frac{24y}{25} + \frac{22y^2}{75}\right] \tag{11}$$

$$L_{0,2}^{-1}(y) = \frac{3y}{1-y^2}.\left[1 - \frac{2y^2}{5} - \frac{y^4}{15} + \frac{2y^6}{15}\right] \tag{12}$$

$$L_{0,3}^{-1}(y) = 3y + \frac{y^2}{5}\sin\left[\frac{7y}{2}\right] + \frac{y^3}{1-y} \tag{13}$$

$$L_{0,4}^{-1}(y) = \frac{y(3-y^2)}{1-y^2} - \frac{y^{10/3}}{2} + 3y^5\left[y - \frac{76}{100}\right](y-1) \tag{14}$$

$$\begin{aligned} L_{0,5}^{-1}(y) &= \frac{3y+a_2y^2+a_3y^3+a_4y^4}{1-y+b_2(y-y^2)+b_3(y^2-y^3)+b_4(y^3-y^4)}, \quad [4/4] \text{ approx.} \\ a_2 &= -6.98408968, \quad a_3 = 5.69026957, \quad a_4 = -1.35415696, \\ b_2 &= -1.33411915, \quad b_3 = 0.03919556, \quad b_4 = 0.64694651, \end{aligned} \tag{15}$$

which, respectively, have been published in [5,8,11,13,14]. The relative error bounds (see definition below) that are associated with these approximations over the interval $[0,1)$, respectively, are: 9.69×10^{-3}, 2.75×10^{-3}, 1.79×10^{-3}, 7.22×10^{-4} and 1.80×10^{-4}.

2.2.1. Relative Error Bound

For an arbitrary function f, defined over the interval $[\alpha, \beta]$, an approximating function f_A has a relative error, at a point x_1, defined according to $\text{re}(x_1) = 1 - f_A(x_1)/f(x_1)$. The relative error bound for the approximating function, over the interval $[\alpha, \beta]$, is defined according to:

$$\text{re}_B = \max\{|\text{re}(x_1)| : \ x_1 \in [\alpha, \beta]\} \tag{16}$$

In this paper, Mathematica has been used to facilitate analysis and to obtain numerical results. In general, relative error results associated with approximations have been obtained by uniformly sampling the interval $[0,1]$ with 1000 points.

3. Relationship between Inverse Langevin and r_0-r_1-Lambert W Function

With the function f defined according to (4), it is the case that $f(x,y) = 0$ when $x = L^{-1}(y)$. The following result then holds:

Theorem 1. *Relationship: r_0-r_1-Lambert W Function and Inverse Langevin Function*

With h_{r_0,r_1} and f defined, respectively, by (9) and (4), the general result

$$h_{r_0(y),r_1(y)}(w) = r_0(y) + r_1(y)w + we^w = k_f(y) f\left[\tfrac{-1}{1+y} - \tfrac{w}{2}, y\right], \quad y \neq -1, \tag{17}$$

holds for the case of $x, y \in \boldsymbol{R}$ when

$$r_0(y) = \tfrac{4e^{-2/(1+y)}}{(1+y)^2}, \quad r_1(y) = \tfrac{(1-y)e^{-2/(1+y)}}{1+y}, \quad k_f(y) = \tfrac{2e^{-2/(1+y)}}{1+y}, \quad y \neq -1. \tag{18}$$

It then follows that

$$h_{r_0(y),r_1(y)}(w) = \tfrac{4e^{-2/(1+y)}}{(1+y)^2} + \tfrac{w(1-y)e^{-2/(1+y)}}{1+y} + we^w, \quad y \neq -1, \tag{19}$$

and the inverse Langevin function can be defined according to

$$L^{-1}(y) = \tfrac{-1}{1+y} - \tfrac{h^{-1}_{r_0(y),r_1(y)}(0)}{2}, \quad 0 \leq y < 1. \tag{20}$$

Proof of Theorem 1. The proof is detailed in Appendix A. □

3.1. Determining Root of h

Consistent with (20), the inverse Langevin function can be defined once the appropriate root of $h_{r_0(y),r_1(y)}$ is known. It can readily be shown that

$$h_{r_0(y),r_1(y)}\left[\tfrac{-2}{1+y}\right] = 0, \quad y \neq -1, \tag{21}$$

which implies $h^{-1}_{r_0(y),r_1(y)}(0) = \frac{-2}{1+y}$. However, this result is associated with a branch of h^{-1} which is not consistent with the validity of the relationship specified by (20). The branch of h^{-1} that is required can readily be seen by considering the nature of h for the case of $0 < y < 1$.

3.1.1. Graphs

For the case of $0 \leq y < 1$, it is the case that $1/e < r_0(y) \leq 4/e^2$ and $0 < r_1(y) \leq 1/e^2$. Graphs of $r_0(y)$ and $r_1(y)$, for the case of $y \in [0,1]$, are shown in Figure 3. The graph of $h_{r_0(y),r_1(y)}(w)$ is shown in Figure 4 for two set values of y to illustrate the nature of the variation of $h_{r_0(y),r_1(y)}(w)$ for set values of y. The variation, with y, of the root of $h_{r_0(y),r_1(y)}(w)$, that is to the left of $-2/(1+y)$, is shown in Figure 5.

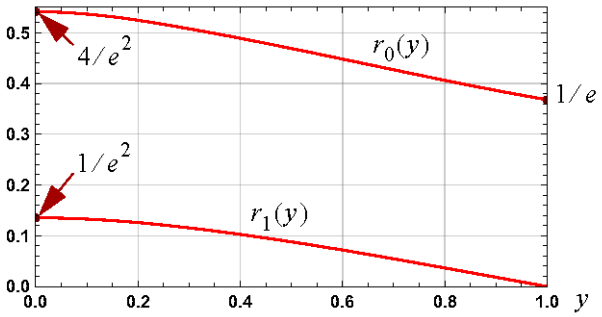


Figure 3. Graphs of $r_0(y)$ and $r_1(y)$ for the case of $y \in (0,1)$.

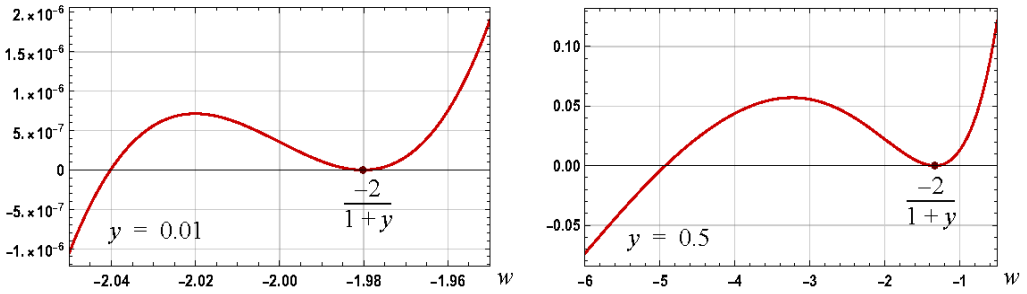


Figure 4. Graphs of $h_{r_0(y),r_1(y)}(w)$ for the set values of y of 0.01 and 0.5.

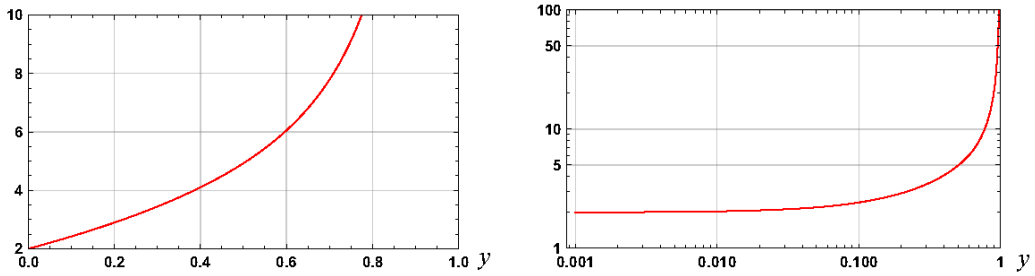


Figure 5. Graphs of $-h^{-1}_{r_0(y),r_1(y)}(0)$, $y \in (0,1)$, which is the negative of the root of $h_{r_0(y),r_1(y)}(w)$ that is to the left of $-2/(1+y)$.

3.1.2. Required Branch for Root of h

The variation of $h_{r_0(y),r_1(y)}(w)$, with w, is illustrated in Figure 4 along with the first negative root of $w = -2/(1+y)$ as defined by (21). The required root is to the left of the local maximum of $h_{r_0(y),r_1(y)}(w)$ that is located left of the first negative root. To establish where this local maximum occurs, consider the derivative of $h_{r_0(y),r_1(y)}(w)$:

$$h^{(1)}_{r_0(y),r_1(y)}(w) = r_1(y) + (1+w)e^w. \tag{22}$$

The root of this function, denoted w_m, occurs when $(1+w_m)e^{w_m} = -r_1(y)$, or, after a change of variable $w_1 = 1 + w_m$, where $w_1 e^{w_1} = -er_1(y)$. As $0 < r_1(y) \leq 1/e^2$, for $0 \leq y < 1$ (see Figure 3), it is the case that

$$w_1 e^{w_1} = \begin{cases} -1/e, & y = 0, r_1(y) = 1/e^2 \\ -er_1(y), & 0 < y < 1,\ 0 < r_1(y) < 1/e^2 \end{cases} \tag{23}$$

with solutions, consistent with the nature of the Lambert W function, e.g., [22], defined according to

$$w_1 = \begin{cases} -1, & y = 0 \\ W_0[-er_1(y)] \quad \text{or} \quad W_{-1}[-er_1(y)]\ , & 0 < y < 1 \end{cases} \tag{24}$$

where W_0 is the principal branch solution and W_{-1} is the negative one branch. It is the negative one branch that provides the required solution and, thus:

$$w_m = \begin{cases} -2, & y = 0 \\ -1 + W_{-1}[-er_1(y)], & 0 < y < 1 \end{cases} \tag{25}$$

Hence, the goal is to establish the root of $h_{r_0(y),r_1(y)}(w)$ for the case of $w \leq w_m$. The nature of $z = h_{r_0(y),r_1(y)}(w)$, and its inverse $w = h^{-1}_{r_0(y),r_1(y)}(z)$, are illustrated in Figure 6 for the case of $w \leq w_m$ and a fixed value of y.

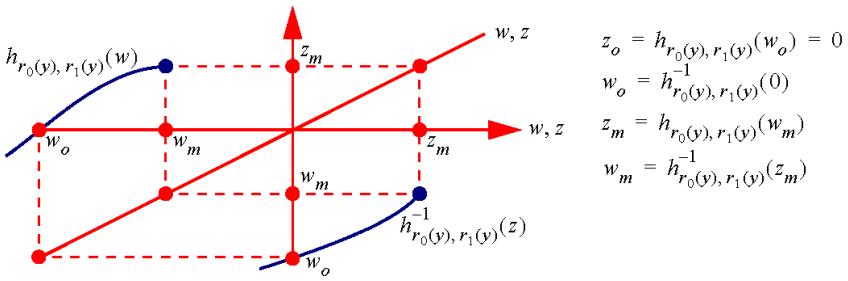


Figure 6. Illustration of the functions $z = h_{r_0(y),r_1(y)}(w)$ and $w = h^{-1}_{r_0(y),r_1(y)}(z)$ for y fixed and $w \leq w_m$.

4. Approximations for Inverse Langevin Function

Consistent with Theorem 1, an approximation to $h^{-1}_{r_0(y),r_1(y)}(0)$, for $0 \leq y < 1$, leads to an approximation for the inverse Langevin function. To establish approximations for $h^{-1}_{r_0(y),r_1(y)}(0)$, first, consider the following two restricted domain approximations which arise from the analysis detailed in Appendix B:

$$\begin{array}{ll} h^{-1}_{r_0(y),r_1(y)}(0) \approx -2 - 4y, & 0 < y \ll 1, \\ h^{-1}_{r_0(y),r_1(y)}(0) \approx \frac{-2}{1-y}, & y = 1 - \Delta, 0 < \Delta \ll 1. \end{array} \tag{26}$$

These approximations are consistent with the first order approximation:

$$h^{-1}_{1,r_0(y),r_1(y)}(0) = \tfrac{-2}{1-y} - 2y = \tfrac{-2}{1-y} \cdot \left[1 + y - y^2\right], \quad 0 \leq y < 1, \tag{27}$$

Consistency arises, first, by noting $-2/(1-y) \approx -2(1+y)$, $0 < y \ll 1$, and thus, the approximation for $y \to 0$ is $-2 - 4y$ as required. Second, for the case of $y = 1 - \Delta$, $0 < \Delta \ll 1$, the approximation approaches $-2/(1-y)$ as required. A measure of the error in this approximation is shown in Figure 7 and the relative error bound associated with the interval $(0, 1)$ is 4.45×10^{-2}.

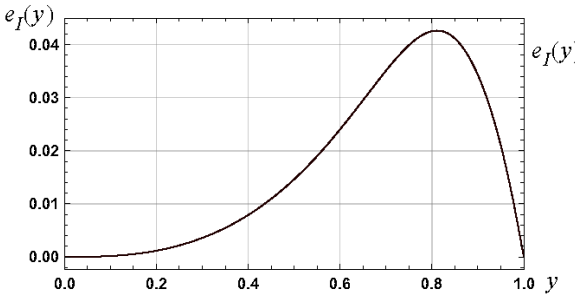


Figure 7. Graph, over the interval $y \in (0,1)$, of a measure of the relative error, e_I, in the approximation $h^{-1}_{1,r_0(y),r_1(y)}(0)$, as defined by (27), to $h^{-1}_{r_0(y),r_1(y)}(0)$.

The approximation defined by (27) and the relative error measure e_I shown in Figure 7 are consistent with

$$h^{-1}_{r_0(y),r_1(y)}(0) = h^{-1}_{1,r_0(y),r_1(y)}(0)\cdot[1 - e_I(y)], \tag{28}$$

and this relationship provides a useful basis for establishing better approximations. Improved approximations include:

$$\begin{gathered} h^{-1}_{2,r_0(y),r_1(y)}(0) = \tfrac{-2}{1-y}\cdot\left[1 + c_{21}\sin(\pi y) + c_{22}\sin\left(\pi y^2\right)\right] \\ c_{21} = \tfrac{281}{1000}, \quad c_{22} = \tfrac{-65}{1000} \end{gathered} \tag{29}$$

$$h^{-1}_{3,r_0(y),r_1(y)}(0) = \frac{-2}{1-y}\cdot\left[1 + y - y^2\right] + y^{7/2} = \frac{-2}{1-y}\cdot\left[1 + y - y^2 - \frac{y^{7/2}}{2} + \frac{y^{9/2}}{2}\right] \tag{30}$$

$$\begin{aligned} h^{-1}_{4,r_0(y),r_1(y)}(0) &= \tfrac{-2}{1-y}\cdot\left[1 + y - y^2\right]\cdot\left[1 - \tfrac{y^4(1-y)}{2}\right] \\ &= \tfrac{-2}{1-y}\cdot\left[1 + y - y^2 - \tfrac{y^4}{2} + y^6 - \tfrac{y^7}{2}\right] \end{aligned} \tag{31}$$

$$h^{-1}_{5,r_0(y),r_1(y)}(0) = \tfrac{-2}{1-y}\cdot\left[1 + y - y^2 - \tfrac{y^{7/2}}{2} + \tfrac{y^{9/2}}{2}\right] + c_{51}y^5(1-y)(y-y_o)e^{-3(y-y_o)^2} \tag{32}$$
$$c_{51} = \tfrac{1133}{200}, \quad y_o = \tfrac{2212}{3125}$$

$$\begin{aligned} h^{-1}_{6,r_0(y),r_1(y)}(0) = \; & \tfrac{-2}{1-y}\cdot\left[1 + y - y^2 - \tfrac{y^4}{2} + y^6 - \tfrac{y^7}{2}\right] \\ & \left[1 + \alpha_0 y^2(1-y)^2\left[1 + \alpha_2 y^2 + \alpha_3 y^3 + \alpha_4 y^4 + \alpha_5 y^5 + \alpha_7 y^7 + \alpha_8 y^8\right]\right] \\ & \alpha_0 = \tfrac{-5813}{1{,}000{,}000}, \alpha_2 = \tfrac{3349}{10}, \alpha_3 = \tfrac{-23467}{10}, \alpha_4 = 6526, \\ & \alpha_5 = -7920, \alpha_7 = 8222, \alpha_8 = -4799. \end{aligned} \tag{33}$$

Graphs of the magnitude of the relative errors in these approximations are shown in Figure 8. The respective relative error bounds associated with these approximations, and for the interval $(0,1)$, are: 4.66×10^{-3}, 3.71×10^{-3}, 2.13×10^{-3}, 2.33×10^{-4} and 4.58×10^{-5}.

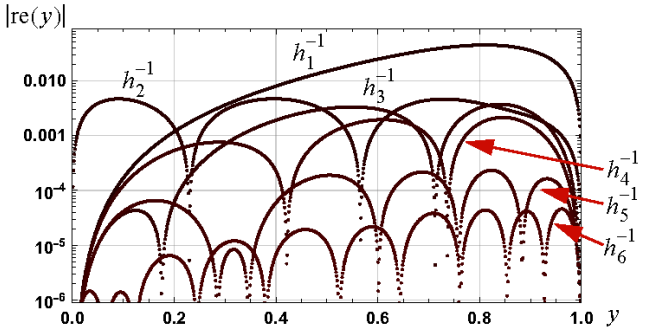


Figure 8. Graphs of the magnitude of the relative error in approximations to $h^{-1}_{r_0(y),r_1(y)}(0)$ as defined by (27) and (29) to (33).

4.1. Associated Langevin Approximations

Given the relationship $L^{-1}(y) = \frac{-1}{1+y} - \frac{h^{-1}_{r_0(y),r_1(y)}(0)}{2}$, as stated in Theorem 1, the above approximations for $h^{-1}_{r_0(y),r_1(y)}(0)$, as specified by (27) and (28) to (33), lead, respectively, to the following approximations for the inverse Langevin function:

$$L_1^{-1}(y) = \frac{3y}{1-y^2}\cdot\left[1 - \frac{y^2}{3}\right] \tag{34}$$

$$\begin{gathered} L_2^{-1}(y) = \frac{2y}{1-y^2} + \frac{d_{21}\sin(\pi y) + d_{22}\sin\left(\pi y^2\right)}{1-y} \\ d_{21} = \frac{301}{1000}, \quad d_{22} = \frac{-92}{1000}. \end{gathered} \tag{35}$$

$$L_3^{-1}(y) = \frac{3y}{1-y^2}\cdot\left[1 - \frac{y^2}{3} - \frac{y^{5/2}}{6} + \frac{y^{9/2}}{6}\right] \tag{36}$$

$$L_4^{-1}(y) = \frac{3y}{1-y^2}\cdot\left[1 - \frac{y^2}{3} - \frac{y^3}{6} - \frac{y^4}{6} + \frac{y^5}{3} + \frac{y^6}{6} - \frac{y^7}{6}\right] \tag{37}$$

$$L_5^{-1}(y) = \frac{3y}{1-y^2}\cdot\left[1 - \frac{y^2}{3} - \frac{y^{5/2}}{6} + \frac{y^{9/2}}{6}\right] - \frac{c_{51}y^5(1-y)(y-y_o)}{2}\cdot e^{-3(y-y_o)^2} \tag{38}$$

$$L_6^{-1}(y) = \frac{3y}{1-y^2}\cdot\left[\begin{array}{l} 1 - \frac{y^2}{3} - \frac{y^3}{6} - \frac{y^4}{6} + \frac{y^5}{3} + \frac{y^6}{6} - \frac{y^7}{6} + \\ \frac{\alpha_0 y\left(1-y^2\right)^2}{3}\cdot\left[1 - y^2 + y^3 - \frac{3y^4}{2} + \frac{3y^5}{2} - \frac{y^6}{2}\right]\cdot \\ \left[1 + \alpha_2 y^2 + \alpha_3 y^3 + \alpha_4 y^4 + \alpha_5 y^5 + \alpha_7 y^7 + \alpha_8 y^8\right] \end{array}\right] \tag{39}$$

Graphs of the magnitude of the relative errors in these approximations are shown in Figure 9. The relative error bounds associated with these approximations, and for the interval $(0,1)$, respectively, are: 4.94×10^{-2}, 1.83×10^{-2}, 4.37×10^{-3}, 2.44×10^{-3}, 2.58×10^{-4} and 4.85×10^{-5}.

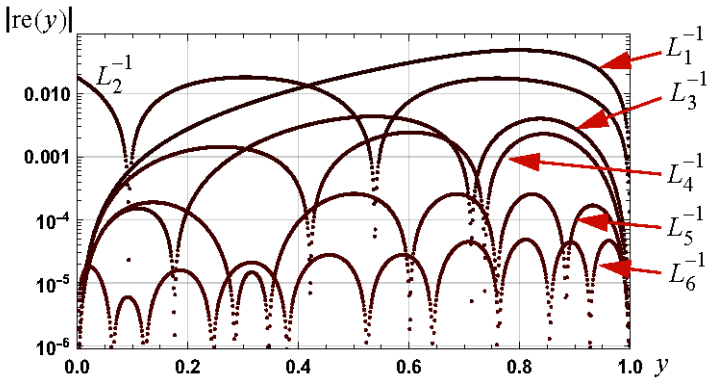


Figure 9. Graphs of the magnitude of the relative error in approximations to $L^{-1}(y)$ as defined by (34) to (39).

Notes

The coefficients utilized in the approximations $h^{-1}_{2,r_0(y),r_1(y)}$ and L_2^{-1} are different and have been chosen to minimize their respective relative error bounds over the interval $[0,1)$.

The fourth approximation, as given by (37), has a relative error bound of 2.44×10^{-3} over the interval $[0,1)$. and is similar in form to that proposed by Kröger [8] and stated in (12). The proposed approximation includes an order seven (rather than six) polynomial and achieves a slightly lower relative error bound.

The fifth (38) and sixth (39) approximations achieve lower relative error bounds, respectively, than the published approximations provided by Nguessong [5] and Marchi [13] and which are stated in (14) and (15).

The proposed approximations complement the list of published approximations for the inverse Langevin function and with relative error bounds consistent with, or lower than, comparable published approximations. Additional approximations remain to be proposed.

5. Schröder Based Higher Order Approximations

The following general approximations for an inverse function arise from the Schröder root approximation of the first kind, e.g., [25] (Equation (21)), [19] (Equation (17)), [20] (Section 3) and [21]:

Theorem 2. *Schröder Based Approximations for an Inverse Function*

Consider a real function f that is differentiable up to at least order n and is also strictly monotonic in an interval around a point x_0 which includes the root x_o of $g(x) = f(x) - y_o$ that is close to x_0. A nth order Taylor series for g^{-1}, based on the point (y_0, x_0), $y_0 = f(x_0) - y_o$, yields the nth order Schröder approximation for $x_o = f^{-1}(y_o)$, denoted x_n, and defined according to

$$x_n = x_0 + \sum_{k=1}^{n} \frac{(-1)^k [f(x_0) - y_o]^k}{k!} \cdot D^{(k)}\left[f^{-1}[f(x_0)]\right], \quad n \in \{1, 2, \cdots\}, \tag{40}$$

where $D^{(k)}$ is the operator denoting kth order differentiation. An explicit form is

$$\begin{aligned}
\mathrm{x}_n = x_0 &- \frac{f(x_0)-y_o}{f^{(1)}(x_0)} - \frac{[f(x_0)-y_o]^2 f^{(2)}(x_0)}{2\left[f^{(1)}(x_0)\right]^3} - \frac{[f(x_0)-y_o]^3 f^{(3)}(x_0)}{6\left[f^{(1)}(x_0)\right]^4} \cdot \left[-1 + \frac{3\left[f^{(2)}(x_0)\right]^2}{f^{(1)}(x_0) f^{(3)}(x_0)}\right] - \\
&\frac{[f(x_0)-y_o]^4 f^{(4)}(x_0)}{24\left[f^{(1)}(x_0)\right]^5} \cdot \left[1 - \frac{10 f^{(2)}(x_0) f^{(3)}(x_0)}{f^{(1)}(x_0) f^{(4)}(x_0)} + \frac{15\left[f^{(2)}(x_0)\right]^3}{\left[f^{(1)}(x_0)\right]^2 f^{(4)}(x_0)}\right] - \\
&\frac{[f(x_0)-y_o]^5 f^{(5)}(x_0)}{120\left[f^{(1)}(x_0)\right]^6} \cdot \left[\begin{array}{l} -1 + \frac{15 f^{(2)}(x_0) f^{(4)}(x_0)}{f^{(1)}(x_0) f^{(5)}(x_0)} + \frac{10\left[f^{(3)}(x_0)\right]^2}{f^{(1)}(x_0) f^{(5)}(x_0)} - \\ \frac{105\left[f^{(2)}(x_0)\right]^2 f^{(3)}(x_0)}{\left[f^{(1)}(x_0)\right]^2 f^{(5)}(x_0)} + \frac{105\left[f^{(2)}(x_0)\right]^4}{\left[f^{(1)}(x_0)\right]^3 f^{(5)}(x_0)} \end{array}\right] - \\
&\frac{[f(x_0)-y_o]^6 f^{(6)}(x_0)}{720\left[f^{(1)}(x_0)\right]^7} \cdot \left[\begin{array}{l} 1 - \frac{21 f^{(2)}(x_0) f^{(5)}(x_0)}{f^{(1)}(x_0) f^{(6)}(x_0)} - \frac{35 f^{(3)}(x_0) f^{(4)}(x_0)}{f^{(1)}(x_0) f^{(6)}(x_0)} + \frac{210\left[f^{(2)}(x_0)\right]^2 f^{(4)}(x_0)}{\left[f^{(1)}(x_0)\right]^2 f^{(6)}(x_0)} + \\ \frac{280 f^{(2)}(x_0)\left[f^{(3)}(x_0)\right]^2}{\left[f^{(1)}(x_0)\right]^2 f^{(6)}(x_0)} - \frac{1260\left[f^{(2)}(x_0)\right]^3 f^{(3)}(x_0)}{\left[f^{(1)}(x_0)\right]^3 f^{(6)}(x_0)} + \frac{945\left[f^{(2)}(x_0)\right]^5}{\left[f^{(1)}(x_0)\right]^4 f^{(6)}(x_0)} \end{array}\right] - \\
&\cdots + \frac{(-1)^n [f(x_0)-y_o]^n}{n!} \cdot D^{(\mathrm{n})}\left[f^{-1}[f(x_0)]\right]
\end{aligned} \tag{41}$$

Proof of Theorem 2. This result arises from a Taylor series approximation for g^{-1} at the point (y_0, x_0) and the inverse function theorem. See, for example, [21]. □

5.1. First, Second and Third Order Approximations for $h^{-1}(0)$

Consider $h_{r_0(y),r_1(y)}(w) = r_0(y) + r_1(y)w + we^w$, with r_0 and r_1 as defined in (18), and an initial approximation for $h^{-1}_{r_0(y),r_1(y)}(0)$ being denoted $h_0^{-1}(0)$ where the argument of y is implicit and assumed to be on the interval $[0, 1)$. It then follows that the first, second, and third order Schröder based approximations for $h^{-1}_{r_0(y),r_1(y)}(0)$, denoted, respectively, $h_1^{-1}(0)$, $h_2^{-1}(0)$, and $h_3^{-1}(0)$ are:

$$h_1^{-1}(0) = \left.\frac{-4 + (1+y)^2 w^2 \exp\left[w + \frac{2}{1+y}\right]}{1 - y^2 + (1+y)^2(1+w)\exp\left[w + \frac{2}{1+y}\right]}\right|_{w = h_0^{-1}(0)} \tag{42}$$

$$h_2^{-1}(0) = \frac{-4+(1+y)^2w^2\exp\left[w+\frac{2}{1+y}\right]}{1-y^2+(1+y)^2(1+w)\exp\left[w+\frac{2}{1+y}\right]} - \left.\frac{\left[4+\left(1-y^2\right)w+(1+y)^2w\exp\left[w+\frac{2}{1+y}\right]\right]^2(2+w)\exp\left[w+\frac{2}{1+y}\right]}{2(1+y)\left[1-y+(1+y)(1+w)\exp\left[w+\frac{2}{1+y}\right]\right]^3}\right|_{w=h_0^{-1}(0)} \tag{43}$$

$$h_3^{-1}(0) = h_2^{-1}(0) - \left.\frac{\left[4+\left(1-y^2\right)w+(1+y)^2w\exp\left[w+\frac{2}{1+y}\right]\right]^3\exp\left[w+\frac{2}{1+y}\right]\cdot\left[-(1-y)(3+w)+(1+y)\left(9+8w+2w^2\right)\exp\left[w+\frac{2}{1+y}\right]\right]}{6(1+y)^2\left[1-y+(1+y)(1+w)\exp\left[w+\frac{2}{1+y}\right]\right]^5}\right|_{w=h_0^{-1}(0)} \tag{44}$$

The proofs for these results are detailed in Appendix C.

Results

The relative error bounds associated with the above approximations, and based on $h_0^{-1}(0)$ as specified by (27) and (29) to (33), are detailed in Table 1. Specific cases are detailed below.

5.2. First Order Approximations: Explicit Cases

The general first order approximation form specified by (42) leads to the approximations detailed below which are valid for $y \in [0,1)$. The associated approximations for the inverse Langevin function can then be defined consistent with Theorem 1.

5.2.1. Case 1

With the initial approximation $h_{1,r_0(y),r_1(y)}^{-1}(0)$ specified by (27), it follows that

$$h_1^{-1}(0) = \frac{-4(1-y)^2+4\left(1+2y-y^3\right)^2\exp\left[\frac{-6y}{1-y^2}\cdot\left[1-\frac{y^2}{3}\right]\right]}{(1+y)(1-y)^3-\left(1-y^2\right)\left(1+4y+y^2-2y^3\right)\exp\left[\frac{-6y}{1-y^2}\cdot\left[1-\frac{y^2}{3}\right]\right]} \tag{45}$$

which has a relative error bound of 2.69×10^{-4}. The relative error bound in the associated approximation for the inverse Langevin function is 3.36×10^{-4}.

5.2.2. Case 2

With the approximation $h_{3,r_0(y),r_1(y)}^{-1}(0)$ defined by (30), it follows, that

$$h_1^{-1}(0) = \frac{-4(1-y)^2+(1+y)^2\left[-2-2y+2y^2+y^{7/2}-y^{9/2}\right]^2\exp[p_E(y)]}{(1-y)^3(1+y)-(1-y)(1+y)^2\left[1+3y-2y^2-y^{7/2}+y^{9/2}\right]\exp[p_E(y)]}$$
$$p_E(y) = \frac{-6y}{1-y^2}\cdot\left[1-\frac{y^2}{3}-\frac{y^{5/2}}{6}+\frac{y^{9/2}}{6}\right] \tag{46}$$

which has a relative error bound of 8.52×10^{-6}. The relative error bound in the associated approximation for the inverse Langevin function is 1.17×10^{-5}.

5.2.3. Case 3

With the approximation $h_{4,r_0(y),r_1(y)}^{-1}(0)$ defined by (31), it follows that

$$h_1^{-1}(0) = \frac{-4(1-y)^2+4(1+y)^2\left[1+y-y^2-\frac{y^4}{2}+y^6-\frac{y^7}{2}\right]^2\exp[p_E(y)]}{(1-y)^3(1+y)-(1-y)(1+y)^2\left[1+3y-2y^2-y^4+2y^6-y^7\right]\exp[p_E(y)]}$$
$$p_E(y) = \frac{-6y}{1-y^2}\cdot\left[1-\frac{y^2}{3}-\frac{y^3}{6}-\frac{y^4}{6}+\frac{y^5}{3}+\frac{y^6}{6}-\frac{y^7}{6}\right] \tag{47}$$

which has a relative error bound of 1.92×10^{-6}. The relative error bound in the associated approximation for the inverse Langevin function is 2.66×10^{-6}.

5.2.4. Case 4

With the approximation $h^{-1}_{6,r_0(y),r_1(y)}(0)$ defined by (33), it follows that

$$h_1^{-1}(0) = \frac{-4(1-y)^2+4(1+y)^2\left[1+y-y^2-\frac{y^4}{2}+y^6-\frac{y^7}{2}\right]^2 p^2(y)\exp[p_E(y)]}{(1-y)^3(1+y)+(1-y)(1+y)^2\left[1-y-2\left[1+y-y^2-\frac{y^4}{2}+y^6-\frac{y^7}{2}\right]p(y)\right]\exp[p_E(y)]} \tag{48}$$

where

$$p(y) = 1+\alpha_0 y^2(1-y)^2\left[1+\alpha_2 y^2+\alpha_3 y^3+\alpha_4 y^4+\alpha_5 y^5+\alpha_7 y^7+\alpha_8 y^8\right] \tag{49}$$

$$p_E(y) = \frac{-6y}{1-y^2}\cdot\left[\begin{array}{l} 1-\frac{y^2}{3}-\frac{y^3}{6}-\frac{y^4}{6}+\frac{y^5}{3}+\frac{y^6}{6}-\frac{y^7}{6}+ \\ \frac{\alpha_0 y\left(1-y^2\right)^2}{3}\cdot\left[1-y^2+y^3-\frac{3y^4}{2}+\frac{3y^5}{2}-\frac{y^6}{2}\right]\cdot \\ \left[1+\alpha_2 y^2+\alpha_3 y^3+\alpha_4 y^4+\alpha_5 y^5+\alpha_7 y^7+\alpha_8 y^8\right] \end{array}\right] \tag{50}$$

This approximation has a relative error bound of 4.07×10^{-10}. The relative error bound in the associated approximation for the inverse Langevin function is 6.27×10^{-10}.

5.3. Second Order Approximations

With the approximation $h^{-1}_{4,r_0(y),r_1(y)}(0)$ defined by (31), it follows, from (43), that:

$$\begin{aligned} h_2^{-1}(0) = \; & h_1^{-1}(0)+ \\ & 8y\left[\begin{array}{l} -(1-y)^3\left[1-y^2-y^3-\frac{y^4}{2}+\frac{y^5}{2}+\frac{y^6}{2}\right]+ \\ (1+y)^3\left[1-y^2+y^3-\frac{3y^4}{2}+\frac{3y^5}{2}-\frac{y^6}{2}\right]\exp[p_E(y)] \end{array}\right]^2\cdot \\ & \frac{\left[1-\frac{y^2}{2}-\frac{y^3}{4}+\frac{y^5}{2}-\frac{y^6}{4}\right]\exp[p_E(y)]}{(1+y)\left[(1-y)^2-(1+y)\left[1+3y-2y^2-y^4+2y^6-y^7\right]\exp[p_E(y)]\right]^3} \end{aligned} \tag{51}$$

where $h_1^{-1}(0)$ and $p_E(y)$ are defined by (47). The relative error bound for this approximation is 9.12×10^{-9}. The relative error bound in the associated approximation for the inverse Langevin function is 1.16×10^{-8}.

5.4. Direct First Order Schröder Approximation to Inverse Langevin Function

Consider the first order Schröder approximation to the inverse Langevin function, based on an initial approximation denoted L_0^{-1}, and defined by

$$L_1^{-1}(y) = x - \left.\frac{L(x)-y}{L^{(1)}(x)}\right]_{x=L_0^{-1}(y)} \tag{52}$$

The nature of L_1^{-1} depends on the form assumed for the Langevin function. For the case of L defined according to $L(x) = \coth(x) - 1/x$, it follows that:

$$L_1^{-1}(y) = \left.\frac{x^3 - x\sinh(x)\left[2\sinh(x) - x\cosh(x) + xy\sinh(x)\right]}{x^2-\sinh(x)^2}\right]_{x=L_0^{-1}(y)} \tag{53}$$

For the case of L defined according to (3), it follows that:

$$L_1^{-1}(y) = \left.\frac{x[2-x+xy]-2x\left[2+2x^2+xy\right]e^{-2x}+x[2+x+xy]e^{-4x}}{1-2[1+2x^2]e^{-2x}+e^{-4x}}\right]_{x=L_0^{-1}(y)} \tag{54}$$

Comparison

The approximations defined by (53) and (54) can be contrasted with approximating the inverse Langevin function according to the relationship stated in Theorem 1, i.e.,

$$L^{-1}(y) \approx \frac{-1}{1+y} - \frac{h_1^{-1}(0)}{2} \tag{55}$$

where $h_1^{-1}(0)$ is the first order approximation defined according to (42). To compare the approximations defined by (54) and (55), consider the initial approximations

$$L_0^{-1}(y) = \frac{3y}{1-y^2} \cdot \left[1 - \frac{2y^2}{5} - \frac{y^4}{15} + \frac{2y^6}{15}\right] \tag{56}$$

$$h_0^{-1}(0) = \frac{-2}{1-y} \cdot \left[1 + y - y^2 - \frac{y^4}{2} + y^6 - \frac{y^7}{2}\right] \tag{57}$$

The first was proposed by Kröger [8] (stated above in (12)) and has a relative error bound of 2.75×10^{-3}. The second approximation is specified in (31) and has a relative error bound of 2.13×10^{-3}. Use of (56) in (54) yields the approximation

$$\begin{array}{l} L_1^{-1}(y) = -p_E(y) \\ \quad \dfrac{\left[\begin{array}{l} (1-y)^2\left[1+\frac{y}{2}-\frac{y^2}{2}+\frac{3y^3}{5}+\frac{3y^4}{5}+\frac{y^5}{10}+\frac{y^6}{10}-\frac{y^7}{5}-\frac{y^8}{5}\right]+ \\ (1+y)^2\left[1-\frac{y}{2}-\frac{y^2}{2}-\frac{3y^3}{5}+\frac{3y^4}{5}-\frac{y^5}{10}+\frac{y^6}{10}+\frac{y^7}{5}-\frac{y^8}{5}\right]e^{2p_E(y)}- \\ 2\left[1+\frac{17y^2}{2}-\frac{83y^4}{10}+\frac{37y^6}{50}+\frac{159y^8}{50}-\frac{28y^{10}}{25}-\frac{4y^{12}}{25}+\frac{4y^{14}}{25}\right]e^{p_E(y)} \end{array}\right]}{\left(1-y^2\right)^2\left[1+e^{2p_E(y)}\right]-2\left[1+16y^2-\frac{67y^4}{5}+\frac{12y^6}{25}+\frac{144y^8}{25}-\frac{46y^{10}}{25}-\frac{8y^{12}}{25}+\frac{8y^{14}}{25}\right]e^{2p_E(y)}} \\ p_E(y) = \frac{-6y}{1-y^2} \cdot \left[1 - \frac{2y^2}{5} - \frac{y^4}{15} + \frac{2y^6}{15}\right] \end{array} \tag{58}$$

with a relative error bound of 7.34×10^{-6}. In contrast, the use of (57) in (42), and then (55), leads to the approximation specified in (47), i.e.,

$$\begin{array}{l} h_1^{-1}(0) = \dfrac{-4(1-y)^2+4(1+y)^2\left[1+y-y^2-\frac{y^4}{2}+y^6-\frac{y^7}{2}\right]^2\exp[p_E(y)]}{(1-y)^3(1+y)-(1-y)(1+y)^2\left[1+3y-2y^2-y^4+2y^6-y^7\right]\exp[p_E(y)]} \\ p_E(y) = \frac{-6y}{1-y^2} \cdot \left[1 - \frac{y^2}{3} - \frac{y^3}{6} - \frac{y^4}{6} + \frac{y^5}{3} + \frac{y^6}{6} - \frac{y^7}{6}\right] \end{array} \tag{59}$$

which has a relative error bound of 1.92×10^{-6}. The associated approximation for $L^{-1}(y)$ is

$$L_1^{-1}(y) = \frac{-1}{1+y} - \frac{h_1^{-1}(0)}{2} \tag{60}$$

and has a relative error bound of 2.66×10^{-6}.

Clearly, (59) and (60) have a simpler form than (58) and yield an approximation for $L^{-1}(y)$ with a lower relative error bound. This, in part, is the rationale for the use of the link between the inverse Langevin function and the r_0-r_1-Lambert W function.

6. Applications

6.1. Limit Approximations

The approximations detailed in (26) lead to the following approximations for the inverse Langevin function after using the relationship defined in Theorem 1:

$$\begin{array}{ll} L^{-1}(y) \approx \frac{3y}{1+y} \cdot \left[1 + \frac{2y}{3}\right], & 0 < y \ll 1, \\ L^{-1}(y) \approx \frac{2y}{1-y^2}, & y = 1 - \Delta, 0 < \Delta \ll 1. \end{array} \tag{61}$$

These approximations can be contrasted with established approximations. First, a Taylor series approximation for the inverse Langevin function, e.g., [4], is

$$L^{-1}(y) \approx 3y + \frac{9y^3}{5} + \frac{297y^5}{175} + \frac{1539y^7}{875} + \cdots \tag{62}$$

Second, a first order approximation for the case of $y = 1 - \Delta$, $0 < \Delta \ll 1$, is $L^{-1}(y) \approx 1/(1-y)$ (see Appendix B of [8]). The graphs of the relative errors associated with these approximations are shown in Figure 10. It is clear from these graphs that the approximations defined in (61) are inferior to established approximations for arguments close to zero or close to one. However, they serve as a useful basis for establishing approximations with low relative error bounds over the interval $(0,1)$ for h^{-1} and, hence, L^{-1}, as summarized in Table 2.

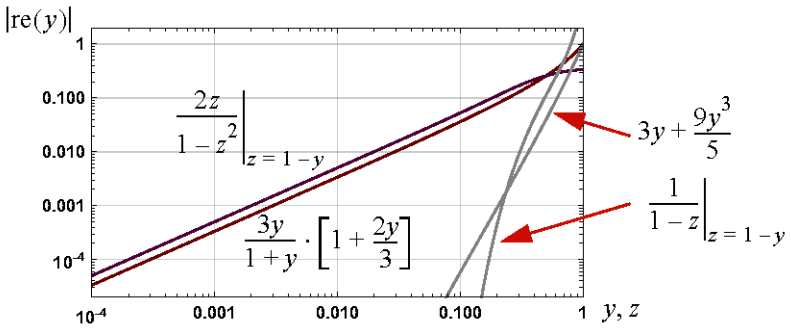


Figure 10. Graph of the relative errors in approximations to the inverse Langevin function for the cases of $0 < y \ll 1$ and $y = 1 - \Delta$, $0 < \Delta \ll 1$.

6.2. Integral Results

As $\int_0^x L(\lambda)d\lambda = \ln[\sinh(x)] - \ln(x)$, the general integral result

$$\int_{y_1}^{y} f^{-1}(\lambda)d\lambda = yf^{-1}(y) - y_1 f^{-1}(y_1) - \int_{f^{-1}(y_1)}^{f^{-1}(y)} f(\lambda)d\lambda \tag{63}$$

assuming f^{-1} and f are well defined, yields

$$\int_0^y L^{-1}(\lambda)d\lambda = yL^{-1}(y) + \ln\left[L^{-1}(y)\right] - \ln\left[\sinh\left[L^{-1}(y)\right]\right], \quad y \in (0,1). \tag{64}$$

The proposed approximations can be utilized to establish approximations for the integral of the inverse Langevin function. For example, use of the approximations specified by (36) and (37), in (64), yield approximations for the inverse Langevin function with the respective relative error bounds of 1.40×10^{-5} and 3.91×10^{-6}. The latter approximation is:

$$\begin{aligned} \int_0^y L^{-1}(\lambda)d\lambda \quad \approx\ & \tfrac{3y^2}{1-y^2}\cdot\left[1 - \tfrac{y^2}{3} - \tfrac{y^3}{6} - \tfrac{y^4}{6} + \tfrac{y^5}{3} + \tfrac{y^6}{6} - \tfrac{y^7}{6}\right] + \\ & \ln\left[\tfrac{3y}{1-y^2}\cdot\left[1 - \tfrac{y^2}{3} - \tfrac{y^3}{6} - \tfrac{y^4}{6} + \tfrac{y^5}{3} + \tfrac{y^6}{6} - \tfrac{y^7}{6}\right]\right] - \\ & \ln\left[\sinh\left[\tfrac{3y}{1-y^2}\cdot\left[1 - \tfrac{y^2}{3} - \tfrac{y^3}{6} - \tfrac{y^4}{6} + \tfrac{y^5}{3} + \tfrac{y^6}{6} - \tfrac{y^7}{6}\right]\right]\right], \quad y \in (0,1). \end{aligned} \tag{65}$$

6.3. Roots of $\alpha + \beta w + we^w$

Consistent with (21) and Theorem 1, the roots of $h_{r_0(y),r_1(y)}(w) = r_0(y) + r_1(y)w + we^w$, $0 < y < 1$ are $\frac{-2}{1+y}$ and $\frac{-2}{1+y} - 2L^{-1}(y)$. A desirable extension to these results would be a specification of the roots of

$$h_{\alpha,\beta}(w) = \alpha + \beta w + we^w \tag{66}$$

for arbitrary values of α and β. Given the definitions of r_0 and r_1, this, in general, is not possible. However, roots can be established for specific cases. For example, the roots of

$$\frac{(1+k)^2}{e^{1+k}} + \frac{kw}{e^{1+k}} + we^w, \quad 0 < k < 1, \tag{67}$$

are $-1-k$ and $-1-k-2L^{-1}\left[\frac{1-k}{1+k}\right]$, and arise from the case of $y = \frac{1-k}{1+k}$, $0 < k < 1$. As an example, the roots of

$$\frac{16}{9e^{4/3}} + \frac{w}{3e^{4/3}} + we^w \tag{68}$$

are $-4/3$ and $-4/3 - 2L^{-1}(1/2) \approx -4.927$, and arise from the case of $y = 1/2$, $k = 1/3$.

Linear System Impulse Response

Note that $\alpha + \beta v_o - v_o e^{-v_o} = 0$ is consistent with $\alpha + \beta w_o + w_o e^{w_o} = 0$ provided $v_o = -w_o$, i.e., the roots of $\alpha + \beta w + we^w = 0$ are the negative of the roots of $\alpha + \beta v - ve^{-v} = 0$.

Consistent with (67), consider a linear system with an impulse response defined by

$$h(t) = 1 - k_1 t - k_2 t e^{-t}, \quad t \geq 0, \tag{69}$$

where $k_1 = \frac{k}{(1+k)^2}$, $k_2 = \frac{e^{1+k}}{(1+k)^2}$, $0 < k < 1$. The transfer function of such a system is defined according to

$$H(s) = \frac{1}{s} - \frac{k_1}{s^2} - \frac{k_2}{(1+s)^2} \tag{70}$$

and is consistent with a system comprising an integrator, a cascade of two integrators, and a second order system. The roots of the impulse response are $1+k$ and $1+k+2L^{-1}\left[\frac{1-k}{1+k}\right]$. The impulse response is shown in Figure 11 for the cases of $k \in \{1/10, 1/4, 1/2\}$.

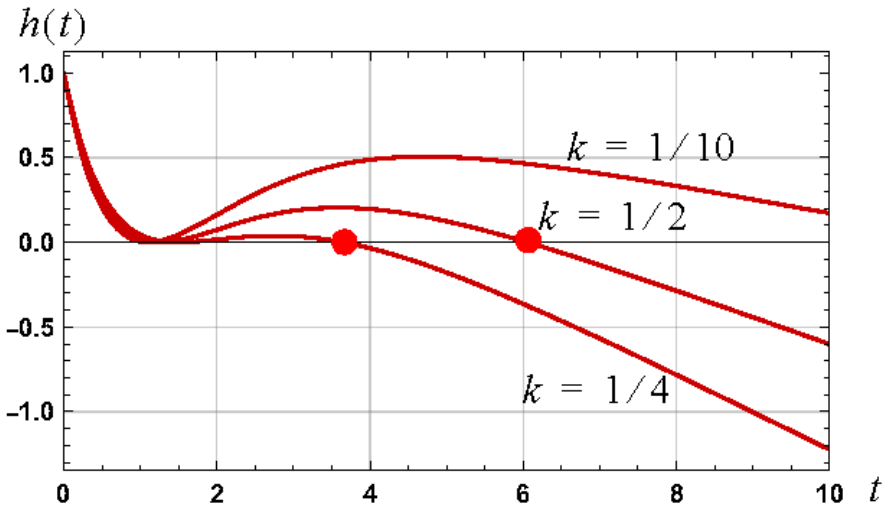


Figure 11. Graph of the impulse response of the transfer function defined by (69) for values of $k \in \{1/10, 1/4, 1/2\}$.

7. Summary of Results and Conclusion

7.1. Summary of Results

The approximations to $h^{-1}_{r_0(y),r_1(y)}(0)$, along with their associated relative error bounds, are detailed in Table 1. The associated approximations for the inverse Langevin function are detailed in Table 2.

Table 1. Summary of approximations for $h^{-1}_{r_0(y),r_1(y)}(0)$, as defined by (27) and (29) to (33), and their relative error bounds for the interval $[0,1)$.

Approximation	Definition	Relative Error Bound	Relative Error Bound: 1st Order—(42)	Relative Error Bound: 2nd Order—(43)	Relative Error Bound: 3rd Order—(44)
$h^{-1}_{1,r_0(y),r_1(y)}(0)$	$\frac{-2}{1-y}\cdot\left[1+y-y^2\right]$	4.45×10^{-2}	2.69×10^{-4}	1.87×10^{-5}	1.32×10^{-6}
$h^{-1}_{2,r_0(y),r_1(y)}(0)$	$\frac{-2}{1-y}\cdot\left[1+c_{21}\sin(\pi y)+c_{22}\sin(\pi y^2)\right]$	4.66×10^{-3}	2.20×10^{-4}	2.42×10^{-5}	3.68×10^{-6}
$h^{-1}_{3,r_0(y),r_1(y)}(0)$	$\frac{-2}{1-y}\cdot\left[1+y-y^2-\frac{y^{7/2}}{2}+\frac{y^{9/2}}{2}\right]$	3.71×10^{-3}	8.52×10^{-6}	7.58×10^{-8}	7.74×10^{-10}
$h^{-1}_{4,r_0(y),r_1(y)}(0)$	$\frac{-2}{1-y}\cdot\left[1+y-y^2-\frac{y^4}{2}+y^6-\frac{y^7}{2}\right]$	2.13×10^{-3}	1.92×10^{-6}	9.12×10^{-9}	4.59×10^{-11}
$h^{-1}_{5,r_0(y),r_1(y)}(0)$	$\frac{-2}{1-y}\cdot\left[1+y-y^2-\frac{y^{7/2}}{2}+\frac{y^{9/2}}{2}\right]+$ $c_{51}y^5(1-y)(y-y_o)e^{-3(y-y_o)^2}$	2.33×10^{-4}	3.15×10^{-8}	1.68×10^{-11}	1.04×10^{-14}
$h^{-1}_{6,r_0(y),r_1(y)}(0)$	$\frac{-2}{1-y}\cdot\left[1+y-y^2-\frac{y^4}{2}+y^6-\frac{y^7}{2}\right]\cdot$ $\left[\begin{array}{l}1+\alpha_0y^2(1-y)^2\\ \left[\begin{array}{c}1+\alpha_2y^2+\alpha_3y^3+\alpha_4y^4+\\ \alpha_5y^5+\alpha_7y^7+\alpha_8y^8\end{array}\right]\end{array}\right]$	4.58×10^{-5}	4.07×10^{-10}	2.43×10^{-14}	1.92×10^{-18}

Table 2. Summary of approximations, based on the approximations to $h^{-1}_{r_0(y),r_1(y)}(0)$ detailed in Table 1, for the inverse Langevin function along with their associated relative bounds for the interval $[0,1)$. The relative error bounds for the first, second, and third order Schröder-based approximations arise from the direct use of the approximations in the general approximation form defined in Theorem 2.

Approx.	Definition	Relative Error Bound	Relative Error Bound: First Order	Relative Error Bound: Second Order	Relative Error Bound: Third Order
$L_1^{-1}(y)$	$\frac{3y}{1-y^2}\cdot\left[1-\frac{y^2}{3}\right]$	4.94×10^{-2}	2.40×10^{-3}	1.20×10^{-4}	5.87×10^{-6}
$L_2^{-1}(y)$	$\frac{2y}{1-y^2}+\frac{d_{21}\sin(\pi y)+d_{22}\sin(\pi y^2)}{1-y}$	1.83×10^{-2}	2.94×10^{-4}	5.29×10^{-6}	9.23×10^{-8}
$L_3^{-1}(y)$	$\frac{3y}{1-y^2}\cdot\left[1-\frac{y^2}{3}-\frac{y^{5/2}}{6}+\frac{y^{9/2}}{6}\right]$	4.37×10^{-3}	1.62×10^{-5}	6.59×10^{-8}	2.64×10^{-10}
$L_4^{-1}(y)$	$\frac{3y}{1-y^2}\cdot\left[1-\frac{y^2}{3}-\frac{y^3}{6}-\frac{y^4}{6}+\frac{y^5}{3}+\frac{y^6}{6}-\frac{y^7}{6}\right]$	2.44×10^{-3}	5.32×10^{-6}	1.24×10^{-8}	2.84×10^{-11}
$L_5^{-1}(y)$	$\frac{3y}{1-y^2}\cdot\left[1-\frac{y^2}{3}-\frac{y^{5/2}}{6}+\frac{y^{9/2}}{6}\right]-$ $\frac{c_{51}y^5(1-y)(y-y_o)}{2}\cdot e^{-3(y-y_o)^2}$ $\frac{3y}{1-y^2}\cdot$	2.58×10^{-4}	6.47×10^{-8}	1.67×10^{-11}	4.22×10^{-15}
$L_6^{-1}(y)$	$\left[\begin{array}{l}1-\frac{y^2}{3}-\frac{y^3}{6}-\frac{y^4}{6}+\frac{y^5}{3}+\frac{y^6}{6}-\frac{y^7}{6}+\\ \frac{\alpha_0y(1-y^2)^2}{3}\cdot\left[\begin{array}{c}1-y^2+y^3-\frac{3y^4}{2}+\\ \frac{3y^5}{2}-\frac{y^6}{2}\end{array}\right]\cdot\\ \left[\begin{array}{c}1+\alpha_2y^2+\alpha_3y^3+\alpha_4y^4+\\ \alpha_5y^5+\alpha_7y^7+\alpha_8y^8\end{array}\right]\end{array}\right]$	4.85×10^{-5}	2.32×10^{-9}	1.14×10^{-13}	5.48×10^{-18}

7.2. Conclusions

The relationship between the inverse Langevin function and the r_0-r_1-Lambert W function was defined. The derived relationship leads to new approximations for the inverse Langevin function with potentially lower relative error bounds than comparable published approximations.

It was shown that improved approximations for the r_0-r_1-Lambert W function, with significantly lower relative error bounds, can be defined by utilizing Schröder's root approximations of the first kind. Approximations for the inverse Langevin function can then be defined by utilizing the established relationship between the two functions. An alternative approach is to directly use an establish approximation for the inverse Langevin function in the general Schröder approximation form. Examples were given to illustrate

that the former approach is likely to lead to a simpler analytical expression for a set relative error bound.

It was shown how the defined approximations for the inverse Langevin function can be utilized to establish approximations for the integral of the inverse Langevin function, with, in general, low relative error bounds.

The link between the inverse Langevin function and the r_0-r_1-Lambert W function was utilized to establish the roots of the impulse response of a custom linear system comprising an integrator, a cascade of two integrators, and a second order system.

In terms of future research, it is noted that iterative approaches can be utilized to establish approximations with lower relative error bounds and several results for the inverse Langevin function have been detailed, e.g., [14,21]. Berinda [26] and Gdawiec [27], for example, provide an overview of iterative approaches that can be utilized for fixed point problems. In general, iterative approaches lead to analytical approximations with a high degree of complexity and rely on an initial approximation. To achieve results with modest complexity, a low order of iteration is required along with an initial approximation that has a low relative error bound and with low complexity. New approximations, thus, are of interest and warrant ongoing research.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

Acknowledgments: The author is pleased to acknowledge the support of A. Zoubir, SPG, Technische Universität Darmstadt, Darmstadt, Germany, who hosted a visit where the fundamental results underpinning this paper were established.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A. Proof of Theorem 1

Consider the function $f(x,y) = 1-(1-y)x-[1+(1+y)x]e^{-2x}$ for the general case of $x, y \in \boldsymbol{R}$, $y \neq -1$ and the transformations

$$z = 1+(1+y)x, \quad x = \tfrac{z-1}{1+y}, \tag{A1}$$

$$w = \tfrac{-2z}{1+y}, \quad z = \tfrac{-(1+y)w}{2}. \tag{A2}$$

The definitions of f_2 and f_1, according to $f_2(w,y) = f_1(z,y) = f(x,y)$, then lead to

$$f_1(z,y) = f\left[\frac{z-1}{1+y}, y\right] = 1+\frac{1-y}{1+y}-\frac{(1-y)z}{1+y}-z\exp\left[\frac{-2z}{1+y}\right]\cdot\exp\left[\frac{2}{1+y}\right] \tag{A3}$$

$$f_2(w,y) = f_1\left[\frac{-(1+y)w}{2}, y\right] = \frac{2}{1+y}+\frac{(1-y)w}{2}+\frac{(1+y)w}{2}\cdot\exp(w)\cdot\exp\left[\frac{2}{1+y}\right] \tag{A4}$$

It then follows that

$$\begin{aligned}\tfrac{2e^{-2/(1+y)}}{1+y}\cdot f_2(w,y) = \; & \tfrac{4e^{-2/(1+y)}}{(1+y)^2}+\tfrac{(1-y)e^{-2/(1+y)}w}{1+y}+we^w \\ & = r_0(y)+r_1(y)w+we^w = h_{r_0(y),r_1(y)}(w)\end{aligned} \tag{A5}$$

when

$$r_0(y) = \tfrac{4e^{-2/(1+y)}}{(1+y)^2}, \quad r_1(y) = \frac{(1-y)e^{-2/(1+y)}}{1+y}. \tag{A6}$$

From (A1) and (A2), it is the case that $x = \frac{-1}{1+y} - \frac{w}{2}$ and it then follows, from the relationship $f_2(w,y) = f(x,y)$, that

$$k_f(y) f_2(w,y) = k_f(y) f\left[\frac{-1}{1+y} - \frac{w}{2}, y\right] = h_{r_0(y),r_1(y)}(w) \tag{A7}$$

when $k_f(y) = \frac{2e^{-2/(1+y)}}{1+y}$, which is the first required result.

To establish the last stated result, consider the specific case of $f(x,y) = 0$, $0 < y < 1$, which defines the path

$$\left\{(x,y): \;\; x = L^{-1}(y), y \in (0,1)\right\}. \tag{A8}$$

Along the path defined by $f(x,y) = 0$, $0 < y < 1$, and with $x = \frac{-1}{1+y} - \frac{w}{2}$, it is the case that

$$k_f(y) f\left[\frac{-1}{1+y} - \frac{w}{2}, y\right] = r_0(y) + r_1(y)w + we^w = h_{r_0(y),r_1(y)}(w) = 0 \tag{A9}$$

i.e., $w = h^{-1}_{r_0(y),r_1(y)}(0)$. Hence, with $w = \frac{-2}{1+y} - 2L^{-1}(w)$, it follows that $L^{-1}(y) = \frac{-1}{1+y} - \frac{h^{-1}_{r_0(y),r_1(y)}(0)}{2}$, which is the last stated result. This result is also valid for the case of $y = 0$.

Appendix B. Restricted Domain Approximations for $h_{r_0(y),r_1(y)}(w)$

Appendix B.1. Case 1: 0 < y << 1

For the case of $0 < y \ll 1$, and consistent with the left graph shown in Figure 4, the root of $h_{r_0(y),r_1(y)}$ that is of interest is close to, but left of, the known root of $-2/(1+y)$ (see (21)). Consider $h_{r_0(y),r_1(y)}(w)$, defined by (19), and the case of $h_{r_0(y),r_1(y)}(w) = 0$, i.e.,

$$\frac{4e^{-2/(1+y)}}{(1+y)^2} + \frac{w(1-y)e^{-2/(1+y)}}{1+y} + we^w = 0. \tag{A10}$$

With the form for w of $w = -2/(1+y) + \varepsilon(y)$, it follows that:

$$\frac{4}{(1+y)^2} + \frac{1-y}{1+y}\cdot\left[\frac{-2}{1+y} + \varepsilon(y)\right] + \left[\frac{-2}{1+y} + \varepsilon(y)\right]e^{\varepsilon(y)} = 0. \tag{A11}$$

Simplification, and a Taylor series expansion for the exponential function around the point zero, leads to:

$$2 + (1-y)\varepsilon(y) + [-2 + (1+y)\varepsilon(y)]\left[1 + \varepsilon(y) + \frac{\varepsilon^2(y)}{2} + \cdots\right] = 0. \tag{A12}$$

First, second, and third order expansions for the exponential function lead to the approximations:

$$\begin{cases} \varepsilon^2(y)[1+y] = 0, & \text{first order} \\ \varepsilon^2(y)\left[1 + \frac{(1+y)\varepsilon(y)}{2}\right] = 0, & \text{second order} \\ \varepsilon^2(y)\left[y + \frac{(1+3y)\varepsilon(y)}{6} + \frac{(1+y)\varepsilon^2(y)}{6}\right] = 0, & \text{third order} \end{cases} \tag{A13}$$

with solutions:

$$\begin{cases} \varepsilon(y) = 0, & \text{first order} \\ \varepsilon(y) = 0, \varepsilon(y) = \frac{-2y}{1+y}, & \text{second order} \\ \varepsilon(y) = 0, \varepsilon(y) = \frac{-1-3y \pm \sqrt{1-18y-15y^2}}{2(1+y)}, & \text{third order} \end{cases} \tag{A14}$$

Consider the first non-zero solution for $\varepsilon(y)$, which is the second order solution of $\varepsilon(y) = -2y/(1+y)$. For the case of $y \to 0$, this solution implies $w = -2$, which is consistent with the root $-2/(1+y)$.

The third order solution, which is consistent with $\varepsilon(y) \to 0$, as $y \to 0$, is $\varepsilon(y) = \frac{-1-3y+\sqrt{1-18y-15y^2}}{2(1+y)} \approx \frac{-6y-24y^2}{1+y}$ and leads to the approximation:

$$h_{r_0(y),r_1(y)}\left[\frac{-2-6y-24y^2}{1+y}\right] \approx 0, \quad 0 < y \ll 1. \tag{A15}$$

Hence:

$$h^{-1}_{r_0(y),r_1(y)}(0) \approx \frac{-2-6y-24y^2}{1+y} \approx -2-4y-20y^2 \tag{A16}$$

after using the approximation $(1+y)^{-1} \approx 1-y+y^2$.

Appendix B.2. Case 2: y -> 1

For $y \approx 1$, Taylor series expansions around the point one yield the approximations:

$$\begin{array}{l} r_0(y) = \frac{4e^{-2/(1+y)}}{(1+y)^2} \approx \frac{1}{e} + \frac{1-y}{2e} + \frac{(1-y)^2}{8e} + \cdots \\ r_1(y) = \frac{(1-y)e^{-2/(1+y)}}{1+y} \approx \frac{1-y}{2e} - \frac{(1-y)^3}{16e} - \cdots \end{array} \tag{A17}$$

Substitution of these approximations into

$$h_{r_0(y),r_1(y)}(w) = r_0(y) + r_1(y)w + we^w = 0 \tag{A18}$$

yields

$$\frac{1}{e} + \frac{1-y}{2e} + \frac{(1-y)^2}{8e} + \cdots + \left[\frac{1-y}{2e} - \frac{(1-y)^3}{16e} - \cdots\right]w + we^w = 0. \tag{A19}$$

As $y \to 1$ from below, the require root of (A18) becomes increasingly negative (this is evident from a comparison of the left and right hand graphs shown in Figure 4). For the case of $|w| \gg 1$, and w negative, the term we^w becomes vanishing small leading to

$$w = -\frac{\frac{1}{e} + \frac{1-y}{2e} + \frac{(1-y)^2}{8e}}{\frac{1-y}{2e} - \frac{(1-y)^3}{16e}} \approx \frac{-2}{1-y}, \quad y = 1-\Delta, 0 < \Delta \ll 1, \tag{A20}$$

and the required result

$$h^{-1}_{r_0(y),r_1(y)}(0) \approx \frac{-2}{1-y}, \quad y = 1-\Delta, 0 < \Delta \ll 1. \tag{A21}$$

Appendix C. Proof for First, Second, and Third Order Schröder Approximations

With $z = h_{r_0(y),r_1(y)}(w) = r_0(y) + r_1(y)w + we^w$, it follows, using the notation $h(w) = h_{r_0(y),r_1(y)}(w)$, that

$$\begin{array}{llll} h^{(1)}(w) = r_1(y) + (1+w)e^w, & & h^{(2)}(w) = (2+w)e^w, \\ h^{(3)}(w) = (3+w)e^w, & \cdots & h^{(k)}(w) = (k+w)e^w. \end{array} \tag{A22}$$

It then follows, from Theorem 2, that the first order Schröder approximation is

$$h_1^{-1}(z) = w - \left.\frac{h(w)-z}{h^{(1)}(w)}\right|_{w=h_0^{-1}(z)} = \left.\frac{-r_0(y) + w^2e^w + z}{r_1(y) + (1+w)e^w}\right|_{w=h_0^{-1}(z)} \tag{A23}$$

assuming an initial approximating function h_0^{-1} for h^{-1}. The approximation for the case of $z = 0$ is

$$h_1^{-1}(0) = \left. \frac{-r_0(y) + w^2 e^w}{r_1(y) + (1+w)e^w} \right|_{w=h_0^{-1}(0)} \tag{A24}$$

Substitution for r_0 and r_1, as specified by (18), yields:

$$h_1^{-1}(0) = \left. \frac{-4 + (1+y)^2 w^2 \exp\left[w + \frac{2}{1+y}\right]}{1 - y^2 + (1+y)^2(1+w)\exp\left[w + \frac{2}{1+y}\right]} \right|_{w=h_0^{-1}(0)} \tag{A25}$$

Appendix C.1. Second Order

The second order Schröder approximation to h, as defined by Theorem 2, is

$$\begin{aligned} h_2^{-1}(z) &= h_1^{-1}(z) - \left. \frac{[h(w)-z]^2 h^{(2)}(w)}{2\left[h^{(1)}(w)\right]^3} \right|_{w=h_0^{-1}(z)} \\ &= \frac{-r_0(y)+w^2e^w+z}{r_1(y)+(1+w)e^w} - \left. \frac{[r_0(y)+r_1(y)w+we^w-z]^2(2+w)e^w}{2[r_1(y)+(1+w)e^w]^3} \right|_{w=h_0^{-1}(z)} \end{aligned} \tag{A26}$$

and the approximation for the case of $z = 0$ is

$$h_2^{-1}(0) = \left. \frac{-r_0(y)+w^2e^w}{r_1(y)+(1+w)e^w} - \frac{[r_0(y)+r_1(y)w+we^w]^2(2+w)e^w}{2[r_1(y)+(1+w)e^w]^3} \right|_{w=h_0^{-1}(0)} \tag{A27}$$

Substitution for r_0 and r_1 yields:

$$\begin{aligned} h_2^{-1}(0) = \frac{-4+(1+y)^2w^2\exp\left[w+\frac{2}{1+y}\right]}{1-y^2+(1+y)^2(1+w)\exp\left[w+\frac{2}{1+y}\right]} - \\ \left. \frac{\left[4+\left(1-y^2\right)w+(1+y)^2w\exp\left[w+\frac{2}{1+y}\right]\right]^2(2+w)\exp\left[w+\frac{2}{1+y}\right]}{2(1+y)\left[1-y+(1+y)(1+w)\exp\left[w+\frac{2}{1+y}\right]\right]^3} \right|_{w=h_0^{-1}(0)} \end{aligned} \tag{A28}$$

Appendix C.2. Third Order

The third order Schröder approximation to h, as defined by Theorem 2, is

$$h_3^{-1}(z) = h_2^{-1}(z) - \frac{[h(w)-z]^3 h^{(3)}(w)}{6\left[h^{(1)}(w)\right]^4} \cdot \left[-1 + \frac{3\left[h^{(2)}(w)\right]^2}{h^{(1)}(w)h^{(3)}(w)} \right] \Bigg|_{w=h_0^{-1}(z)} \tag{A29}$$

and the approximation for the case of $z = 0$ is

$$\begin{aligned} h_3^{-1}(0) = h_2^{-1}(0) - \frac{[r_0(y)+r_1(y)w+we^w]^2(3+w)e^w}{6[r_1(y)+(1+w)e^w]^4} \cdot \\ \left. \left[-1 + \frac{3(2+w)^2e^{2w}}{[r_1(y)+(1+w)e^w](3+w)e^w} \right] \right|_{w=h_0^{-1}(0)} \end{aligned} \tag{A30}$$

Substitution for r_0 and r_1 yields the required result:

$$\begin{aligned} h_3^{-1}(0) = h_2^{-1}(0) - \\ \left[4 + \left(1-y^2\right)w + (1+y)^2 w \exp\left[w + \frac{2}{1+y}\right]\right]^3 \exp\left[w + \frac{2}{1+y}\right] \cdot \\ \left. \frac{\left[-(1-y)(3+w)+(1+y)\left(9+8w+2w^2\right)\exp\left[w+\frac{2}{1+y}\right]\right]}{6(1+y)^2\left[1-y+(1+y)(1+w)\exp\left[w+\frac{2}{1+y}\right]\right]^5} \right|_{w=h_0^{-1}(0)} \end{aligned} \tag{A31}$$

References

1. Langevin, P. Sur la théorie du magnétisme. *J. Phys. Theor. Appl.* **1905**, *4*, 678–693. [CrossRef]
2. Fiasconaro, A.; Falo, F. Analytical results of the extensible freely jointed chain model. *Phys. A Stat. Mech. Appl.* **2019**, *532*, 121929. [CrossRef]
3. Iliafar, S.; Vezenov, D.; Jagota, A. Stretching of a Freely Jointed Chain in Two-Dimensions. *arXiv* **2013**, arXiv:1305.5951.
4. Itskov, M.; Dargazany, R.; Hornes, K. Taylor expansion of the inverse function with application to the Langevin function. *Math. Mech. Solids* **2012**, *17*, 693–701. [CrossRef]
5. Nguessong, A.N.; Beda, T.; Peyraut, F. A new based error approach to approximate the inverse Langevin function. *Rheol. Acta* **2014**, *53*, 585–591. [CrossRef]
6. Darabi, E.; Itskov, M. A simple and accurate approximation of the inverse Langevin function. *Rheol. Acta* **2015**, *54*, 455–459. [CrossRef]
7. Jedynak, R. Approximation of the inverse Langevin function revisited. *Rheol. Acta* **2015**, *54*, 29–39. [CrossRef]
8. Kröger, M. Simple, admissible, and accurate approximants of the inverse Langevin and Brillouin functions, relevant for strong polymer deformations and flows. *J. Non-Newton. Fluid Mech.* **2015**, *223*, 77–87. [CrossRef]
9. Marchi, B.C.; Arruda, E.M. An error-minimizing approach to inverse Langevin approximations. *Rheol. Acta* **2015**, *54*, 887–902. [CrossRef]
10. Rickaby, S.R.; Scott, N.H. A comparison of limited-stretch models of rubber elasticity. *Int. J. Non-Linear Mech.* **2015**, *68*, 71–86. [CrossRef]
11. Petrosyan, R. Improved approximations for some polymer extension models. *Rheol. Acta* **2017**, *56*, 21–26. [CrossRef]
12. Jedynak, R. A comprehensive study of the mathematical methods used to approximate the inverse Langevin function. *Math. Mech. Solids* **2018**, *24*, 1992–2016. [CrossRef]
13. Marchi, B.C.; Arruda, E.M. Generalized error-minimizing, rational inverse Langevin approximations. *Math. Mech. Solids* **2019**, *24*, 1630–1647. [CrossRef]
14. Howard, R.M. Analytical approximations for the inverse Langevin function via linearization, error approximation and iteration. *Rheol. Acta* **2020**, *59*, 521–544. [CrossRef]
15. Mező, I.; Keady, G. Some physical applications of generalized Lambert functions. *Eur. J. Phys.* **2016**, *37*, 065802. [CrossRef]
16. Mező, I. On the structure of the solution set of a generalized Euler–Lambert equation. *J. Math. Anal. Appl.* **2017**, *455*, 538–553. [CrossRef]
17. Mező, I.; Baricz, Á. On the generalization of the Lambert *W* function. *Trans. Am. Math. Soc.* **2017**, *369*, 7917–7934. [CrossRef]
18. Jamilla, C.; Mendoza, R.; Mező, I. Solutions of neutral delay differential equations using a generalized Lambert W function. *Appl. Math. Comput.* **2020**, *382*, 125334. [CrossRef]
19. Petković, M.S.; Petković, L.D.; Herceg, Đ. On Schröder's families of root-finding methods. *J. Comput. Appl. Math.* **2010**, *233*, 1755–1762. [CrossRef]
20. Dubeau, F. Polynomial and rational approximations and the link between Schröder's processes of the first and second kind. *Abstr. Appl. Anal.* **2014**, *2014*, 719846. [CrossRef]
21. Howard, R.M. Schröder based inverse function approximation. *Axioms* **2023**, *12*, 1042. [CrossRef]
22. Corless, R.M.; Gonnet, G.H.; Hare, D.E.G.; Jeffrey, D.J.; Knuth, D.E. On the Lambert W function. *Adv. Comput. Math.* **1996**, *5*, 329–359. [CrossRef]
23. Howard, R.M. Analytical Approximations for the Principal Branch of the Lambert W Function. *Eur. J. Math. Anal.* **2022**, *2*, 14. [CrossRef]
24. Lóczi, L. Guaranteed-and high-precision evaluation of the Lambert W function. *Appl. Math. Comput.* **2022**, *433*, 127406.
25. Schröder, E. Über unendlich viele Algorithmen zur Auflösung der Gleichungen. *Math. Ann.* **1870**, *2*, 317–365. [CrossRef]
26. Berinde, V.; Takens, F. *Iterative Approximation of Fixed Points*; Springer: Berlin/Heidelberg, Germany, 2007.
27. Gdawiec, K.; Kotarski, W.; Lisowska, A. Polynomiography based on the nonstandard Newton-like root finding methods. *Abstr. Appl. Anal.* **2015**, *2015*, 797594. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

AppliedMath

MDPI

Article

Applications of Differential Geometry Linking Topological Bifurcations to Chaotic Flow Fields

Peter D. Neilson [1,*] and Megan D. Neilson [2]

[1] School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, NSW 2052, Australia
[2] Independent Researcher, Late School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, NSW 2052, Australia; megan.neilson@gmail.com
* Correspondence: p.neilson@unsw.edu.au

Abstract: At every point p on a smooth n-manifold M there exist $(n+1)$ skew-symmetric tensor spaces spanning differential r-forms ω with $r = 0, 1, \cdots, n$. Because $d \circ d$ is always zero where d is the exterior differential, it follows that every exact r-form (i.e., $\omega = d\lambda$ where λ is an $(r-1)$-form) is closed (i.e., $d\omega = 0$) but not every closed r-form is exact. This implies the existence of a third type of differential r-form that is closed but not exact. Such forms are called *harmonic* forms. Every smooth n-manifold has an underlying topological structure. Many different possible topological structures exist. What distinguishes one topological structure from another is the number of holes of various dimensions it possesses. De Rham's theory of differential forms relates the presence of r-dimensional holes in the underlying topology of a smooth n-manifold M to the presence of harmonic r-form fields on the smooth manifold. A large amount of theory is required to understand de Rham's theorem. In this paper we summarize the differential geometry that links holes in the underlying topology of a smooth manifold with harmonic fields on the manifold. We explore the application of de Rham's theory to (i) visual, (ii) mechanical, (iii) electrical and (iv) fluid flow systems. In particular, we consider harmonic flow fields in the intracellular aqueous solution of biological cells and we propose, on mathematical grounds, a possible role of harmonic flow fields in the folding of protein polypeptide chains.

Keywords: simplicial complexes; singular homology groups; singular cohomology groups; singular chain complexes; Grassmann algebra; Laplacian and harmonic forms; de Rham cohomology groups; de Rham's theorem; protein folding

MSC: 53Z10; 53Z30

check for updates

Citation: Neilson, P.D.; Neilson, M.D. Applications of Differential Geometry Linking Topological Bifurcations to Chaotic Flow Fields. *AppliedMath* **2024**, *4*, 763–790. https://doi.org/10.3390/appliedmath4020041

Academic Editor: Massimo Ferri

Received: 23 April 2024
Revised: 11 June 2024
Accepted: 11 June 2024
Published: 15 June 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

A well-known quote of Einstein says that as far as the laws of mathematics refer to reality, they are not certain, and as far as they are certain, they do not refer to reality. While this may be seen as controversial, it certainly is the case that mathematical theories of physical processes require interpretation in terms of experimental observations. Differential geometry is a logical structure of provable propositions, theorems and lemmas whose truth derives from axioms, definitions and the elementary connectives of formal logic. To apply differential geometry to physical systems requires the construction of bridges between events and variables in the physical world (preferably, but not necessarily, measurable) and well-defined notions in the deductively logical structure of differential geometry. These bridges are intuitive hypotheses linking reality to mathematics and their value derives from the quality of the intuition of the practitioner. If the hypothesis is true, then certain laws of mathematics can be applied to the real-world system. But even if it is demonstrated experimentally that the mathematical laws do describe the real-world behavior, this does not prove the hypothesis. Other hypotheses may do equally well.

Indeed, it is not possible to prove if-then conjectures, although they can be easily proven false. In applied mathematics, the 'if-then bridges' between reality and mathematics remain forever intuitive. With respect to certainty, it is the mathematics that is primary.

That said and understood, there is value in showing that the same mathematics can apply usefully to a range of real-world systems.

De Rham's cohomology theorem of differential forms is a remarkable piece of mathematics that links together areas of mathematics that were previously thought unrelated. We believe that the mathematical facts provided by de Rham's theorem have great relevance to the nonlinear behavior and the presence of critical points and singularities in the state spaces of many different real-world systems. In this paper, we consider an interpretation of the theorem in terms of experimental observations of the dynamics of fluid flow. In essence, de Rham's cohomology theory relates the existence of holes of various dimensions in the underlying topology of smooth and Riemannian manifolds with a third type of differential form flow field on the manifold, one that is closed but not exact. Such a flow field is called an *harmonic* field. In other words, our aim is to demonstrate the links between the harmonic fields of mathematics and physical flows of real-world nonlinear systems with singularities and critical points observed in nature.

To illustrate the sort of nonlinear dynamical processes in the world that can be understood through this mathematical linking, consider the bubbles and fluid flow associated with pouring a glass of champagne, or the generation and flow of foam associated with an ocean wave breaking on a beach, or the chaotic turbulent flow of a mountain stream around rocks. Even the metamorphosis of a caterpillar into a moth may be within the scope of the theory. Certain singularities in the nonlinear dynamics of visual, mechanical, electrical, and fluid-flow systems can be related to harmonic fields. In particular, we will examine the possible role of vortices and eddies in harmonic flow fields induced by the presence of organelles and other molecules immersed in the intracellular aqueous solution of biological cells on the kinetics of protein folding.

Our search for this linking was inspired by two seemingly unrelated happenings. Firstly, in our analysis of the geometry of 3D binocular visual space [1–3], we came to realize that three different 3D geometries have to be taken into account. There is of course the 3D Euclidean geometry of the outside world. But the geometry of the 3D visual space encoded by sensory receptors is warped (or curved) by the fact that the size of images on the retinas of the eyes vary in inverse proportion to the Euclidean distance between the object in the environment and the nodal point of the eye [4]. This intrinsic 3D space corresponds to a curved Riemannian manifold with a metric proportional to the inverse of the Euclidean distance squared [1]. This gives rise to a singularity at the origin where the metric goes to infinity. The consciously perceived 3D visual space is different again because many different cognitive mechanisms of depth perception are brought into play. The singularity at the origin is taken care of by the existence of a hole about the egocenter in visual space. Existence of such a hole is confirmed by the fact that it is not possible to see your own head, let alone your own egocenter. This raises the question, what effect does such a hole in visual space have on the perceived flow of visual images?

The second event concerned an unusual cluster of breast cancer diagnoses that occurred in women working together at an Australian television studio [5,6]. It was speculated that the cancers might be caused by leakage of electromagnetic radiation from the TV transmitter. Experts were brought in to test this hypothesis. They found no evidence of any problematic radiation levels in the women's workplace. In physical theories of electromagnetism, there are at least two types of electric field. One is given by voltage gradients and the curl of the other is given by changing magnetic fields. These two types of field are independent because the curl of a gradient is always zero. The question arises, is there a mysterious third type of electric field that would have gone undetected by the standard measurement techniques? Differential geometry suggests that there is. According to de Rham's theorem, the curl of an electric field can be zero without it being the gradient

of a voltage. The question becomes, does such a third type of field exist in nature and if so, what properties does it have?

Applying differential geometry to physical systems requires a thorough knowledge of the mathematical theory, because the sequence in which it is applied does not necessarily follow the beautiful logical sequence in which the theorems were developed and proven. Application often requires jumping from one point in the mathematical theory to a distant point with a large amount of mathematics linking the two points. Moreover, an understanding of de Rham's theorem is not trivial. It requires familiarity with a large amount of differential geometry. Consequently, in Appendix A we provide detail of all the steps in our proposed mathematical linking between the appearance and disappearance of holes in the underlying topology of smooth and Riemannian manifolds and the appearance and disappearance of harmonic flow fields on the manifold. Meanwhile, in the following section we provide the reader with a more conceptual introduction to the differential geometry used in relating topological bifurcations to chaotic flow fields in nature.

2. A Summary of Relevant Differential Geometry Concepts

2.1. Topological Structure

An important property of topological spaces is that they are not equipped with a metric. They have no measure of size. Size and topology can be thought of as independent properties. A topological space, like the set of all points on a 2-sphere, can be expanded to the size of a planet or contracted to the size of an atom without changing its topological structure. So long as it is not cut, torn, punctured or any parts of it glued together, it can be deformed in any way by any amount without changing its topological structure. For example, the 2-sphere can be stretched into a long thin rope and tied into a knot without changing its topology.

All the different shapes a topological space can be deformed into without cutting, tearing, puncturing or gluing can be mapped in a continuous, one-to-one, onto and continuously invertible fashion to each other. This is a special type of map called a *homeomorphism*. Any two homeomorphic spaces have all the same topological properties. Indeed, a topological property can be defined as one that is preserved by homeomorphic maps. Transformations that leave the topological structure unchanged are called *homotopies*, a well-known example being the transformation of a coffee mug with a handle into a donut and vice versa [7]. The mathematical mapping ([8], p. 348]) is shown in Figure 1.

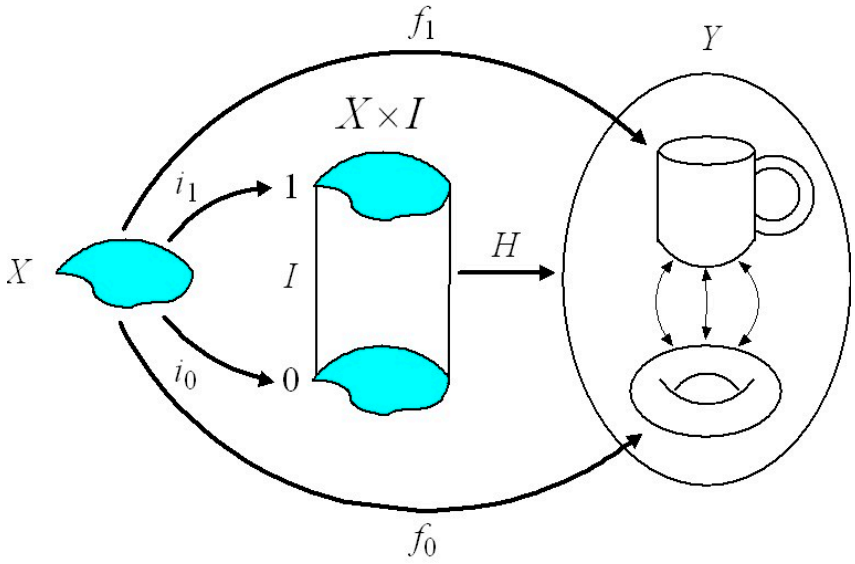


Figure 1. A diagram illustrating homotopy. The 2-torus (surface of a donut) in the topological space Y and the coffee mug with handle in the topological space Y are equivalent topological structures. These are not only homeomorphically related (i.e., continuous, one-to-one, onto and invertible) but

there are continuous homotopic maps $f_0, f_1 : X \to Y$ that map the topological space X to both the 2-torus and the mug. The fact that the donut can be continuously transformed into the mug implies the existence of all the intermediate shapes in the transformation. The hole in the donut morphs into the hole between the mug and its handle. This is illustrated in the diagram by the special case in which $Y = X \times I$, $f_0 = i_0$, $f_1 = i_1$, where the interval I varies from 0 to 1. For each point $x \in X$, the map $f_0(x) = i_0(x) = (x, 0)$ corresponds to the 2-torus, the map $f_1(x) = i_1(x) = (x, 1)$ corresponds to the mug and all the other points along the interval I correspond to the intermediate shapes. This is represented in the diagram by the homotopy H. More generally, while the topological spaces X and Y can include holes of various dimensions, they must have the same topological structure in order for the homotopy H to exist. In other words, for $f_0, f_1 : X \to Y$ to be homotopic maps, the induced homomorphisms $(f_0)_*, (f_1)_* : H_r(X) \to H_r(Y)$ for all $r \geq 0$ between the homology groups for X and Y must be equal, i.e., $(f_0)_* = (f_1)_*$.

There is no way, however, that a 2-sphere can be transformed into a 2-torus without puncturing holes into the sphere. The 2-sphere and the 2-torus do not have the same topological structure. It was shown by A.A. Markov in 1958 (see [8]) that there is no algorithm that can determine all the different topological structures a topological space with dimension 4 or greater can have. The best we can say is that such spaces have a large number of possible different topological structures. Basically, what makes topological structures different is the different number of holes of various dimensions that exist in the topological space (e.g., the difference between the topology of a sphere and the topology of a torus is attributable to the hole(s) in the torus).

2.2. Topological Bifurcations

Based on Smale [9], Abrahams and Marsden [10] provide a description of bifurcations in the underlying topology of a smooth manifold. A smooth map $f : M \to N$ between smooth manifolds M and N is locally trivial at $y_0 \in N$ if there is a neighborhood U of $y_0 \in N$ such that $f^{-1}(y)$ is a smooth properly embedded submanifold (i.e., a level set) in M for each $y \in U$ in N and there are diffeomorphic maps $h|_{f^{-1}(y)} : f^{-1}(y) \to f^{-1}(y_0)$ for each $y \in U$ (i.e., h projects nearby level sets $f^{-1}(y)$ in M diffeomorphically onto the level set $f^{-1}(y_0)$ in M).

The bifurcation set of $f : M \to N$ is $\Sigma_f := \{y_0 \in N | f \text{ fails to be locally trivial at } y_0\}$. In other words, as y passes through Σ_f , $f^{-1}(y)$ changes its topological structure. Let $\sigma(f)$ denote the closed set of critical points x of f in M, i.e.,

$$\sigma(f) = \Big\{ x \in M |\ d_x f : T_x M \to T_{f(x)} N \text{ is not surjective} \Big\}.$$

By Sard's theorem, the singular values $f(\sigma(f))$ in N have measure zero (i.e., they are isolated points with empty interior and their union is nowhere dense in N). The singular values of $f(\sigma(f))$ are the bifurcation values of f in N [10]. In other words, the level sets in M of singular bifurcation values in N have a topological structure that differs from the topological structures of neighboring level sets in M. The number of holes in the level set in M of a bifurcation value in N differs from the number of holes in the neighboring level sets. This can be related to isotropy subgroups, the slice theorem, the tube theorem, the (H)-orbit type submanifold, the (H)-isotropy type submanifold and the (H)-fixed point type submanifold described by Ortega and Ratiu [11] that extend the notion of critical points in submanifolds.

2.3. De Rham's Theory

We provide here only a brief description. A more expanded summary of differential geometry linking bifurcations with harmonic flow fields is given in Appendix A.

Every exact differential r-form ω with $r = 0, 1, \cdots, n$ on a smooth n-dimensional manifold M or an n-dimensional Riemannian manifold (M, g) is closed (i.e., $d\omega = 0$) because $d \circ d$ is always zero. But a closed differential r-form is not necessarily exact. Differential r-forms that are closed but not exact give rise to a third type of differential r-form. There are thus three types of differential r-forms, (i) r-forms ω whose exterior derivative is not zero (i.e., $d\omega \neq 0$), (ii) closed r-forms that are exact differentials (i.e., $\omega = d\lambda$ where λ is an $(r-1)$-form) and (iii) r-forms ω that are closed but not exact. This third type of r-form is called an *harmonic* r-form and is contained in the r-th harmonic space $\mathcal{H}^r(M)$ of the manifold with $r = 0, 1, \cdots, n$. It is also an element in the r-th de Rham cohomology group $H^r_{dR}(M)$ with $r = 0, 1, \cdots, n$. $H^r_{dR}(M)$ is isomorphic with $\mathcal{H}^r(M)$. Both $\mathcal{H}^r(M)$ and $H^r_{dR}(M)$ are vector subspaces as well as being groups.

If ω is a closed r-form (i.e., $d\omega = 0$) on a smooth manifold M, and z_r is an r-dimensional cycle in the underlying topology X of the manifold, then the integral of ω over z_r is called *a period of* ω (the term period derives from the periods of elliptic integrals). From Stokes' theorem, if the r-cycle z_r happens to be a boundary b_r of an $(r+1)$-chain c_{r+1} (i.e., $b_r = \partial c_{r+1}$), then the period vanishes since

$$\int_{z_r} \omega = \int_{b_r} \omega = \int_{\partial c_{r+1}} \omega = \int_{c_{r+1}} d\omega = 0 \text{ (since } d\omega = 0).$$

(See Appendix A.1 for definitions of chains, cycles and boundaries).

From this we obtain de Rham's first theorem: a closed form is globally exact if and only if all of its periods vanish [12]. Knowing which forms are globally exact and which are not has important consequences. For example, (i) a smooth 1-form is conservative (i.e., its line integral is path-independent) if and only if it is exact, (ii) the integral of every globally exact form over a compact oriented smooth manifold without boundary is zero, (iii) the integral of globally closed forms over boundaries is zero and (iv) if the integral of a closed r-form over an r-dimensional submanifold without boundary is not zero, then the r-form is not exact and the submanifold is not the boundary of a compact oriented submanifold with boundary.

A remarkable property of harmonic spaces and de Rham cohomology groups is that they are homotopically and topologically invariant. They are isomorphic to singular cohomology groups $H^r(M)$, and these in turn are isomorphic to singular homology groups $H_r(M)$ on the underlying topological structure of the smooth manifold M (see Appendix A.12). This is depicted in Figure 2.

This is indeed a remarkable property, because it is not immediately obvious that closeness and exactness of differential forms on a smooth manifold should have anything to do with the underlying topological structure of the manifold. But de Rham's theory shows mathematically that they do! The r-th homology group $H_r(M)$ is isomorphic to the r-th de Rham cohomology group $H^r_{dR}(M)$ with $r = 0, 1, \cdots, n$. $H_r(M)$ detects the presence of r-dimensional holes in the underlying topology of the smooth manifold M. The dimension of $H_r(M)$ equals the number of r-dimensional holes in the topology, and this is known as the r-th Betti number b_r of the manifold. Thus, de Rham's theory of differential forms provides the link we are looking for between holes in the topological structure of a smooth or Riemannian manifold and harmonic flow fields on the manifold.

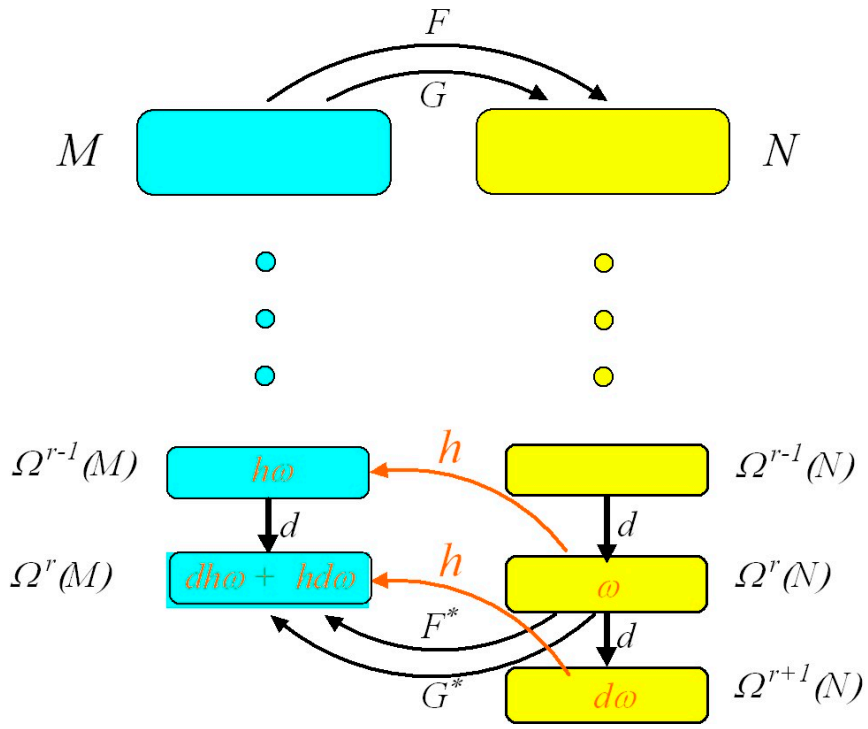


Figure 2. A diagram illustrating the fact that if $F, G : M \to N$ are diffeomorphic maps between smooth manifolds M and N, then the pull-back maps $F^*, G^* : H^r_{dR}(N) \to H^r_{dR}(M)$ between the de Rham cohomology groups $H^r_{dR}(N)$ and $H^r_{dR}(M)$ for all values of r are equal, i.e., $F^* = G^*$. In other words, if the maps F and G are diffeomorphic, they have the same homeomorphic topological structure and so they are homotopy-invariant. This means that homotopy-equivalent manifolds have isomorphic de Rham groups. The labels $\Omega^{r-1}(M)$, $\Omega^r(M)$ represent the skew-symmetric tensor spaces of degree $(r-1)$ and degree (r) for differential forms at each point on the smooth manifold M. The labels $\Omega^{r-1}(N)$, $\Omega^r(N)$, $\Omega^{r+1}(N)$ represent the skew-symmetric tensor spaces of degree $(r-1)$, degree (r) and degree $(r+1)$ for differential forms at each point on the smooth manifold N. The maps $h: \Omega^{r+1}(N) \to \Omega^r(M)$ for all values of r are called homotopy operators. Examination of the diagram shows that if the r-form ω is closed, i.e., $d\omega = 0$, then $d(h\,\omega) = G^*\omega - F^*\omega$. In other words, $G^*\omega$ and $F^*\omega$ differ by an exact differential $d(h\,\omega)$. The maps $F^*, G^* : H^r_{dR}(N) \to H^r_{dR}(M)$ are said to be cohomologous. This is exactly how equal cohomology groups are defined. Therefore, cohomologous groups have isomorphic de Rham cohomology groups.

2.4. The Hodge Laplacian Operator

By definition, the Hodge Laplacian operator on a Riemannian manifold is $\Delta : dd^* + d^*d$, where d is the exterior derivative $d : \Omega^r \to \Omega^{r+1}$, $d^* = (-1)^{rn+n+1} * d*$, where $*$ is the Hodge star operator $* : \Omega^r \to \Omega^{n-r}$ and Ω^r is the skew-symmetric tensor space of r-forms for $r = 0, 1, \cdots, n$ at every point p on an n-dimensional smooth manifold M or a Riemannian manifold (M, g). The Laplacian operator Δ plays an important role in de Rham's theory of differential forms because it is the only operator that can be used to compute harmonic forms on a Riemannian manifold. The harmonic space $\mathcal{H}^r(M) = \ker(\Delta\Omega^r)$ for $r = 0, 1, \cdots, n$. (Keep in mind that harmonic r-forms for all values of r on a smooth manifold or a Riemannian manifold are associated with the presence of r-dimensional holes in the underlying topology of the manifold).

2.5. Harmonic 0-Forms and 1-Forms

Harmonic 0-forms are given by the equation $\Delta\,\varphi = d^*d\,\varphi = div\ grad\ \varphi$, where φ is a smooth real-valued function on the smooth manifold M or Riemannian manifold (M, g). Potential fields in a fluid are irrotational and incompressible (i.e., divergence-free) harmonic 0-forms. We focus here on computing harmonic 1-forms on Riemannian manifolds because

of their importance in the nonlinear dynamics of the physical systems to be discussed below. Let λ be a differential 1-form in $\Omega^1 = T_p^*M$ at each point $p \in (M, g)$. First, we compute the Hodge Laplacian of a 1-form λ on a Riemannian manifold (M, g), i.e., $\Delta\lambda := dd^*\lambda + d^*d\lambda$, and then we compute the harmonic 1-form λ by setting the Laplacian $\Delta\lambda = 0$.

The Ricci curvature tensor can be described geometrically as follows. Let (M, g) be a Riemannian manifold and p a point belonging to (M, g) (i.e., $p \in (M, g)$). For every unit vector $\xi \in T_pM$, $\mathrm{Ric}(\xi, \xi)$ is the sum of the sectional curvatures of the 2-planes in T_pM spanned by $(\xi, \xi_2), \cdots, (\xi, \xi_n)$, where $(\xi_1, \cdots, \xi_n)$ is any g-orthonormal basis for T_pM with $\xi_1 = \xi$. Given any unit vector $\xi \in T_pM$, choose a g-orthonormal basis $(\xi_1, \cdots, \xi_n)$ for T_pM such that $\xi_1 = \xi$. Then $\mathrm{Ric}(\xi, \xi)$ is given by

$$R_{11} = R_{i11}{}^i = \sum_{i=1}^n R_m(\xi_i, \xi_1, \xi_1, \xi_i) = \sum_{i=1}^n \langle(\nabla_{\xi_i}\nabla_\xi - \nabla_\xi\nabla_{\xi_i})\xi, \xi_i\rangle.$$

Following Lang [13] (pp. 423–424), the inner product of the Ricci curvature tensor with the vector ξ (i.e., $\langle \mathrm{Ric}(\lambda), \xi\rangle$) is defined to be the scalar-valued form such that, with respect to an orthonormal frame $\xi_1, \cdots, \xi_n$, and any unit vector field ξ with $\xi_1 = \xi$, we have

$$\langle \mathrm{Ric}(\lambda), \xi\rangle = \sum_{i=1}^n \langle(\nabla_\xi\nabla_{\xi_i} - \nabla_{\xi_i}\nabla_\xi)\rangle\lambda, \xi_i\rangle$$

where we denote by $\langle\lambda, \xi\rangle$ the value of the 1-form λ on the vector field ξ. As an operator on 1-forms, the Laplacian $\Delta : \Omega^1 \to \Omega^1$ is given by $((-\sum_{i=1}^n \nabla^2_{\xi_i}) - \mathrm{Ric})$. Written in terms of variables, this means $\langle\Delta\lambda, \xi\rangle = -\sum_{i=1}^n \langle\nabla_{\xi_i}\nabla_{\xi_i}\lambda, \xi\rangle - \sum_{i=1}^n \langle(\nabla_\xi\nabla_{\xi_i} - \nabla_{\xi_i}\nabla_\xi)\lambda, \xi_i\rangle$.

To compute the Hodge Laplacian inner product $\langle\Delta\lambda, \xi\rangle = \langle(dd^* + d^*d)\lambda, \xi\rangle$, we compute (i) $\langle dd^*\lambda, \xi\rangle$ and (ii) $\langle d^*d\lambda, \xi\rangle$ separately and then add:

(i) We have $d^*\lambda = -\mathrm{div}\lambda^\vee = -tr(\nabla\lambda) = -\sum_{i=1}^n \langle\nabla_{\xi_i}\lambda, \xi_i\rangle$, where $\lambda^\vee$ is the dual of the 1-form field λ and ξ_i for $i = 1, \cdots, n$ is a g-orthonormal frame of vector fields. These orthonormal vectors have to be computed using the Riemannian metric Gram–Schmidt orthogonalizing algorithm at each point. We see that $d^*\lambda$ is a real-valued function $f = -\sum_{i=1}^n \langle\nabla_{\xi_i}\lambda, \xi_i\rangle$ on (M, g). For each point $p \in (M, g)$, we have $\nabla_\xi f = \xi f = \langle df, \xi\rangle$. Thus we can write

$$\nabla_\xi(-\sum_1^n \langle\nabla_{\xi_i}\lambda, \xi_i\rangle) = \langle d(-\sum_{i=1}^n \langle\nabla_{\xi_i}\lambda, \xi_i\rangle), \xi\rangle = \langle dd^*\lambda, \xi\rangle = -\sum_{i=1}^n \langle\nabla_\xi\nabla_{\xi_i}\lambda, \xi_i\rangle. \quad (1)$$

(ii) To obtain $d^*d\lambda$, we first note that if ω is an r-form, we can use the proposition

$$d\omega(\xi_0, \xi_1, \cdots, \xi_r) = \sum_{i=0}^r (-1)^i \nabla_{\xi_i}\omega(\xi_1, \cdots, \xi_{i-1}, \xi_0, \xi_{i+1}, \cdots, \xi_r) \text{ to obtain}$$

$$d\lambda(\xi, \eta) = \langle\nabla_\xi\lambda, \eta\rangle - \langle\nabla_\eta\lambda, \xi\rangle \quad (2)$$

where λ is a 1-form and ξ and η are vectors. Using the fact that for any r-form ω

$$d\omega = \sum_{i=0}^r (\lambda_i \wedge \nabla_{\xi_i}\omega) \text{ and } d(*\,\omega) = \sum_{i=1}^n (\lambda_i \wedge \nabla_{\xi_i}(*\,\omega))$$

where ξ_i for $i = 1, \cdots, n$ is an orthonormal frame and λ_i for $i = 1, \cdots, n$ is an orthonormal coframe, it follows that

$$d^*\omega = *d*\omega = \sum_{i=1}^n (\nabla_{\xi_i}\omega \circ \xi_i) \quad (3)$$

where $(\nabla_{\xi_i}\omega \circ \xi_i)$ is notation for $\nabla_{\xi_i}\omega$ contracted by ξ_i [13].
From Equations (3) and (2) with $\eta = \xi_i$, we obtain [13] (p. 424)

$$\langle d^*d\lambda, \xi\rangle = \sum_{i=1}^n \langle\nabla_{\xi_i}\nabla_\xi\lambda, \xi_i\rangle - \sum_{i=1}^n \langle\nabla_{\xi_i}\nabla_{\xi_i}\lambda, \xi\rangle \quad (4)$$

Finally, we add Equations (1) and (4) to obtain

$$\langle\Delta\lambda, \xi\rangle = \langle(dd^* + d^*d)\lambda, \xi\rangle = -\sum_{i=1}^n \langle\nabla_{\xi_i}\nabla_{\xi_i}\lambda, \xi\rangle - \sum_{i=1}^n \langle(\nabla_\xi\nabla_{\xi_i} - \nabla_{\xi_i}\nabla_\xi)\lambda, \xi_i\rangle. \quad (5)$$

Thus, for a 1-form on an n-dimensional Riemannian manifold (M, g), the Hodge Laplacian operator has the form $\Delta = (-\sum_{i=1}^{n} \nabla^2_{\xi_i}) - \mathrm{Ric}$, where Ric is the Ricci curvature operator. The harmonic 1-form is then obtained by setting the Laplacian $\Delta\ \lambda$ equal to zero.

3. Physical Examples

There exists an endless variety of behaviors in nature of nonlinear dynamical systems associated with different singularities, critical points and degeneracies in their differential equations. Most of these behaviors have not yet been described, let alone solved. One family of such behaviors considered in this paper is associated with the existence of holes of various dimensions in the underlying topology of the state space of the system. Such holes create differential forms and vector fields that are closed but not exact and create 'no-go' spaces in the dynamical state of the system. Such holes can dramatically change the dynamical behavior of the system. Isotropy subgroups in the action of a Lie group on a dynamical system, for example, create dynamic states that a vector field (or subset of vector fields) can flow into but not out of; a type of equilibrium point dynamical-state trap. But nonlinear dynamical systems can transition to neighboring vector fields and flow around the trap, creating novel nonlinear behaviors [11]. In what follows, we consider the consequences on dynamical behavior of singularities in the differential equations that give rise to differential forms that are closed but not exact in visual, mechanical, electrical and fluid-flow physical systems.

3.1. The Human Visual System

Human binocular 3D visual space encoded by visual, proprioceptive and vestibular sensory receptors is warped (or curved) because the size of 2D images on the retina of each eye varies in inverse proportion to the Euclidean distance between the nodal point of the eye and the object in the environment. Geometrically speaking, 3D visual space corresponds to an egocentric negatively curved Riemannian manifold with a 3D ball of points missing from about the origin (i.e., egocenter) [1,2]. This intrinsic hole corresponds to the fact that it is not possible to see your own head. What influence does such a hole about the origin in 3D warped visual space have on the perceived flow of visual images? Differential geometry and de Rham's theorem provide a possible answer to this question. (Note that other holes in 3D visual space are possible due to neural pathology. Known as scotomas, these are not considered here.)

Given a 2D hole about the origin (i.e., a missing 3D ball of points about the origin enclosed within a 2-cycle), de Rham's theorem implies that closed differential r-forms for $r = 0, 1, 2, 3$ are not globally exact. The unit volume form $\Omega = dx^1 \wedge dx^2 \wedge dx^3$ is a closed 3-form on the warped 3D visual space, and therefore the hole about the origin prevents this volume 3-form from being globally exact. This is just as well, because if there was no hole about the origin and the closed volume form Ω was globally exact, then all the periods $\int_{c_3} \Omega$ equal to the volume form integrated over all 3-chains (i.e., volumes) enclosed by 2-cycles would vanish. The underlying topology of the warped 3D visual space would be such that the perceived volume $\int_{c_3} \Omega$ of any enclosed 3D space would equal zero. In other words, it would not be possible to sense the size (i.e., volume) of enclosed 3D spaces. We know from everyday experience that we can visually sense the size of the room we are in. This is as predicted by de Rham's theorem.

3.2. Mechanical Systems

In the absence of all external forces, Hamiltonian mechanics plays out on the $2n$-dimensional cotangent bundle $T^*Q = P$ (or momentum phase space) of an n-dimensional configuration manifold Q. Using position (configuration) q and momentum p as canonical coordinates, a canonical 1-form $\theta = p_i dq^i$ and a canonical closed 2-form $\omega = dq^i \wedge dp_i$ exist at every point (q, p) on $T^*Q = P$ (i.e., on the momentum phase space of the Hamiltonian system). The symplectic structure of the 2-form $\omega = dq^i \wedge dp_i$ defines a symplectic geometry on the system. Contraction of the symplectic 2-form field ω by a vector field ξ

gives a 1-form field $\lambda = C_{\xi}\omega$ on $T^*Q = P$. However, there is a wonderful discovery that applies to all Hamiltonian mechanical systems (and to all systems whose dynamics can be described by 2nd-order differential equations). There exists a special vector field ξ_h for which the 1-form field $\lambda = C_{\xi_h}\omega$ is an exact 1-form field equal to dh, where h is a smooth real-valued function on the momentum phase space $T^*Q = P$. The real-valued function h is called the Hamiltonian function, and the vector field ξ_h is called the Hamiltonian vector field of the Hamiltonian mechanical system. If the real-valued function h is equal to the total energy of the mechanical system (i.e., equal to the sum of the kinetic energy and potential energy of the system expressed as a function of position and momentum), then the integral flow of the Hamiltonian vector field ξ_h in the momentum phase space $T^*Q = P$ is equal to the free motion of the mechanical system. Since by definition there is no energy-injecting or energy-dissipating external force, the integral flow of the Hamiltonian vector field ξ_h conserves the total energy h of the system. The integral flow is confined to a $(2n-1)$-dimensional constant energy (i.e., constant h) contour shell in $T^*Q = P$. Different energy levels correspond to flows in different constant energy contour shells in the momentum phase space of the system.

However, if there exists a 1-dimensional hole in the momentum phase space, then, from de Rham's theorem, the 1-form $\lambda = C_{\xi}\omega$ cannot be globally exact, and hence neither the Hamiltonian function h nor the Hamiltonian vector field ξ_h can be globally defined! The exact 1-form field $\lambda = dh$ is replaced by an harmonic 1-form field

$$\lambda \in \mathcal{H}^1(P) \cong H^1_{dR}(P)$$

that is closed but not exact.

3.3. Electrodynamical Systems

In electrodynamics on Minkowski 4D spacetime, Maxwell's equations can be written in terms of differential forms as

$$\begin{gathered} *F^{\mu\nu}_{,\nu} = F_{\mu\nu,\rho} + F_{\nu\rho,\mu} + F_{\rho\mu,\nu} = 0 \text{ (source free)} \\ F^{\mu\nu}_{,\nu} = \tfrac{4\pi}{c} J^{\mu} \text{ (source coupled)}, \end{gathered}$$

where F is a skew-symmetric electromagnetic 2-form at each point in 4D Minkowski spacetime, J is a current 4-vector at each point, c is the speed of light in a vacuum and $*$ is the Hodge star operator (all in Gaussian units).

The source-free equations are true if and only if there exists a 4-covector field A (i.e., a 1-form field) called the electromagnetic potential field such that $F = dA$ (i.e., F is a closed, globally exact 2-form field on 4D flat Minkowski spacetime).

Such a 1-form field A can be altered by the addition of an exterior differential $d\psi$ of a smooth real-valued function ψ on the 4D Minkowski manifold (i.e., $\widetilde{A} = A + d\psi$) without changing the closed 2-form electromagnetic field F. F is unchanged because $d \circ d = dd$ is always zero, and F can be written as

$$F = d\widetilde{A} = dA + dd\psi = dA$$

Any such transformation of the electromagnetic potential A that has no effect on the electromagnetic field F is called a *gauge transformation*. Any choice of ψ that satisfies the inhomogeneous wave equation is called a *Lorentz gauge*; i.e.,

$$\begin{gathered} \Box\psi = -A^{\mu}_{,\mu} \text{ so that} \\ \widetilde{A}^{\mu}_{,\mu} = A^{\mu}_{,\mu} + \Box\psi = A^{\mu}_{,\mu} - A^{\mu}_{,\mu} = 0 \end{gathered}$$

where the operator $\Box$ is the wave operator or the d'Alembertian

$$\Box\psi = \nabla^2\psi - \frac{1}{c^2}\frac{\partial^2\psi}{\partial t^2}$$

This shows that an electromagnetic potential A in the Lorentz gauge can carry an electromagnetic wave $\Box\psi$ propagating at the speed of light c in a vacuum without changing the electromagnetic field F. The electromagnetic field F is the medium in which the wave propagates with no need of an aether. It is important to notice that

$$\Box\psi = -A^{\mu}_{,\mu} = -\text{div} \circ \text{grad}\ \psi = -\nabla^2\psi = -\Delta\psi,$$

i.e., the d'Alembertian wave operator $\Box$ on 4D flat Minkowski spacetime is equal to the Laplacian Δ on a flat 4D spacetime.

However, if a 1-dimensional hole exists in the 4D spacetime manifold, then, from de Rham's theorem, the electromagnetic potential A cannot be globally exact, and consequently the Lorentz gauge cannot be defined. The global electromagnetic potential 1-form field A is replaced by an harmonic 1-form field. This leaves the question about the existence of an aether uncertain.

3.4. Fluid Flow

The principles of fluid mechanics have long been a standard in the teaching of physics [14–16]. Consider a slowly flowing stream of water. Water particles close to the boundary adhere to the boundary and so have zero downstream velocity. Water molecules in this boundary layer form hydrogen bonds with each other that increase surface tension and alter the dipole distribution of electric charge. Points away from the boundary towards the center of the stream have higher average downstream flow velocities. This pattern of flow is known as *lamina flow*. The random motion of water particles due to thermal energy and electric fields causes an interchange of water particles between the various laminae, with the higher-velocity particles speeding up the slower ones and vice versa. This gives rise to shear stresses and transfer of kinetic energy and momentum between the various layers producing a property of the water known as its dynamic viscosity μ. Viscosity varies between different fluids and provides a measure of the ease of flow. Honey, for example, has a higher viscosity than water and flows less easily.

Now consider a fast-flowing mountain stream flowing around rocks. Water particles cannot flow through the rocks, so they have to flow around them. The rocks form 2D holes of various sizes in the space of the water (i.e., the 2D surface of each rock is a 2-cycle that forms a boundary about a hole in the space of water flow). If the stream is flowing fast enough (due to an ample supply of kinetic energy provided by gravity), flow around the rocks induces large eddies and vortices in the flow. These large eddies propagate outwards, transferring kinetic energy to smaller eddies and vortices. These in turn generate even smaller eddies. This cascade of kinetic energy from larger to smaller eddies depends on the velocity of flow (i.e., available kinetic energy) and the rate of energy dissipation in the water due to viscosity. The eddies of various sizes appear and disappear in an unpredictable fashion, causing local tangential velocities and pressures to vary from point to point in a chaotic unpredictable fashion. This increases the amount of mixing. Increased local flow velocities reduce local pressure, allowing gases in the water to vaporize and produce bubbles. Formation and popping of bubbles dissipate energy and produce a 'gurgling' sound. The chaotic pattern of flow associated with such a cascade of energy from large to small eddies with minimal energy dissipation is called *turbulent flow*. Turbulent flow is observed in many fluid flows in nature, such as ocean waves, storm clouds, smoke, bubbles in champagne and more.

3.4.1. Navier–Stokes Equations

The Navier–Stokes equations are nonlinear partial differential equations that describe the motion of viscous fluids in 3D Euclidean space. They were developed over many years during the first half of the 1800s by Navier and Stokes. In essence, they are an expression of Newton's second law of motion; that is, force equals mass times acceleration. The

incompressible Navier–Stokes equations can be derived from the convective form of the Cauchy momentum equation and have the form

$$\nabla \cdot u = 0 \tag{6}$$

$$\partial u/\partial t + (u \cdot \nabla)u = -\frac{1}{\rho}\nabla p + \mu \nabla^2 u + F. \tag{7}$$

Equation (6) is a continuity equation expressing conservation of momentum and applies to incompressible fluids. Solution of the equations gives the flow velocity u as a vector field over the 3D space of the fluid. The vector u gives the direction and magnitude of the velocity of the fluid at every point in the 3D space of the fluid at every moment in time. $\partial u/\partial t$ is time acceleration. $(u \cdot \nabla)u$ or $u \cdot (\nabla u)$ is spatial or convective acceleration (e.g., eddies and vortices). Convective acceleration is nonlinear, describing the change in velocity u with position. It makes the equations difficult to solve and is the predominant cause of turbulence. ∇u is the velocity gradient. It is a Jacobian matrix or a second-order tensor. Its symmetrical component $1/2(\nabla u + (\nabla u)^{\mathrm{T}})$ describes stretching and shearing, while its skew-symmetric component $1/2(\nabla u - (\nabla u)^{\mathrm{T}})$ describes the rate of rotation. ρ is fluid density, while ∇p is the pressure gradient. $\mu \nabla^2 u$ is the diffusion divergence of shear stress due to viscosity, and μ is the dynamic viscosity of the fluid. The symmetrical shear stress tensor τ equals $\mu(\nabla u + (\nabla u)^{\mathrm{T}})$, and the divergence of τ is given by

$$2\mu \nabla \cdot \tau = 2\mu \nabla \cdot (1/2(\nabla u + (\nabla u)^{\mathrm{T}})) = \mu \nabla^2 u = \mu \Delta u \tag{8}$$

where $\Delta u = \nabla^2 u$ is the Laplacian of u in Euclidean space. F represents all the external forces acting on the fluid, including gravity and electromagnetic forces.

The Navier–Stokes equations are difficult if not impossible to solve analytically. However, computer simulation studies using simplifying assumptions such as temporal and spatial averaging or band-pass filtering do a good job of matching simulated fluid flows with those measured experimentally [16].

3.4.2. Reynolds Number

The Reynolds number is a dimensionless number defined by Osborne Reynolds (1842–1917) to be

$$\mathrm{Re} = \frac{inertial\ forces}{viscous\ forces} = \frac{\rho u L}{\mu}$$

where ρ = density, u = velocity, L = characterisitc length (e.g., size of rocks in a mountain stream) and μ = dynamic viscosity.

The Reynolds number plays an important role in interpreting the Navier–Stokes equations and de Rham's theorem because as Re is increased experimentally from a small number to a large number, fluid flow is observed to undergo a transition from lamina flow to turbulent flow. The transition can be divided into stages depending on the size of the Reynolds number. When viscous forces due to shear stresses in the fluid dominate the inertial forces, the Reynolds number is small (e.g., Re for honey ≈ 0.001). When Re is small, fluid flow is dominated by energy dissipation due to the viscous shear stresses, and consequently eddies and vortices are strongly damped. The result is a stable lamina flow. As Re is increased, outward-flowing 3D pressure waves are observed between lamina layers. This increases mixing and increases energy-momentum transfer between layers. For large Reynolds numbers, inertial forces dominate and viscous damping is minimal. Turbulent flow results from the relatively low level of energy dissipation and damping. When injection of kinetic energy exceeds energy dissipation, instabilities in nonlinear convective interactions create a cascade of energy transfer from large to small eddies and vortices.

For very large Reynolds numbers, the diffusion divergence of shear stress due to viscosity $\mu \nabla^2 u$ in the Navier–Stokes equation effectively can be ignored. Extreme turbulent

flow results, and the cascade of energy produces chaotic rotational flows at the small scale of individual molecules. Ignoring $\mu\nabla^2 u$ is equivalent to setting the Laplacian of the velocity vector field in Euclidean space equal to zero (i.e., $\Delta u = \nabla^2 u = 0$), where $\nabla^2 u$ equals the Laplacian Δu in Euclidean space where the Ricci curvature is zero. According to de Rham's cohomology theory, such a velocity vector field is a global harmonic flow field associated with the existence of holes in the underlying topology of the 3D Euclidean fluid space. Given an ample supply of kinetic energy, the existence of holes (such as particles, rocks and bubbles) in the space of fluid flow is associated with extreme turbulence, due to the lack of energy dissipation in the fluid and the presence of chaotic rotational kinetic energy at molecular levels.

4. A Theoretical Proposal

4.1. Biological Cells

Biological processes regulate human body temperature to 37 °C. There is thus an ample supply of thermal kinetic energy for motion of aqueous solution within biological cells. Biological cells are densely packed with organelles and other molecules. Intracellular fluid cannot flow through these organelles, so it has to flow around them. In other words, the 3D space of the intracellular aqueous solution contains a large number of holes of various shapes and sizes. As described above, de Rham's theorem shows that the presence of such holes in the underlying topology of the 3D space of a fluid has a dramatic influence on the vector and 1-form flow fields. Based on mathematical theory, we propose that the thermal motion of intracellular fluid around organelles creates a transfer of kinetic energy with minimal energy dissipation from large to very small eddies and vortices at a molecular level, and that these rotations are appropriately scaled to play a role in protein folding.

4.2. Protein Folding

The study of protein folding began some 60 years ago. Much is known, but much is yet to be learned [17–19]. Proteins are polypeptide chains of amino acids assembled by ribosomes within biological cells. Hundreds of thousands of atoms can be involved in long peptide chains. The single-file order of amino acid side chains in each peptide chain is specified by codons within mRNA molecules. The sequence of amino acids determines the 3D structure that the peptide chain spontaneously folds into. Folding happens within milliseconds and takes place even as the peptide chain is being assembled by a ribosome. The native 3D conformation of the peptide chain is critically important. It determines the function of the protein. Different functions of thousands of different proteins provide the molecular machinery for living systems. If the proteins are not folded correctly, then they do not function properly, causing many disorders such as amyotrophic lateral sclerosis, sickle cell anemia, Alzheimer's disease and mad cow syndrome, to name a few. But what are the mechanisms controlling the folding of proteins and what causes it to go wrong?

Given the extremely large number of possible bonding angles between the atoms and molecules within the peptide chain, the fact that correct folding happens very quickly indicates that it is not a random trial-and-error process. There must be a physical process directing the folding pathway. There is also a paradox. The large number of possible configurations of the unfolded peptide chain folds into a unique 3D structure of the functioning protein. This violates the second law of thermodynamics. The conformational entropy of the protein decreases as its folded structure becomes more ordered!

The paradox can be resolved by considering the intracellular aqueous solution in which the peptide chains are immersed. This is known to play a crucial role in protein folding. If a folded protein is immersed in urea (which acts as a detergent), it unfolds, and it refolds correctly when returned into intracellular fluid. Ion concentrations, pH and temperature of the intracellular fluid have to be carefully regulated to appropriate levels by biological processes. If the regulation fails, the ribosome-assembled peptide chains do not fold correctly. A prevailing thermodynamic theory for protein folding invokes the role for intracellular aqueous solution as follows. It proposes that the lack of polarization of

hydrophobic amino acid side chains distributed along the peptide chain cause the water molecules surrounding these side chains to become structured, driven by their electrical dipoles and hydrogen bonding between water molecules. Thermodynamic free-energy gradients cause the peptide chain to fold so as to enclose its hydrophobic side chains within its core, thereby hiding them from the surrounding water molecules. This releases the water molecules from the ordered structure imposed by the hydrophobic side chains, thereby increasing the entropy of the water. The increase in the entropy of the water exceeds the decrease in the configurational entropy associated with protein folding, so spontaneous folding is thermodynamically feasible.

In addition to thermodynamic free-energy gradients within the intracellular fluid, there are many other chemical and electrical potential energy fields surrounding the peptide chain that can influence its folding, including covalent, ionic and hydrogen bonds, voltage gradients and van der Waals forces between atoms in the protein. Different configurations of the peptide chain bring attractive sites along the chain into alignment. These alignments allow the formation of hydrogen bonds, disulphide bonds, salt bridges and others. Associated potential energy fields drive the folding of the protein towards the minimum energy configuration specified by the sequence of amino acids in the peptide chain. There is no guarantee, however, that a particular set of molecular bonds and salt bridges will correspond to the functional minimum-energy configuration. If it is only a local minimum-energy configuration, then work has to be done on the molecule to push its configuration out of the local energy trap and into a deeper energy valley.

4.3. The Proposal

It does not seem unreasonable to suggest that the jiggling (Brownian motion) of atoms and molecules in peptide chains caused by motion of the intracellular aqueous solution plays an important role in the kinematics of protein folding. De Rham's theorem shows that holes in the underlying topology of the intracellular fluid are associated with global *harmonic* flow fields in the 3D space of the fluid. The experimental study of fluid flow together with Reynolds number, the Navier–Stokes equations and de Rham's theorem provide a connection (consistent with experimental observation) among holes, extreme turbulent flow and harmonic flow fields. The term $\mu\nabla^2 u = \mu\Delta u$ in the Navier–Stokes equation describes the diffusion divergence of shear stress due to dynamic viscosity μ of the fluid. Experimental studies show that large ratios of inertial forces to viscous forces (i.e., large Reynolds number) are associated with extreme turbulent flow, and the term $\mu\nabla^2 u = \mu\Delta u$ in the Navier–Stokes equation effectively can be ignored. This is equivalent to setting the Laplacian of the velocity field Δu equal to zero. In differential geometry, this is exactly how harmonic vector and 1-form fields are defined (i.e., in Euclidean space $\Delta u = \nabla^2 u = 0$). In other words, experimental observations of fluid flow imply that harmonic flow fields associated with flow around organelles and other molecules (including protein molecules) correspond to extreme turbulent flow. Existence of holes in the 3D fluid space prevents closed 1-form and vector flow fields from being globally exact and replaces them with 1-form harmonic flow fields. In other words, given an ample supply of kinetic energy, de Rham's theory provides a formal description of intracellular fluid flow responsible for the Brownian motion of molecules in proteins and other structures within biological cells.

As described in Section 3.4, mathematically speaking, harmonic flow fields associated with holes in the underlying topology of the fluid space can be computed by equating the Laplacian of the flow field to zero (i.e., $\Delta\lambda := dd^*\lambda + d^*d\lambda = 0$). But in Euclidean space, the Laplacian of a vector flow field is given by $\Delta u = \nabla^2 u$, and according to the Navier–Stokes Equations (6) and (7), setting $\Delta u = \nabla^2 u = 0$ is equivalent to the absence of viscous shear stress and absence of energy dissipation in the fluid. Extreme turbulent flow results. An ample supply of thermal kinetic energy with negligible energy dissipation in the fluid causes a cascade of kinetic energy transfer from large eddies and vortices to smaller and smaller ones until finally a large amount of kinetic energy is available in

the form of chaotic 3D rotations of individual water molecules. Remembering that water molecules form hydrogen bonds to each other in the neighborhood of hydrophobic amino acid side chains and adhere to hydrophilic amino acid side chains, it seems reasonable to propose that this kinetic energy is available to do work (i.e., inject energy) on the atoms and molecules of protein polypeptide chains.

The rotational torques applied to the protein molecules are at the right energy level and scale to operate on the bonding angles between atoms and molecules within the peptide chain. Moreover, they operate simultaneously on all parts of the peptide chain and, when combined with thermodynamic, chemical and electrical potential energy gradients surrounding the peptide chain, can play an important role in protein folding. They not only supply the energy needed to break away from local energy traps but also can accelerate folding towards minimum energy configurations. Like 'annealing' in minimum-variance search algorithms for neural networks, the chaotic harmonic flow fields within intracellular fluids enable the folding to find the unique absolute minimum energy configuration of the protein. Of course, this requires the appropriate biologically regulated level of available energy. If body temperature is too low, the intracellular fluid flow has insufficient energy to break peptide chains away from local energy traps, and if body temperature is too high, it can break bonds and denature proteins.

A nice analogy for the Brownian motion of protein molecules and side chains caused by turbulent flow of intracellular fluid about the densely packed organelles within the cell is given by the chaotic movement of the leaves and branches of a tree generated by the flow of wind through the tree.

5. Discussion

To recap, de Rham's theorem shows that holes in the underlying topology of a smooth manifold are associated with closed differential form fields on the manifold that are not exact. Known as harmonic form fields, these can be computed on a Riemannian manifold by setting their Hodge Laplacian $\Delta := dd^* + d^*d$ equal to zero, and on a Euclidean space by setting the Laplacian $\Delta = \nabla^2$ equal to zero. The question is, do such harmonic form fields exist in nature and, if so, what properties do they possess?

The mathematics tells us that the integral flow fields on a sphere, for example, will be very different from the flow fields on a torus, the difference being attributable to the hole in the torus. Likewise, for the physical systems described in Section 3, the same mathematics predicts *with certainty* that the behavior of vector and differential form fields on their smooth manifolds should be profoundly altered by the appearance of any hole or holes in the underlying topology of the manifold. Do we have experimental evidence that this is the case? And what are the implications for modeling these systems?

5.1. Mechanical Systems

In Section 3.2, we showed mathematically that, according to de Rham's theorem, the presence of holes in the momentum phase space of a mechanical system prevents the Hamiltonian function h and the Hamiltonian vector field ξ_h from being globally defined on the phase space. However, despite this being a definitive mathematical prediction, until appropriate experimental data become available, it remains uncertain as to how exactly a physical mechanical system with holes in its momentum phase space might behave. What type of flow field will replace the globally exact 1-form field $\lambda = C_{\xi_h}\omega$ in a Hamiltonian system is yet to be discovered. In other words, de Rham's theorem implies that changes in the topology of the state space of a mechanical system require a different mathematical model from that of Hamiltonian mechanics.

This problem applies similarly to Lagrangian mechanics. While Hamiltonian mechanics plays out on the cotangent bundle (T^*Q) over the configuration manifold (Q), Lagrangian mechanics plays out on the tangent bundle (TQ) over the configuration manifold (Q). Legendre's transformation $FL = TQ \rightarrow T^*Q$ maps points in TQ to points in T^*Q. Lagrangian and Hamiltonian mechanics are often described as being dual to each

other because, if the equations are everywhere regular, both derive the same trajectory of motion in the configuration manifold Q. However, duality between Lagrangian and Hamiltonian systems requires the Legendre transformation $FL(\dot{q}^i) = \frac{\partial L}{\partial \dot{q}^i} \rightarrow p_i$ to be a smooth one-to-one diffeomorphism, where $L(q, \dot{q})$ is the Lagrangian equal to the difference between the kinetic energy and potential energy, $\dot{q}^i$ is the $i - th$ component of the velocity vector $\dot{q}$ and p_i is the $i - th$ component of the momentum vector p. Duality also requires that both the symplectic Lagrangian second differential form on TQ and the Hamiltonian symplectic second differential form on T^*Q be nondegenerate. However, many real-world mechanical systems do not satisfy these requirements. Indeed, it is this fact that underlies the importance of Emmy Ntre's theorem; namely, if a Lagrangian $L(q, \dot{q})$ is degenerate and does not vary along a coordinate q^i, then the associated component of momentum p_i is conserved. Removing degeneracies and singularities from the manifolds introduces holes, and these, as shown by de Rham's theory, change the behavior of the mechanical system away from that predicted by classical Lagrangian and Hamiltonian mechanics. As far as we know, the modeling of real-world mechanical systems to account for holes in the manifold is yet to come. Hopefully this setting out of the theoretical background can assist in stimulating the collection of the necessary experimental data.

5.2. Electrodynamical Systems

In Section 3.3, we showed mathematically that the presence of holes in Minkowski 4D spacetime prevents Maxwell's source-free equations from being true. The closed 2-form electromagnetic field F is prevented from being globally exact, and the electromagnetic 1-form potential field cannot be defined globally. Consequently, global gauge transformations of the potential field, including the Lorentz gauge, cannot exist. Will the 1-form electromagnetic field be replaced with a harmonic 1-form field and, if so, what properties will it possess? Again, appropriate experimental observations of physical electrodynamical systems are required to answer these questions.

Indeed, the implications go to the very nature of electromagnetic radiation. The churning motion of charged particles in the sun creates changing electric currents J and changing charge densities ρ. But, as shown by Maxwell's equations of electromagnetism, J and ρ are exactly the energy sources for magnetic fields B and electric fields E. In the vacuum of space between the sun and the earth, J and ρ are both zero. Nevertheless, using the curl of the curl relationship and the div and curl relationships between E and B in Maxwell's equations, even with both J and ρ set equal to zero, the coupled E and B fields satisfy wave equations $\nabla^2 E - 1/c^2 \partial^2 E/\partial t^2$ and $\nabla^2 B - 1/c^2 \partial^2 B/\partial t^2$. In other words, E and B form coupled electromagnetic waves that can travel at the speed of light c through a vacuum despite both J and ρ being zero.

But in physics there remain unresolved questions. By analogy with waves in water and sound waves in air, it can be seen that waves require a medium in which to travel. But what is the medium in which an electromagnetic wave in a vacuum travels? In the late 1800s, it was proposed by Lorentz that something called the *aether* filled all of space and that electromagnetic waves travel at the speed of light in the aether. However, the 1887 interferometer experiment of Michelson and Morley showed that there was no difference in the speed of light coming from a star when the earth was moving towards the star compared to when the earth was moving away from the star. This challenged the idea that both the light and the earth are moving in space relative to a fixed aether. More recently, the mathematical approach to particle physics known as *gauge theory* offers an alternative explanation, namely that the electric and magnetic fields themselves provide the medium in which light travels. Yet as described in Section 3.3, de Rham's theory of differential forms shows that gauge theory itself fails to hold in the vicinity of holes in the manifold. And this is just what the presence of stars and planets create in the fabric of spacetime.

Life on earth depends fundamentally on the ability of electromagnetic waves to carry energy through empty space. Maxwell was quick to assert that light is an electromagnetic

wave. Via photosynthesis, plants absorb energy from the light that falls upon them, transforming this into chemical energy stored in sugar molecules. This stored chemical energy in turn provides most of the biological energy needed to sustain life on earth. The understanding of electromagnetic fields and the consequences of the occurrence of holes within them is therefore an important pursuit.

5.3. The Human Visual System

For both mechanical and electrodynamic systems, the implications of de Rham's theorem are yet to be addressed by experiment. For the human visual system examined in Section 3.1, however, the situation is somewhat different. In this case, we have extra relevant information derived from our own conscious awareness. We know, for example, that there is a hole about the egocenter in our perceived binocular 3D visual space because we are unable to see our own head. We also know that the size of a perceived visual image increases as the Euclidean distance between the egocenter and the object in the environment decreases. This is consistent with experimental data showing that the size of an image on the retina changes in inverse proportion to the Euclidean distance between the nodal point of the eye and the object in the environment [4]. This creates a singularity at the egocenter in 3D visual space where the size of the image on the retina becomes infinite. But this singularity is not a problem, because it is contained within the hole about the egocenter. De Rham's theorem shows that such a hole should prevent perceived image flow fields from being globally exact, replacing them with harmonic flow fields.

The question as to whether or not such harmonic flow fields exist in the human visual system and what properties they possess can be answered to some extent from personal experience. As an object in the outside world is brought increasingly close to the egocenter, we can keep it in the binocular visual field by converging the eyes. The perceived size of the image gets bigger and bigger and starts to exceed the size of the perceived visual space. Eventually, we reach a limit to cross-eyed vision. The perceived visual image becomes defocused and disintegrates in a chaotic fashion. It can no longer be perceived as a coherent single image. It becomes physically impossible to bring the object any closer to the egocenter. It has reached the edge of the hole in perceived 3D binocular visual space. Thus, personal experience shows that harmonic visual-image flow fields in 3D visual space disintegrate in an unpredictable chaotic fashion in a region close to the boundary of the hole about the egocenter; in other words, this is a boundary layer phenomenon.

At this point, it is salient to note that experimentation in human visual science tells us that visual space can have multiple mappings. We argue that the fundamental mapping is that revealed by the surprising fact that there are people who, as the result of brain damage, are visually blind but nevertheless can move about in their 3D environment and perform visually guided movements with precision just as if they could see. This phenomenon has been called blind sight [20]. It demonstrates that the brain has a system able to form a subconscious egocentric 3D visual space that has a one-to-one mapping with the Euclidean outside world. This subconscious visual system has to take into account the warping of this 3D visual space, a warping that occurs because image detection by the retinas occurs on a 2D surface. More precisely, the warping is due to the size of an image on the 2D retinas varying in proportion to the angle subtended at the nodal point of the eye by the corresponding object in the 3D Euclidean outside world [4]. As a result of this warping, changes in the place of the head in the environment and changes in the posture of the body change not only the computed size and position of encoded visual images of objects in the environment and of the body in that environment, but they also change the outline, occlusions, shape, curvatures, relative velocities and relative accelerations of what is seen.

In other words, there exists a conformal mapping between the Euclidean outside world and the warped subconscious 3D visual space. A conformal mapping preserves the infinitesimal shape of objects and angles between coordinates, but sizes of the encoded images are not preserved. Images of objects shrink in size as the relative distance between the egocenter and the object in Euclidean space increases. Nevertheless, despite this

warping, the conformal mapping with the outside world is one-to-one, onto and smooth, and the same warping applies to visual images of the environment and visual images of the body or parts of the body in that environment. Thus, this warped, subconscious, egocentric 3D visual space can be used for the planning and execution of visually guided movements.

That said, there is much evidence that the subconscious and conscious perceptions of the visual world are not the same. The latter can be mediated by top-down cognitive processing of visual images, held to be based on Bayesian statistical hypotheses derived from experience over time. This introduces all sorts of illusions into the consciously perceived visual world. A favorite is the concave face-mask illusion. Because of a lifetime of looking at faces that are convex in shape, when a person looks at the inside concave surface of a face mask, they see it consciously as being convex. Moreover, the illusory convex face seems to turn and follow them as they move about. A consciously perceived visual space containing illusions is not what is needed for the planning and control of precision visually guided movements. The subconscious 3D visual space, derived directly from afferent signals and free of statistical cognitive intervention, is the more appropriate for visually guided movement [1–3].

Experimental support for this description of the human visual system comes from an experiment first carried out by Taylor in 1941 but repeated many times since (see [3]). In the dark, a person holds their hand in front of their face. A brief flash of bright light then burns (or bleaches) an image of the hand onto the retinas. This gives rise to a perceived afterimage of the hand that can persist for several tens of seconds. If, in the dark, the person moves the hand towards the face or moves the face towards the hand, a remarkable thing happens. The afterimage of the hand appears to decrease in size even though the image bleached onto the retinas remains unchanged. This is the opposite of what usually happens because, as the hand is moved closer to the face in light, it appears to increase in size. Our explanation is that the nervous system can transform proprioceptively sensed movements of the body into anticipated changes in its visual image [3]. If the anticipated visual image is subtracted from the actual visual image, the perceived body image does not appear to change its size but only its position in space (so called size-constancy). During the afterimage experiment, the movement of the hand generates the anticipated change in the retinal image of the hand, but the actual retinal image is burnt onto the retinas and so does not change. Subtracting the anticipated size of the image from the actual retinal image generates an illusory increase in the size of the afterimage.

We have constructed a Riemannian geometry description of the calculations required to transform movements of the body sensed proprioceptively into anticipated changes in 3D visual images of the body [1–3]. This is not a simple calculation, because the relationship is a multivariable nonlinear one. It requires 116 variables to specify the configuration of the body moving in a local 3D environment. For each place and posture of the body there exists a 3D warped visual space, and at each point in that visual space there exists an encoded visual image of the environment. These visual images change from one place and posture to another. Geometrically speaking, each 3D visual space corresponds to an egocentric negatively curved Riemannian manifold.

This relationship between movement and vision is complicated by the fact that the visual space is missing the 3D ball of points from about the egocenter [1,2]. This intrinsic hole is what inspired us in the first place to consider how differential geometry and de Rham's theorem in particular might illuminate this hole's influence on the perceived flow of visual images. The chaotic breakdown of visual images near the boundary of the hole was described above.

5.4. Fluid Flow

Turning to Section 3.4, our best interpretation to date of the properties of real-world harmonic flow fields comes from the experimental studies of fluid flow. As set out in that section, Navier–Stokes equations describe fluid flow in 3D Euclidean space where the Laplacian of the vector flow field u is given by $\Delta u = \nabla^2 u$. The harmonic vector

field associated with the presence of holes (e.g., particles, rocks, bubbles, etc.) in the Euclidean space of a fluid is obtained by setting the Laplacian to zero (i.e., $\Delta u = \nabla^2 u = 0$). When this is substituted into the Navier–Stokes equation, extreme turbulence is obtained consistent with experimental observations of extreme turbulence associated with large Reynolds numbers. Inertial forces ρuL, velocities, pressures, temperatures and kinetic energy all change in a chaotic fashion in time and from point to point in space. As the Reynolds number of a fluid is increased experimentally, the energy-dissipating viscous forces within the fluid become negligible and the behavior of the fluid flow approaches that of a theoretical harmonic flow field. Because of the lack of viscous damping within the fluid, a cascade of kinetic energy transfer occurs from large to small eddies and a chaotic flow field results. Rotational vortices and rotational energy become available at the small scale of individual molecules in the fluid. The presence of such small-scale eddies in extreme turbulent fluid flow created by the presence of holes in the space of the fluid suggests a role in protein folding for extreme turbulence in the intracellular fluid of biological cells due to holes in the fluid created by the densely packed organelles within the fluid.

Inertial forces changing from point to point in the fluid give rise to a symmetrical mass-inertia matrix corresponding to a kinetic energy Riemannian metric tensor [21]. This provides an opportunity to improve on the Navier–Stokes equations. A Riemannian kinetic energy metric allows the fluid flow to be described by vector fields on a 3D curved Riemannian manifold rather than in the flat 3D Euclidean space of the Navier–Stokes equations. As shown in Section 3.4, the harmonic 1-form flow field λ on a Riemannian manifold is obtained by setting the Laplacian of the 1-form λ on the Riemannian manifold to zero (i.e., $\Delta\lambda = (-\sum_{i=1}^{n} \nabla^2_{\xi_i})\lambda - \mathrm{Ric}\,(\lambda) = 0$). The Ricci curvature at each point on the Riemannian manifold takes into account nonlinear dynamics attributable to inertial forces changing from point to point in the fluid. These are the nonlinear forces within the fluid that are predominantly responsible for turbulent flow. Thus, describing fluid flow on a Riemannian manifold endowed with an appropriate Riemannian metric will give a more detailed description of fluid flow than given by the Navier–Stokes equations. However, because of the chaotic nature of turbulent flow, stochastic methods are required to obtain estimates of an appropriate kinetic energy Riemannian metric as a function of position in the manifold. A stochastic time series model of the changing Riemannian metric appears necessary.

Deep learning neural networks have been used to model turbulent fluid flow [16]. Unlike a "black box" solution, however, differential geometry offers a more general mathematical account of turbulent flow that can provide an understanding of the underlying physical phenomena.

5.5. Protein Folding

There exists ample experimental evidence showing that intracellular aqueous solution plays an important role in protein folding. However, the availability of data that can be used to verify or refute working hypotheses about the precise nature of water–protein interactions is limited by the experimental difficulties involved. That said, available results provide no support for the popular hydrophobic theory of protein folding. Instead, the data support water forming hydrogen bonds between points in the peptide chain and then pulling those points towards each other. Work to measure the role of water in protein folding by measuring neutron diffraction patterns is promising, but the technique is not an easy one, requiring stochastic analysis of measured data and use of computer simulations to find the model that best fits the data. Meanwhile, the general conclusion that movement of water does play a role in protein folding seems beyond doubt [19,22,23].

Our proposal regarding the Brownian motion of bonding angles within peptide chains associated with turbulent rotation of water molecules flowing around organelles within the cell is a novel hypothesis. It is based solely on the mathematics of fluid flow (i.e., the Navier–Stokes equations) in conjunction with de Rham's theorem about the influence of holes on flow fields. It awaits experimental testing that is beyond the resources of the

present authors. Meanwhile we are convinced of the following. The logical formalism and certainty within the mathematical formulation set out here make it highly likely that the experimentally observed Brownian motion in the intracellular aqueous solution during protein folding is contributing fundamentally to the folding process. We hope that this can provide a guide to future research.

6. Conclusions

An aim of this paper has been to demonstrate the worth of applying differential geometry to gain understanding of any physical system where holes can occur in a surrounding flow field. According to de Rham's theory, holes alter the character of the surrounding fields, transforming them into harmonic fields that are closed but not exact. As a result, the consequent flow of a fluid becomes chaotic. We have discussed this phenomenon in relation to 3D binocular vision, which breaks down due to a hole in the visual flow field about the egocenter. We now point out that the presence of electromagnetic screening about a source of electromagnetic radiation, such as a television transmitter, creates a hole in the surrounding 3D space of electromagnetic fields. We suggest that it was the creation of chaotic flow fields that underlaid the television studio issue [5,6] that in part led us to examine this theory in the first place. No cause of the cancer cluster came to light in the investigations carried out, but the studio was ultimately abandoned [24].

For mechanical and electrodynamical systems where flow data are apparently lacking, we hope that the ideas presented here will be of interest to those working directly in those fields. In particular, we are optimistic that our proposal about protein folding might stimulate a desire to know more about differential geometry and de Rham using the tutorial information written here. In summary, we trust that our demonstration of the reach of de Rham's cohomology theorem across diverse real-world systems can help to focus future attention in general on this extraordinary piece of mathematics. After all, the notion that the same mathematics can be applied across a range of systems is what underlies a wealth of modeling and simulation, be it with mechanical, electrical, chemical, economic, physiological, sociological or cosmological variables, etc. Differential geometry has found a key place in physics for those on a quest to understand nature. For those elsewhere who have mathematical inclinations along with a foothold in a relevant field, the work involved to understand and apply differential geometry promises ultimately to be very rewarding.

Author Contributions: Conceptualization, P.D.N.; methodology, P.D.N.; writing—original draft preparation, P.D.N. and M.D.N.; writing—review and editing, M.D.N. and P.D.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

As indicated in the Introduction, a substantial knowledge of geometrical theory is required in order to appreciate exactly how de Rham's theorem links holes in the topology of smooth manifolds with the chaotic behavior of integral flow fields on a manifold. A full understanding of the link depends on familiarity with (i) simplicial complexes, (ii) singular homology groups, (iii) singular chain complexes, (iv) chain maps, (v) computing singular homology groups, (vi) the Mayer–Vietoris theorem, (vii) homotopy invariance, (viii) singular cohomology and the coboundary operator, (ix) skew-symmetric tensors, (x) Grassmann algebra and (xi) de Rham cohomology groups and the Hodge star oper-

ator, as well as knowledge of the Laplacian operator and harmonic flow fields. Some of these have been touched on in the text. In this Appendix A, we provide a more detailed overview of the differential geometry involved. Proofs are available in a variety of texts [7,11,12,25–30].

Appendix A.1. Simplicial Complexes

In 3D Euclidean spaces, (P_0), (P_0, P_1), (P_0, P_1, P_2), (P_0, P_1, P_2, P_3), where the P_i are points in $\mathbb{R}^3$, are the simplicial building blocks that can be glued together along their boundary faces to form 0-, 1-, 2- and 3-dimensional oriented connected compact subspaces in $\mathbb{R}^3$. These glued structures are called simplicial complexes, on which 0, 1, 2 and 3rd degree differential forms, respectively, can be integrated. A standard 3-simplex $\Delta_3 = (e_0, e_1, e_2, e_3)$ in $\mathbb{R}^3$ is defined such that $e_0 = (0,\ 0,\ 0)$, $e_1 = (1,\ 0,\ 0)$, $e_2 = (0,\ 1,\ 0)$ and $e_3 = (0,\ 0,\ 1)$. This standard 3-simplex Δ_3 defines a Cartesian coordinate system on $\mathbb{R}^3$ where e_0 defines the origin and e_1, e_2, e_3 define unit-length vectors along mutually orthogonal Cartesian coordinates x^1, x^2, x^3, respectively. A 2-simplex (P_0, P_1, P_2) in $\mathbb{R}^3$ defines a triangle in 3D space that includes all the points along its oriented boundary faces (P_1, P_2), (P_0, P_2), (P_0, P_3) as well as all the internal points. Given vectors (v_0, v_1, v_2) from the origin e_0 to each of the points P_0, P_1 and P_2, respectively, if the vectors $v_1 - v_0$ and $v_2 - v_0$ are linearly independent, we say that the points P_0, P_1 and P_2 are affinely independent (unlike a vector space, an affine space does not span the origin). As the vector v_0 is moved about in $\mathbb{R}^3$, the affinely independent 2-simplex (P_0, P_1, P_2) can be moved to any place in $\mathbb{R}^3$. By gluing together a collection of 2-simplices along common boundary faces, a 2-dimensional compact oriented surface known as a simplicial complex can be constructed in $\mathbb{R}^3$. Such a surface is referred to as a 2-chain, denoted by c_2. Many different 2-chains can be constructed in $\mathbb{R}^3$. Many different 0-chains (discrete points), 1-chains (wire-frame models), 2-chains (2-dimensional surfaces) and 3-chains (3-dimensional subspaces) can be constructed in $\mathbb{R}^3$. In general, simplicial complexes are closed and bounded polyhedrons.

Given a 3D smooth manifold M, a 2-simplex (P_0, P_1, P_2) in $\mathbb{R}^3$ can be mapped by a smooth map σ_2 into M. Such a smooth map is a nonlinear map giving rise to a singular 2-simplex in M. Because of the nonlinear mapping, such a singular 2-simplex may no longer look like a triangle. It can lose rank and may even appear as a single point. For this reason, it is referred to as a *singular* simplex. The same applies for one, two and three singular simplices in 3D manifolds.

Of all the possible 2-chains c_2 that can be constructed in $\mathbb{R}^3$, there are some 2-chains that close on themselves (like 2-spheres) and have empty boundaries (i.e., $\partial c_2 = 0$). Such 2-chains with empty boundaries are called 2-cycles, represented by the notation z_2. A similar property applies to 0-, 1-, 2- and 3-chains in $\mathbb{R}^3$, giving 0-, 1-, 2- and 3-cycles with zero boundaries.

Appendix A.2. Singular Homology Groups

Singular 2-cycles in the underlying topology X of M can be used to detect the existence of 2-dimensional holes in M. The singular 2-homology group on the underlying topology of M is defined to be all of the 2-cycles (represented by the free Abelian group $Z_2(M)$) on the underlying topology of M that are not boundaries of 3-chains c_3. Such 2-cycles enclose 2-dimensional holes in the underlying topology. Define $B_2(M)$ to be all of the 2-cycles in $Z_2(M)$ that are boundaries of 3-chains c_3. The singular 2-homology group is then defined by the quotient $H_2(M) = {}^{Z_2(M)}/_{B_2(M)}$ (i.e., the group of all 2-cycles that are not boundaries of 3-chains). The dimension of the free Abelian group $H_2(M)$ equals the number of 2-dimensional holes in the underlying topology X of M. This can be generalized to r-th degree homology groups $H_r(M) = {}^{Z_r(M)}/_{B_r(M)}$ with $r = 0,\ 1,\ \cdots, n$ in the underlying topology X of an n-dimensional smooth manifold M. This intuitive description of singular homology groups can be formalized by defining a sequence of singular homology groups attached to the underlying topology of smooth manifolds.

Appendix A.3. Singular Chain Complexes

A long sequence of Abelian groups connected by homomorphisms

$$\cdots \to G_{r+1} \overset{\alpha_{r+1}}{\to} G_r \overset{\alpha_r}{\to} G_{r-1} \to \cdots$$

is said to be exact if im $\alpha_{r+1} = \ker \alpha_r$ for all r.

The first isomorphism theorem for groups implies that a short 5-term sequence

$$0 \to A \overset{\alpha}{\to} B \overset{\beta}{\to} C \to 0$$

is exact if and only if the homomorphism $\alpha : A \to B$ is injective and the homomorphism $\beta : B \to C$ is surjective. This implies that for a short 4-term sequence

$$0 \to A \overset{\alpha}{\to} B \to 0$$

to be exact, the homomorphism $\alpha : A \to B$ has to be both injective and surjective. This fact is useful when computing singular homology groups of higher-dimensional topological spaces (e.g., high-dimensional spheres described below).

A long sequence of free Abelian groups (chains) together with homomorphisms

$$\cdots \to C_{r+1} \overset{\partial_{r+1}}{\to} C_r \overset{\partial_r}{\to} C_{r-1} \to \cdots$$

is called a singular chain complex, denoted by C_*, if the composition of any two consecutive homomorphisms $\partial_r \circ \partial_{r+1}$ is zero. This is the same as requiring im $\partial_{r+1} = \ker \partial_r$.

The r-th singular homology group of the singular chain complex C_* is

$$H_r(C_*) = {}^{\ker \partial_r}/_{\text{im } \partial_{r+1}} = {}^{Z_r(X)}/_{B_r(X)}$$

where $Z_r(X)$ is the space of all r-cycles in X, and $B_r(X)$ is the space of all r-cycles in X that are boundaries of $(r+1)$-chains C_{r+1} in X. It follows that the singular chain complex C_* is exact if and only if $H_r(C_*)$ is zero for all values of r. In other words, the singular homology groups $H_r(C_*)$ provide a precise measure of the failure of the chain complex C_* to be exact. This occurs when there is an r-dimensional hole(s) in any of the r-chains.

Appendix A.4. Chain Maps $F_\#$

A chain map $F_\# : C_*(X) \to D_*(Y)$ between singular chain complexes associated with topological spaces X and Y, respectively, is a collection of homomorphisms $F_\# : C_r \to D_r$ for all r such that $F_\# \circ \partial_r = \partial_r \circ F_\#$, as illustrated below:

$$\begin{array}{ccccccccc} X \to & C_*(X) & \cdots & \overset{\partial_{r+1}}{\to} & C_r(X) & \overset{\partial_r}{\to} & C_{r-1}(X) & \to \cdots \\ & \downarrow F & \downarrow F_\# & & \downarrow F_\# & & \downarrow F_\# & \\ Y \to & D_*(Y) & \cdots & \overset{\partial_{r+1}}{\to} & D_r(Y) & \overset{\partial_r}{\to} & D_{r-1}(Y) & \to \cdots . \end{array}$$

A continuous map $F : X \to Y$ between topological spaces X and Y induces a singular chain map $F_\# : C_*(X) \to D_*(Y)$ between the singular chain complex $C_*(X)$ and the singular chain complex $D_*(Y)$. This map is defined by the homomorphisms $F_\# : C_r(X) \to D_r(Y)$ for all values of r, as illustrated. Since $F_\#$ maps $\ker \partial_r$ to $\ker \partial_r$ and $\text{im } \partial_{r+1}$ to $\text{im } \partial_{r+1}$, it follows that the continuous map $F : X \to Y$ between topological spaces X and Y induces a singular homology group homomorphism $F_* : H_r(C_*(X)) \to H_r(D_*(Y))$ for all values of r.

Because the image of any one singular simplex is connected, it has to lie entirely in one path-connected component of the topological space X. So it follows that the r-th singular homology group $H_r(X)$ for X equals the direct sum of the r-th singular homology groups $H_r(X_\alpha)$ for each of the path-connected components X_α (i.e., $\bigoplus_\alpha : H_r(X_\alpha) \to H_r(X)$).

Appendix A.5. Computing Singular Homology Groups

The definition given above of singular homology groups involves the notion of r-chain complexes consisting of every possible r-chain that can be constructed in a topological space. This is a very large number of r-chains, so it does not help much in telling us how we can actually compute singular homology groups. However, powerful tools based on the so-called Mayer–Vietoris exact sequence and the zigzag lemma have been developed for computing the homology groups for most if not all topological spaces. Some singular homology groups can be deduced directly from the definition, in particular (i) degree zero singular homology groups $H_0(X)$ for any topological space X and (ii) all singular homology groups $H_r(X)$ for all values of r on a discrete topological space X.

Appendix A.5.1. Degree Zero Singular Homology Groups

For every topological space X, the boundary ∂C_0 of the zero-dimensional chain complex C_0 is zero because there is no C_{-1} chain. It follows that every singular 0-chain is a 0-cycle (i.e., $C_0(X) = Z_0(X)$). On the other hand, a singular 0-chain is a formal linear combination of points x_i in X with integer coefficients (i.e., $c_0 = \sum_i a_i x_i$). Assuming that X is path-connected, we can define a surjective map $\epsilon : C_0(X) \to \mathbb{Z}$ by $\epsilon : \sum_i a_i x_i \to \sum_i a_i \in \mathbb{Z}$, where $\mathbb{Z}$ is the group of all integers. The kernel of this map corresponds to all points $B_0(X)$ in X that are boundary points of a 1-chain, and consequently, from the first isomorphism theorem for groups,

$$Z_0(X)/\ker \epsilon = Z_0(X)/B_0(X) = H_0(X).$$

Thus, for any topological space X, the zero singular homology space $H_0(X)$ is isomorphic to the group of integers $\mathbb{Z}$, and the dimension of $H_0(X)$ is equal to the number of path-connected components in X. This dimension is the b_0 Betti number.

Appendix A.5.2. Homology Groups for Discrete Topological Spaces

For a discrete space X, each isolated point corresponds to a connected component of X, so $H_0(X)$ is a free Abelian group isomorphic to the group of integers $\mathbb{Z}$ with dimension b_0 equal to the number of discrete points in X. Although each connected component of X is a single isolated point, the boundary $\partial\sigma_r$ for a singular r-simplex σ_r is given by the alternating sum $\partial\sigma_r = \sum_{i=0}^{r}(-1)^r \sigma_r \circ F_{i,r} = \sum_{i=0}^{r}(-1)^r \sigma_{r-1}$, where $\sigma_r \circ F_{i,r}$ represents the singular i-th boundary face. For single point spaces, the singular faces $\sigma_r \circ F_{i,r}$ are all equal to σ_{r-1}. Thus, $\partial : C_r \to C_{r-1}$ is an isomorphism when r is even and zero when r is odd. Consequently, we obtain a singular chain complex

$$\cdots \to C_3 \to C_2 \to C_1 \to C_0 \to 0$$

where the boundary maps $\partial : C_r \to C_{r-1}$ are 0-maps for odd values of r and are isomorphisms for even values of r. It follows that the sequence is exact at each Abelian group C_r except C_0. But we have already dealt with $H_0(X)$ for C_0. Since the singular homology groups $H_r(X)$ provide a measure of the deviation from exactness, it follows that the singular homology group $H_r(X)$ for each isolated point of a discrete topological space X is equal to zero for all $r > 0$.

Appendix A.6. Mayer–Vietoris Theorem

The Mayer–Vietoris theorem is an important tool for computing singular homology groups for topological spaces. If X is a topological space and U, V are open subsets of X whose union is X, then for each r there is a homomorphism called the *connection homomorphism* $\partial_* : H_r(X) \to H_{r-1}(U \cap V)$, such that the following long sequence of homology groups is exact:

$$\cdots \to H_r(U \cap V) \overset{i_* \oplus j_*}{\to} H_r(U) \oplus H_r(V) \overset{k_* - l_*}{\to} H_r(X) \overset{\partial_*}{\to} H_{r-1}(U \cap V) \to \cdots$$

where $i_* \bigoplus j_* : H_r(U \cap V) \to H_r(U) \oplus H_r(V)$ with $(i_* \bigoplus j_*)[c] = (i_*[c], j_*[c])$ and $k_* - l_* : H_r(U) \oplus H_r(V) \to H_r(X)$ with $(k_* - l_*)([c], [c']) = k_*[c] - l_*[c']$ where $[c]$ and $[c']$ represent singular homology classes [8].

To illustrate how the Mayer–Vietoris theorem can be combined with other properties of topological spaces to compute singular homology groups, consider computing the singular homology groups for all values of r and for all dimensions n of a unit sphere. Consider a unit n-sphere $\mathbb{S}^n$ with open subset U equal to $\mathbb{S}^n$ with the south pole removed and open subset V equal to $\mathbb{S}^n$ with the north pole removed. A part of the Mayer–Vietoris exact sequence is

$$H_r(U) \bigoplus H_r(V) \to H_r(\mathbb{S}^n) \xrightarrow{\partial_*} H_{r-1}(U \cap V) \to H_{r-1}(U) \oplus H_{r-1}(V)$$

where the connecting homomorphism is $\partial_* : H_r(\mathbb{S}^n) \to H_{r-1}(U \cap V)$. All points in U can be contracted to the north pole and all points in V can be contracted to the south pole. This means that both U and V are contractible and therefore simply connected (i.e., contain no holes). It follows that

$$H_r(U) \bigoplus H_r(V) = 0 \text{ and } H_{r-1}(U) \oplus H_{r-1}(V) = 0$$

Consequently, when $r > 1$ and $n \geq 1$, the exact sequence can be rewritten as

$$0 \to H_r(\mathbb{S}^n) \xrightarrow{\partial_*} H_{r-1}(U \cap V) \to 0.$$

This is an exact 4-term sequence as described above and, consequently, the connecting homomorphism $\partial_* : H_r(\mathbb{S}^n) \to H_{r-1}(U \cap V)$ is a bijective map (i.e., an isomorphism). Just as the equator of a unit 2-sphere $\mathbb{S}^2$ is topologically equivalent to a unit circle $\mathbb{S}^1$, the equatorial sphere $(U \cap V)$ of a unit n-sphere $\mathbb{S}^n$ is topologically equivalent to a unit $(n-1)$-sphere $\mathbb{S}^{n-1}$. This means $H_r(\mathbb{S}^n) \cong H_{r-1}(U \cap V) \cong H_{r-1}(\mathbb{S}^{n-1})$ for $r > 1$ and $n \geq 1$ and $H_r(\mathbb{S}^n) \cong 0$ if $0 < r < n$. We already know that $H_0(\mathbb{S}^n) \cong \mathbb{Z}$ and $H_r(\mathbb{S}^n) \cong 0$ if $r > n$, so we have all the singular homology groups for higher-dimensional spheres.

Appendix A.7. Homotopy Invariance

A 2-torus $\mathbb{T}^2$ has two one-dimensional holes and one two-dimensional hole. In other words, it has a two-dimensional degree-one singular homology group $H_1(\mathbb{T}^2)$ and a one-dimensional degree-two singular homology group $H_2(\mathbb{T}^2)$. The rubber sheet geometry of the 2-torus allows it to be deformed into many different sizes and shapes. As long as it is not cut, torn, punctured or glued, all the different sizes and shapes of $\mathbb{T}^2$ can be mapped homeomorphically onto each other and the topological structure remains unchanged (i.e., the number of holes and hence the singular homology groups $H_1(\mathbb{T}^2)$ and $H_2(\mathbb{T}^2)$ remain unchanged). A similar property holds for all topological spaces. The number of holes of various dimensions, and hence their singular homology groups, are homotopically and topologically invariant. Let $f_0, f_1 : X \to Y$ be continuous homotopic maps (i.e., $f_0(X)$ in Y and $f_1(X)$ in Y can be morphed into each other without changing their topological structure), then $f_0(X)$ and $f_1(X)$ (together with all the intermediate shapes along the homotopy transformation pathway between $f_0(X)$ and $f_1(X)$) have the same singular homology groups. In other words, continuous homotopy maps $f_0, f_1 : X \to Y$ induce singular homology group homomorphisms $(f_0)_*, (f_1)_* : H_r(X) \to H_r(Y)$ for all values of r. A continuous map $F\colon X \to Y$ between topological spaces X and Y is said to be a homotopy equivalence if there is a continuous pull-back map $G\colon Y \to X$ such that the composition $F \circ G$ is homotopic to the identity map I_Y on Y and the composition $G \circ F$ is homotopic to the identity map I_X on X. If $F\colon X \to Y$ is a homotopy equivalence between topological spaces X and Y, then for each $r \geq 0$, the map $F_* : H_r(X) \to H_r(Y)$ is an isomorphism.

Appendix A.8. Singular Cohomology

Singular homology groups are covariant functors from the category of topological spaces to the category of free Abelian groups. For many applications, including de Rham's theory of differential forms, it is more useful to have contravariant functors. Singular cohomology is a pull-back version of singular homology. It does not give any new information but simply presents the same information in a different way.

The transformation from covariant functor to contravariant functor is based on the following fact. If A and B are Abelian groups under pointwise addition, then the set $\mathrm{Hom}(A, B)$ of group homomorphisms from A to B is itself an Abelian group under pointwise addition. For any fixed Abelian group G, each homomorphism $f_{\#}: A \to B$ gives an induced pull-back homomorphism $f^{\#}: \mathrm{Hom}(B, G) \to \mathrm{Hom}(A, G)$. If $\varphi^B \in \mathrm{Hom}(B, G)$ and $\varphi^A \in \mathrm{Hom}(A, G)$, then $f^{\#}(\varphi^B) = \varphi^A$. This pull-back map can be used to relate the free Abelian group of all singular r-chains $C_r(X)$ on a topological space X to the free Abelian group of all singular r-cochains $C^r(X; G) = \mathrm{Hom}(C_r(X), G)$ on the same topological space X. For free Abelian groups, the group G equals the group of integers $\mathbb{Z}$. A continuous map f: $X \to Y$ between topological spaces X and Y induces a homomorphism $f_{\#}: C_r(X) \to C_r(Y)$ between the free Abelian groups of all possible r-chains on X and Y (i.e., $c_r^Y = f_{\#} c_r^X$ where we use the notation c_r^Y to represent a single r-chain in Y). It also induces a pull-back homomorphism $f^{\#}: C^r(Y; G) \to C^r(X; G)$, or equivalently, $f^{\#}: \mathrm{Hom}(C_r(Y), G) \to \mathrm{Hom}(C_r(X), G)$ (i.e., $f^{\#}: \varphi^Y(f_{\#} c_r^X) \to (f^{\#}\varphi^Y)(c_r^X)$ where $\varphi^Y \in \mathrm{Hom}(C_r(Y), G)$).

Coboundary Operator

The boundary operator $\partial: C_{r+1}(X) \to C_r(X)$ between free Abelian groups of all possible $(r+1)$-chains and all possible r-chains on X induces a pull-back coboundary operator $\delta: C^r(X; G) \to C^{r+1}(X; G)$ between all possible r-cochains and all possible $(r+1)$-cochains on X (notice increasing values of r). For the coboundary of a single r-chain c_r^X on X, we obtain $f^{\#}: (\delta\varphi^Y)(f_{\#} c_r^X) \to (f^{\#}\delta\varphi^Y)(c_r^X)$, where $(\delta\varphi^X)(c_r^X) = (\varphi^X)(\partial c_r^X)$. More simply, this shows that $f^{\#} \circ \delta = \delta \circ f^{\#}$. It also follows that $\delta \circ \delta = \delta\delta = 0$, because $(\delta\delta\varphi^X)(c_r^X) = (\varphi^X)(\partial\partial c_r^X)$ and $\partial\partial = 0$. Because $\delta\delta = 0$, there exists an exact long cochain complex and cochain map $f^{\#}$ between topological spaces X and Y (notice decreasing r).

$$\begin{array}{ccccccc} \cdots \to & C^{r+1}(X; G) & \xrightarrow{\delta} & C^r(X; G) & \xrightarrow{\delta} & C^{r-1}(X; G) & \to \cdots \\ & \downarrow f^{\#} & & \downarrow f^{\#} & & \downarrow f^{\#} & \\ \cdots \to & D^{r+1}(Y; G) & \xrightarrow{\delta} & D^r(Y; G) & \xrightarrow{\delta} & C^{r-1}(Y; G) & \to \cdots \end{array}$$

An r-cochain $\varphi^X \in C^r(X; G)$ is an r-cocycle $\varphi^X \in Z^r(X; G)$ if $\delta\varphi^X = 0$ (i.e., it has an empty coboundary), and it is a bounding cocycle $\varphi^X \in B^r(X; G)$ if there exists a cocycle $\psi^X \in C^{r-1}(X; G)$ such that $\delta\psi^X = 0$ (i.e., it is a coboundary of an $(r-1)$-cochain). Both the r-cocycles $Z^r(X; G)$ and the r-bounding cocycles $B^r(X; G)$ are preserved by cochain maps $f^{\#}$, so we can define the r-th singular cohomology group of X with coefficients in G to be the quotient $H^r(X; G) = {}^{Z^r(X;G)}/_{B^r(X;G)}$.

It can be concluded that a continuous map $f: X \to Y$ between topological spaces X and Y induces a pull-back homomorphism $f^{\#}: C^r(Y; G) \to C^r(X; G)$ between r-th singular cochains on Y and r-th singular cochains on X. This in turn induces a pull-back homomorphism $f^*: H^r(Y; G) \to H^r(X; G)$ between the r-th singular cohomology group $H^r(Y; G)$ on Y and the r-th singular cohomology group $H^r(X; G)$ on X. Just as singular homology groups are homotopically and topologically invariant, so are singular cohomology groups. In other words, singular cohomology groups are isomorphic to their corresponding singular homology groups.

Appendix A.9. Skew-Symmetric Tensors on a Smooth n-Manifold M

At every point p on a smooth n-manifold M, there exists an n-dimensional tangent space T_pM and an n-dimensional cotangent space T_p^*M. A covector or 1-form λ in T_p^*M

acts on a vector v in T_pM and transforms it into a real number (i.e., $\lambda(v) = \lambda, v \to \mathbb{R}$). A type $(0,\ 2)$ skew-symmetric tensor $T(X,\ Y)$ operating on two vectors X and Y in T_pM to give a real number changes sign if the order of the two vectors is reversed (i.e., $T(X,\ Y) = -T(Y,\ X)$). If $(e_1, \cdots, e_n)$ are unit basis vectors spanning the tangent vector space T_pM at $p \in M$ and $(\varepsilon^1, \cdots, \varepsilon^n)$ are unit basis covectors dual to $(e_1, \cdots, e_n)$ spanning the cotangent vector space T_p^*M, then we can write the type $(0,\ 2)$ skew-symmetric tensor $T(X,\ Y)$ in the form $T = T_{j_1 j_2}\varepsilon^{j_1 j_2}$, where

$$\varepsilon^{j_1 j_2} = \ (\varepsilon^{j_1} \bigotimes \varepsilon^{j_2} - \varepsilon^{j_2} \bigotimes \varepsilon^{j_1}) = \varepsilon^{j_1} \wedge\ \varepsilon^{j_2} \text{ with } j_1 < j_2$$

This defines the wedge product $\varepsilon^{j_1} \wedge \varepsilon^{j_2}$ to be a basis for the type $(0,\ 2)$ skew-symmetric tensor space Ω^2 at $p \in M$, and $T_{j_1 j_2}$ is the component of the type $(0,\ 2)$ skew-symmetric tensor (vector) projected onto the basis vector $\varepsilon^{j_1} \wedge\ \varepsilon^{j_2}$. This can be generalized to a skew-symmetrization (or alternating) operator $\mathbf{A}$ for a type $(0,\ s)$, $s = 0, 1, \cdots, n$ skew-symmetric tensor T by the formula

$$\mathbf{A}T(X_1, \cdots, X_s) = \frac{1}{s!} \sum_{(a_1 \cdots a_s)} (-1)^{\kappa} T(X_{a_1}, \cdots, X_{a_s})$$

where $s!$ is the number of permutations $P(n,r) = \frac{n!}{(n-r)!}$ of n objects taken r at a time when order matters and with $n = s$ and $r = s$ in this case, $\sum_{(a_1 \cdots a_s)}$ is the sum over all $s!$ permutations of the s vectors $(X_1, \cdots, X_s)$, and κ is the parity of each permutation (i.e., $\kappa = 0$ if the number of interchanges of two vectors is even and $\kappa = 1$ if the number of interchanges is odd).

If Φ is an r_1-form and Π is an r_2-form, then $\Phi \wedge \Pi = \mathbf{A}(\Phi \otimes \Pi)$ is an $(r_1 + r_2)$-form. The wedge product $\wedge$ is

(i) associative $\Phi \wedge (\Pi \wedge \Psi) = (\Phi \wedge \Pi) \wedge \Psi$,
(ii) (anti)-commutative $(\Phi \wedge \Pi = (-1)^{r_1 r_2} \Pi \wedge \Phi)$, and
(iii) distributive $(\Phi + \Psi) \wedge \Pi = \Phi \wedge \Pi + \Psi \wedge \Pi$ where Φ and Ψ are both r_1-forms here.

Basis vectors (skew-symmetric tensors) spanning the r-th skew-symmetric tensor space $\Omega^r(M)$ are $\varepsilon^{j_1 \cdots j_r} = \left\{\varepsilon^{j_1} \wedge \cdots \wedge \varepsilon^{j_r} : j_1 < \cdots < j_r\right\}$.

Appendix A.10. Grassmann Algebra

At each $p \in M$ on a smooth n-manifold M, there exists a family of skew-symmetric tensor spaces $\Omega^0, \Omega^1, \Omega^2, \cdots, \Omega^n$. The skew-symmetric tensors in these skew-symmetric tensor spaces are called differential forms. An r-form in Ω^r acts on r vectors $v_1, \cdots, v_r$ in T_pM and transforms them into a real number for all values of $r = 0, \cdots, n$. The dimension of each skew-symmetric tensor space Ω^r is given by $\frac{n!}{r!(n-r)!}$. From this formula, the skew-symmetric tensor spaces Ω^0 and Ω^n are both one-dimensional. The dimension increases linearly with r from Ω^0 to $\Omega^{\frac{n}{2}}$ and then decreases linearly to Ω^n. This implies that the dimension of Ω^r is equal to the dimension of Ω^{n-r}. Indeed, the skew-symmetric tensor space Ω^r is isomorphic to the skew-symmetric tensor space Ω^{n-r}. An isomorphic operator $*$ known as the Hodge star operator transforms r-forms in Ω^r into $(n-r)$-forms in Ω^{n-r} for all values of r. For $r < 0$ and $r > n$, the skew-symmetric tensor spaces Ω^r are empty spaces and the differential forms are set to zero. An operator d, known as the exterior derivative, transforms an r-form into an $(r+1)$-form. Importantly, the composition $d \circ d = dd$ is always zero. It follows that there exists a sequence of skew-symmetric tensor spaces connected by the exterior differential operator

$$0 \ \xrightarrow{d} \Omega^0 \xrightarrow{d} \cdots \xrightarrow{d} \Omega^n \xrightarrow{d} 0$$

where each arrow corresponds to an exterior differential d. An operator d^* transforms an r-form into an $(r-1)$-form. More generally, $d^* = (-1)^{nr+n+1} * d *$, where $*$ is the Hodge

star operator, that is, an isomorphism between Ω^r and Ω^{n-r}. The operators d and d^* are not inverses of each other, they are adjoints. The direct sum of skew-symmetric tensor spaces forms a graded algebra $\Omega = \Omega^0 \bigoplus \cdots \bigoplus \Omega^n$, known as an exterior algebra or Grassmann algebra. Any r-form ω for which $d\omega = 0$ is said to be a closed r-form, while any r-form ω equal to $d\lambda$ where λ is an $(r-1)$-form is said to be an exact r-form.

Appendix A.11. De Rham Cohomology Groups

From Grassmann algebra and the fact that $d \circ d = dd = 0$, it follows that every exact differential form is closed. It does not follow, however, that every closed differential form is exact. De Rham's theory of differential forms makes it clear that whether or not a closed differential form is exact depends uniquely on the underlying topological structure of the manifold. Specifically, it depends on the presence or absence of holes of various dimensions in the underlying topology. As illustrated by the amount of theory summarized in this appendix, a significant amount of differential geometry is needed to appreciate this remarkable fact. Because the exterior differential $d: \Omega^r \to \Omega^{r+1}$ is linear, its kernel $\ker(d_r) = Z^r(M) =$ closed r-forms and its image $\mathrm{im}(d_{r-1}) = B^r(M) =$ exact r-forms are linear vector spaces. Since every exact form is closed, it follows that the exact r-forms are a subspace of the closed r-forms ($B^r(M) \subseteq Z^r(M)$). Thus, the r-th de Rham cohomology group can be defined to be the quotient vector space

$$H^r_{dR}(M) = {}^{Z^r(M)}/_{B^r(M)}$$

(i.e., the space of all closed forms that are not exact).

While $H^r_{dR}(M)$ is a real vector space, it is also a Lie group under addition. Since $\Omega^r(M) = 0$ for $r < 0$ and $r > n$, it follows that $H^r_{dR}(M)$ equals zero for these values of r. It also follows that $H^r_{dR}(M) = 0$ for $0 \leq r \leq n$ if and only if every globally closed r-form on M is globally exact.

The vector quotient space ${}^{Z^r(M)}/_{B^r(M)}$ partitions the space of closed r-forms into equivalence classes. For any closed r-form ω on M that is not exact, we let $[\omega]$ denote the equivalence class of ω. If ω and ω' belong to the same equivalence class, $[\omega]$ then they differ by an exact r-form (i.e., $\omega - \omega' = d\eta$), and we say that ω and ω' are cohomologous. This is easy to visualize in terms of homology groups, where it simply means that there can exist multiple r-cycles that contain the same r-dimensional hole. Just as a continuous map $f\colon X \to Y$ between topological spaces X and Y induces a pull-back homomorphism $f^*: H^r(Y;G) \to H^r(X;G)$ between the r-th singular cohomology group $H^r(Y;G)$ on Y and the r-th singular cohomology group $H^r(X;G)$ on X, a smooth map $F: M \to N$ between smooth manifolds M and N induces a pull-back $F^*: H^r_{dR}(N) \to H^r_{dR}(M)$ between the r-th de Rham cohomology groups of N and M, respectively. This corresponds to a contravariant functor between the category of smooth maps between smooth manifolds and the category of de Rham cohomology groups.

We can form an interesting picture of a relation between degree-one de Rham cohomology groups $H^1_{dR}(M)$ on a connected smooth manifold and the underlying topology X of the manifold as measured by the fundamental group $\Pi_1(X,q)$. All the 1-cycles in all the possible 2-chains in the underlying topology of the smooth manifold M that contain a 1-dimensional hole correspond not only to $H^1_{dR}(M)$ but also to the fundamental group $\Pi_1(X,q)$. The fundamental group consists of all the loops associated with the point q in the underlying topology X of M that cannot be contracted to a point. The set of group homomorphisms $\mathrm{Hom}(\Pi_1(X,q),\mathbb{R})$ from the fundamental group $\Pi_1(X,q)$ to the additive group of real numbers $\mathbb{R}$ is a vector space under addition and multiplication by constants. We can define a linear map $\Phi: H^1_{dR}(M) \to \mathrm{Hom}(\Pi_1(X,q),\mathbb{R})$ as follows: Given a cohomology class $[\omega] \in H^1_{dR}(M)$, define $\Phi[\omega]: \Pi_1(X,q) \to \mathbb{R}$ by $\Phi[\omega] = \int_{\widetilde{\gamma}} \omega =$ a real number, where $[\gamma]$ is any path homotopy class in $\Pi_1(X,q)$ and $\widetilde{\gamma}$ is any piecewise smooth curve representing the same loop path class $[\gamma]$. The map $\Phi: H^1_{dR}(M) \to \mathrm{Hom}(\Pi_1(X,q),\mathbb{R})$ is a well-defined homomorphism between groups. It follows that $H^1_{dR}(M) = 0$ when the topol-

ogy X underlying the smooth manifold M is simply connected (i.e., has no 1-dimensional holes and no loops that cannot be contracted to a point).

Appendix A.12. De Rham's Theorem

De Rham's theorem hinges on the fact that, just like singular homology and singular cohomology groups, de Rham's cohomology groups $H^r_{dR}(M)$ for all values of r are homotopically and topologically invariant. Assume that smooth maps $F : M \to N$ and $G : M \to N$ between smooth manifolds M and N are homotopic (i.e., $F(M)$ in N can be morphed into $G(M)$ in N and vice versa). Define a homotopy operator $h : \Omega^r(N) \to \Omega^{r-1}(M)$ for all values of r. Suppose $\omega \in \Omega^r(N)$ and $\eta = h\omega \in \Omega^{r-1}(M)$ so that $d(h\omega) + h(d\omega) = G^*\omega - F^*\omega$, where G^* and F^* are smooth pull-back maps from $\Omega^r(N)$ to $\Omega^r(M)$ for all values of r. If we let ω be a closed r-form on N (i.e., $d\omega = 0$), we find $d(h\omega) = G^*\omega - F^*\omega$. Consequently, $G^*\omega$ and $F^*\omega$ differ by an exact form $d(h\omega)$ in M, so it follows that G^* and F^* are cohomologous, and the induced pull-back maps $F^*, G^* : H^r_{dR}(N) \to H^r_{dR}(M)$ for all values of r are isomorphic. De Rham cohomology groups for all values of r are homotopically and topologically invariant. It can be concluded that if M and N are homotopy-equivalent smooth manifolds with or without boundary, then $H^r_{dR}(N) \cong H^r_{dR}(M)$ for each value of r. Smooth homotopy equivalence $F : M \to N$ between smooth manifolds induces isomorphisms between de Rham cohomology groups.

From Stokes' theorem and the definition of a period $\int_{z_r} \omega$ where z_r is an r-cycle and ω is an r-form, we can define a bracket $\langle\ ,\rangle : C_r(M) \times C^r(M) \to \mathbb{R}$, where $C_r(M)$ is the free Abelian group of all possible r-chains $c_r \in C_r(M)$ in M and $C^r(M)$ is the free Abelian group of all possible r-cochains (i.e., r-forms) $\omega \in C^r(M)$ in M, by setting $\langle c_r, \omega\rangle := \int_{c_r} \omega$. For every r-form ω, the map $c_r \longmapsto \langle c_r, \omega\rangle$ is linear on $C_r(M)$, and for every $\omega \in C^r(M)$, the map $\omega \longmapsto \langle c_r, \omega\rangle$ is linear on $C^r(M)$. This induces a bracket on the singular r-homology group $H_r(M)$ and the r-th de Rham cohomology group $H^r_{dR}(M)$ by setting

$$\langle [z_r], [\omega]\rangle = \langle z_r, \omega\rangle = \int_{z_r} \omega$$

where $[z_r] \in H_r(M)$ is an r-homology class, $[\omega] \in H^r_{dR}(M)$ is an r-th de Rham cohomology class, z_r is any representative of $[z_r]$ and ω is any representative of $[\omega]$. The period $\int_{z_r} \omega$ only equals zero if the closed r-form ω is exact, otherwise it is a real non-zero number. For any fixed de Rham r-th cohomology class $[\omega]$, the map $f_{[\omega]} : H_r(M) \to \mathbb{R}$ given by $f_{[\omega]}([z_r]) = \langle [z_r], [\omega]\rangle = \int_{z_r} \omega$ is a well-defined linear functional on $H_r(M)$. De Rham's theorem asserts that this correspondence between linear functionals on $H_r(M)$ and de Rham cohomology classes $[\omega] \in H^r_{dR}(M)$ is bijective. The bilinear map $\langle\ ,\rangle : H_r(M) \times H^r_{dR}(M) \to \mathbb{R}$ is non-degenerate in both arguments [27]. Essentially, de Rham's theorem says that the r-th de Rham cohomology group is isomorphic with the dual space of the r-th homology group (i.e., with the space of the singular cohomology group). This implies that an r-form ω is globally exact if and only if all of its periods vanish. This can only happen if there are no r-dimensional holes in the underlying topology of the smooth manifold.

References

1. Neilson, P.D.; Neilson, M.D.; Bye, R.T. A Riemannian geometry theory of three-dimensional binocular visual perception. *Vision* **2018**, *2*, 43. [CrossRef] [PubMed]
2. Neilson, P.D.; Neilson, M.D.; Bye, R.T. A Riemannian geometry theory of synergy selection for visually-guided movement. *Vision* **2021**, *5*, 26. [CrossRef] [PubMed]
3. Neilson, P.D.; Neilson, M.D.; Bye, R.T. The Riemannian geometry theory of visually-guided movement accounts for afterimage illusions and size constancy. *Vision* **2022**, *6*, 37. [CrossRef] [PubMed]
4. Katz, M.; Kruger, P.B. The human eye as an optical system, Chapter 33. In *Duane's Clinical Ophthalmology*; Tasman, W., Jaeger, E.A., Eds.; Lippincott, Williams and Wilkins: Philadelphia, PA, USA, 2006; Volume 1.

5. Armstrong, B.; Aitken, J.; Sim, M.; Swan, N. *Final Report of the Independent Review and Scientific Investigation Panel*; Australian Broadcasting Corporation: Sydney, NSW, Australia, 2007.
6. Sitas, F.; O'Connell, D.L.; van Kemenade, C.H.; Short, M.W.; Zhao, K. Breast cancer risk among female employees of the Australian Broadcasting Corporation in Australia. *Med. J. Aust.* **2010**, *192*, 651–654. [CrossRef] [PubMed]
7. Vieira, L. Torus to Mug Morphing Animation. 2007. Available online: https://commons.wikimedia.org/wiki/File:Mug_and_Torus_morph.gif (accessed on 7 June 2024).
8. Lee, J. *Introduction to Topological Manifolds*, 2nd ed.; Graduate Texts in Mathematics; Springer: New York, NY, USA, 2011; ISBN 978-1-4419-7939-1.
9. Smale, S. Topology and mechanics. I. *Inventiones Math.* **1970**, *10*, 305–331. [CrossRef]
10. Abraham, R.; Marsden, J.E. *Foundations of Mechanics*, 2nd ed.; AMS Chelsea Pub./American Mathematical Society: Providence, RI, USA, 2008; ISBN 978-0-8218-4438-0.
11. Ortega, J.-P.; Ratiu, T. *Momentum Maps and Hamiltonian Reduction*; Progress in Mathematics; Birkhäuser: Basel, Switzerland, 2004; ISBN 978-0-8176-4307-2.
12. Flanders, H. *Differential Forms with Applications to the Physical Sciences*; Revised ed.; Dover Publications: Mineola, NY, USA, 1989; ISBN 978-0-486-66169-8.
13. Lang, S. *Fundamentals of Differential Geometry*; Graduate Texts in Mathematics; Springer: New York, NY, USA, 1999; Volume 191, ISBN 978-1-4612-6810-9.
14. Feynman, R.P.; Leighton, R.B.; Sands, M. The Feynman Lectures on Physics Vol. II Ch. 41: The Flow of Wet Water. 1963/2013. Available online: https://www.feynmanlectures.caltech.edu/II_41.html (accessed on 21 April 2024).
15. Smits, A.J. *Lectures in Fluid Mechanics: Viscous Flows and Turbulence*; Smits, A.J., Ed.; Department of Mechanical Engineering, Princeton University: Princeton NJ, USA, 2009.
16. Brunton, S.L.; Noack, B.R.; Koumoutsakos, P. Machine learning for fluid mechanics. *Annu. Rev. Fluid Mech.* **2020**, *52*, 477–508. [CrossRef]
17. Dobson, C.M. Protein folding and misfolding. *Nature* **2003**, *426*, 884–890. [CrossRef] [PubMed]
18. Dill, K.A.; MacCallum, J.L. The protein-folding problem, 50 years on. *Science* **2012**, *338*, 1042–1046. [CrossRef] [PubMed]
19. The Royal Institution. The Physics of Life: How Water Folds Proteins—With Sylvia McLain; YouTube Video, 45:07. 2017. Available online: https://www.youtube.com/watch?v=SaSntOkK-Vk (accessed on 7 June 2024).
20. Goodale, M.A.; Milner, D. *Sight Unseen: An Exploration of Conscious and Unconscious Vision*; Oxford University Press: Oxford, UK, 2013; ISBN 978-0-19-175300-8.
21. Bullo, F.; Lewis, A.D. *Geometric Control of Mechanical Systems: Modeling, Analysis, and Design for Simple Mechanical Control Systems*; Texts in Applied Mathematics; Springer: New York, NY, USA, 2005; ISBN 978-0-387-22195-3.
22. The Royal Institution. Protein Folding and Particle Accelerators: A New Solution—With Sylvia McLain and Alan Soper; YouTube Video, 7:21. 2016. Available online: https://www.youtube.com/watch?v=uSErKi_HWg8 (accessed on 7 June 2024).
23. Rhys, N.H.; Soper, A.K.; Dougan, L. The hydrogen-bonding ability of the amino acid glutamine revealed by neutron diffraction experiments. *J. Phys. Chem. B* **2012**, *116*, 13308–13319. [CrossRef] [PubMed]
24. Stewart, B.W. "There will be no more!": The legacy of the Toowong breast cancer cluster. *Med. J. Aust.* **2007**, *187*, 178–180. [CrossRef] [PubMed]
25. Lee, J.M. *Riemannian Manifolds: An Introduction to Curvature*; Graduate Texts in Mathematics; Springer: New York, NY, USA, 1997; ISBN 978-0-387-98271-7.
26. Morita, S. *Geometry of Differential Forms*; American Mathematical Society: Providence, RI, USA, 2001; ISBN 978-0-8218-1045-3.
27. Szekeres, P. *A Course in Modern Mathematical Physics: Groups, Hilbert Space and Differential Geometry*; Cambridge University Press: Cambridge, UK, 2004; ISBN 978-0-511-26329-3.
28. Lee, J.M. *Introduction to Smooth Manifolds*, 2nd ed.; Graduate Texts in Mathematics; Springer: New York, NY, USA, 2013; ISBN 978-1-4419-9981-8.
29. Tu, L.W. *Differential Geometry: Connections, Curvature, and Characteristic Classes*; Graduate Texts in Mathematics; Springer International Publishing: Cham, Switzerland, 2017; ISBN 978-3-319-55082-4.
30. Lee, J.M. *Introduction to Riemannian Manifolds*, 2nd ed.; Graduate Texts in Mathematics; Springer International Publishing: Cham, Switzerland, 2018; ISBN 978-3-319-91754-2.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

AppliedMath

MDPI

Article

Innovative Ways of Developing and Using Specific Purpose Alternatives for Solving Hard Combinatorial Network Routing and Ordered Optimisation Problems

Santosh Kumar [1,*] and Elias Munapo [2]

[1] Department of Mathematical and Geospatial Sciences, School of Sciences, RMIT University, Melbourne, VIC 3000, Australia
[2] School of Economics and Decision Sciences, North West University, Mafikeng Campus, Mafikeng 2735, South Africa; emunao@gmail.com
* Correspondence: santosh.kumarau@gmail.com

Abstract: This paper reviews some recent contributions by the authors and their associates and highlights a few innovative ideas, which led them to address some hard combinatorial network routing and ordered optimisation problems. The travelling salesman, which is in the NP hard category, has been reviewed and solved as an index-restricted shortest connected graph, and therefore, it opens a question about its 'NP Hard' category. The routing problem through 'K' specified nodes and ordered optimum solutions are computationally demanding but have been made computationally feasible. All these approaches are based on the strategic creation and use of an alternative solution in that situation. The efficiency of these methods requires further investigation.

Keywords: shortest connected network; minimum travelling salesman tour; NP Hard problems; path through K specified nodes; index-balancing theorem; index-restricted connected network; creation of an alternative solution; ordered optimal solutions

check for updates

Citation: Kumar, S.; Munapo, E. Innovative Ways of Developing and Using Specific Purpose Alternatives for Solving Hard Combinatorial Network Routing and Ordered Optimisation Problems. *AppliedMath* **2024**, *4*, 791–805. https://doi.org/10.3390/appliedmath4020042

Academic Editor: Carlo Bianca

Received: 18 April 2024
Revised: 6 June 2024
Accepted: 6 June 2024
Published: 18 June 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Mathematics of OR has provided support in many real-life industrial, business, and government decision-making processes. Changing technology has always been a source of new challenges, and accordingly, new mathematical methods and tools have been developed for providing necessary decision-making support for better and informed decisions. Routing problems and their variants in graph theory have attracted a lot of attention from researchers due to various industrial applications. Shortest route, shortest connected graph, and the travelling salesman are just a few of them. Variants of these problems have been identified from time to time. Ordered optimisation is another field where obtaining an ordered optimum, i.e., the k^{th} *best*, $k \geq 2$ solution, can be a highly challenging task. In this paper, we review a few specific innovative ideas for developing an alternative in a given situation, and use that alternative solution to solve some difficult combinatorial hard optimisation situations; for example, we find the following:

(1) The minimum travelling salesman tour, which is in the 'NP Hard' category, is solved as an index-restricted minimum connected graph;
(2) The path passing through a set of specified nodes, which is computationally demanding due to the path being constrained to pass through a set of specified nodes, is also solved as an index-restricted minimum connected graph;
(3) The second-best optimal solution for the shortest route and the assignment problem.

All these approaches develop an alternative solution and use that solution in a strategic way to solve the problem under consideration. These new approaches need further investigation to establish their efficiency, if any. That work is currently being conducted. Among the above problems, the minimum travelling salesman tour (MTST) has been classified

as 'NP Hard'; see [1]. Similarly, determination of the shortest route through 'K' specified nodes and finding an ordered optimum solution have been classified as computationally demanding; see [2,3]. approach based on the index-restricted shortest connected graph for the travelling salesman problem raises a question—"Is the TSP really an 'NP-Hard' problem?" A similar question arises again for the path through K specified nodes. The question is, "Is the path through specified nodes computationally demanding?" The development of the restricted minimum spanning tree is based on the idea of developing an alternative and making use of that alternative to find the index-restricted minimum spanning tree.

We extend the idea of developing an alternative solution for the shortest route and an assignment problem and find the second-best optimal solution for these problems, without excessively increasing computations. This approach is in its initial stage, and more computational work is necessary to establish its computational efficiency for different problems.

Many real-life situations can be represented in the form of a network, which is made of nodes and links. These links are formed by joining the nodes, and weights associated with these links provide a kind of measure of the relationship between the nodes. The network representing a physical situation can be a directed or a non-directed network. Analysis of a non-directed network is usually more demanding compared to a directed graph, simply because directions in a directed graph provide additional information, which is missing in a non-directed network. Many industrial situations have been modelled as directed networks and were analysed by using dynamic programming techniques; see [4–6]. These problems, once transformed into a network model, were analysed by using the network-related theory and tools. The outcome of the analysis can be interpreted in terms of the real-life situation that was modelled as a network. For recent developments, see [7,8].

In general, a network can be represented by $G(N, L)$, where N represents a set of some entities represented by the nodes $N = \{1, 2, \ldots, n\}$ and L is the set of links joining various nodes representing properties associated with pairs of these nodes. These links are denoted by (i, j), where i and j are the nodes $1, 2, \ldots, n$ and $i \neq j$. In weighted networks, we associate a number with each link, providing a quantitative measure of the relationship between those two nodes. For example, it may just be the distance between the two nodes, or it may represent the time of travel between the pair of nodes, i and j. This weight associated with the link (i, j) is denoted by c_{ij}, which can be a positive or a negative quantity, depending on the situation.

Some ideas used in this paper have been reviewed in Section 2, and their applications, in the form of the travelling salesman and routing problems, have been presented in Section 3. The ordered optimum problem has been discussed in Section 4. Finally, concluding remarks are presented in Section 5.

2. Modification of a Classical Shortest Connected Network Problem and Its Applications

The shortest connected network (SCN) of any given undirected connected network $G(N, L)$ is a classical network problem. It comprises $(n - 1)$ links, where n represents the number of nodes in the given network $G(N, L)$. Since the given network is connected, there must be a unique path from every node to every other node, and that path will be free of loops. In such a connected graph, if any link is removed, it will make the graph disconnected. This SCN can be obtained by a greedy approach according to [9–11]. A modification of the shortest connected graph problem was discussed in [12], which is briefly presented below.

2.1. Modification of the Shortest Connect Network (SCN)

2.1.1. Statement of the Modified Shortest Connected Network

For a network $G(N, L)$, the SCN can be obtained by any known method. From the SCN, one can find a node index for each node. The node index is the number of links emanating from that node in the SCN. Let n_i represent the index of the node i, where $1 \leq n_i \leq (n - 1)$. In the modification discussed in [12] in which the SCN is restricted, the

node index n_i satisfies $1 \leq n_i \leq 2$ for each node. This means that if the original SCN does not satisfy the restriction on the index value on any node, the SCN obtained by using the classical approach must be modified to satisfy the condition $1 \leq n_i \leq 2$, $\forall i$. In the given index-restricted network, two nodes will have an index value of 1 and the other nodes will have an index value of 2; thus, they will form a path joining two nodes and passing through all other remaining nodes of the network. We may assume that in the given connected network, all nodes are such that at least two links emanate from each node.

2.1.2. Some Essential Terms and a Theorem to Modify the SCN Obtained by the Greedy Approach to an Index-Restricted SCN (IRSCN)

The necessary terms are the following:

T1. Index of a node: Since each selected link generates two indices, the total number of indices of the selected $(n-1)$ links forming the SCN will be $2(n-1)$. For a path passing through all nodes, two of these nodes will have only one index value (from where the path starts to where it ends), and the remaining $(n-2)$ nodes will have an index of 2. The index-restricted SCN or IRSCN will form a path joining two nodes, which passes through all other remaining nodes of the network.

T2. Basic and non-basic links: A link joining two nodes *i and j* is said to be basic if $x_{ij} = 1$, i.e., that link belongs to the SCN and, if $x_{ij} \neq 1$, then it is a non-basic link.

T3. High- and low-index nodes: Once an SCN has been identified using any known method, we know the values of node indices for all nodes. These values are given by $1 \leq n_i \leq (n-1)$. They must be changed to satisfy the condition $1 \leq n_i \leq 2$ $\forall i$, i.e., for the origin and the destination nodes, the *index* $n_i = 1$, and for all other remaining $(n-2)$ nodes, $n_i = 2$. The high- and low-index nodes can be easily identified by their current value against the desired node index value for that node.

T4. Neighbouring nodes and links: A node *i* is said to be a neighbouring node to node *j*, if the nodes *i* and *j* are connected by a single link. In a completely connected network, all nodes are neighbouring nodes as all nodes are connected by a direct link. A **neighbouring link** is defined as a link that emanates from a neighbouring node.

Theorem 1. *Index-balancing theorem: Adding the same constant to all links emanating from the same node does not change the relative merit of any given selected link in the SCN obtained by the greedy approach. However, the index-balancing theorem can* ***create alternatives****. From the SCN obtained by the greedy approach, we know the index for each node; therefore, low- and high-index nodes can be easily identified. The index-balancing theorem commences its work by adding constants to all links that emanate from the same node, which has been classified as a high-index node.* ***The idea is to create alternatives for the selection of a link, and thus, it reduces the index at a high-index node and increases the index of a low-index node; when a link from a high-index node is removed and placed on a node with a low index value, the imbalance is reduced*** *This process is repeated until all nodes attain the required index and satisfy all index restrictions.*

Theorem 2. *For any given SCN solution, a basic link can be altered by adding a constant to all emanating links from that node, i.e., changing the status of a link from basic to non-basic and that of some other link from non-basic to basic. Repeated applications can generate an SCN path joining two nodes passing through all other nodes in a network, i.e., it will satisfy the index condition on each node.*

2.1.3. Numerical Illustration

Consider a simple but interesting trivial example: the c_{ij} values given in Table 1.

Table 1. A 6-node completely connected network with weighted links.

i\j	**1**	**2**	**3**	**4**	**5**	**6**
1	-	1	1	1	1	1
2	1	-	2	2	2	2
3	1	2	-	3	3	3
4	1	2	3	-	4	4
5	1	2	3	4	-	5
6	1	2	3	4	5	-

The SCN can be obtained by using the greedy approach, which is shown in Figure 1. One can start from any node, say node 6, and select the link (6,1) as the first link; the next selected links can be (1,2), (1,3), and so on. The SCN will be as given in Figure 1.

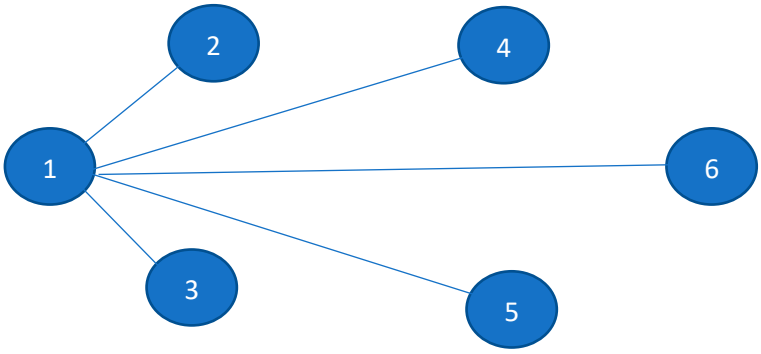


Figure 1. Shortest connected graph of the network from Table 1.

The incidence matrix for the SCN in Figure 1 is shown in Table 2.

Table 2. Incidence matrix of the SCN in Figure 1.

Nodes	1	2	3	4	5	6
Incidence	5	1	1	1	1	1

This means that if we want a path, say, between nodes 5 and 6, nodes 5 and 6 are balanced with a node index value of 1, and all other nodes are imbalanced. Node 1 is a high-index node, and nodes 2, 3, and 4 are low-index nodes. As an application of the index-balancing theorem, add a constant value of 1 to all links emanating from the high-index node, which is node 1; the relative merits of the links (1,2), (1,3), (1,4), (1,5), and (1,6) will remain unchanged. However, an alternative has been developed by adding 1 to all links emanating from node 1. The alternative is a link starting from node 2, instead of a link starting from node 1. Therefore, the link (1,4) can be removed and replaced by the link (2,4). This change is possible as an alternative has been created to start from node 2. Note that it will lower the index of the high-index node 1 by 1, and it will increase the index value of the low-index node 2 by 1. This is shown in Figure 2. Similarly, again, if 1 is added to all links emanating from nodes 1 and 2, the link (1,5) can be removed and replaced by the link (3,5). One more application of the index-balancing theorem will result in the SCN as shown in Figure 3.

However, nodes 3 and 4 are still low-index nodes and node 1 is a high-index node. Another application gives the index-restricted SCN as shown in Figure 3.

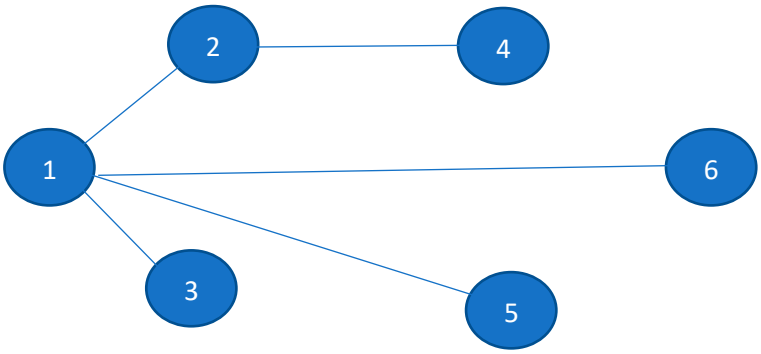


Figure 2. Improved SCN.

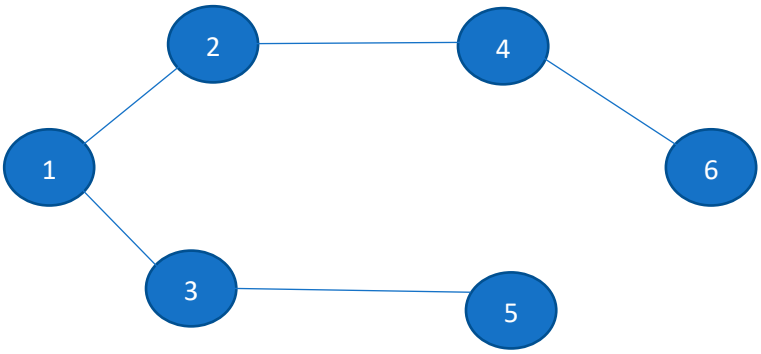


Figure 3. Index-restricted SCN of the network in Table 1.

Suppose our interest was to find the shortest connected graph with an index restriction of 2 on all nodes except the starting and finishing nodes, where the index is restricted to 1. The total incidence of the selected SCN will be $2(n-1) = 10$, as in this example, $n = 6$. In other words, the IRSCN will be a path between two nodes, and it will pass through $(n-2)$ number of nodes. The connected network in Figure 3 is a required IRSCN, and it may not be unique.

3. The Index-Restricted SCN and Its Applications

3.1. The Travelling Salesman Problem

The minimum travelling salesman tour is a well-known problem in graph theory. It deals with the determination of a minimum-cost tour passing through all nodes and returning to the home city. This classical travelling salesman problem has a lot of industrial applications, and it has been classified as an NP Hard problem; see [1,13–15]. Another problem dealing will all nodes in graph theory is the shortest connected graph, which has been obtained by the greedy approach; see [11]. In an earlier paper [12], the authors developed an approach to convert the shortest connected network (SCN) to an index-restricted shortest connected graph, as discussed in Section 2. The index-restricted SCN can be used to find the minimum travelling salesman tour. The index-restricted SCN is developed by using the index-balancing theorem [12]. It, therefore, raises a question: **is the travelling salesman problem really NP Hard?** The SCN in a modified form has been used to find the travelling salesman tour in different ways, [16,17], and finally, the minimum spanning tree under the index restriction has been established as equal to the travelling salesman tour [18]. We briefly explain these approaches in Sections 3.1.1–3.1.3.

3.1.1. A Minimum Spanning Tree-Based Approach to Find the Minimum Travelling Salesman Tour (MTST) of the Network $G(N, L)$: Approach 1 (Complexity r_{C_2})

We propose starting the minimum travelling salesman tour from the node with a minimum node index value, i.e., the node from where a minimum number of links are emanating. It is also true that every node will have two links, which will form a part

of the minimum travelling salesman tour. If the minimum-index node is denoted by the node p and its index value is r, i.e., r links are emanating from the p, two of these links will form a part of the minimum salesman tour. We examine all such possibilities. Let us denote these r links emanating from the node p by p_r. The minimum travelling salesman tour can be identified by evaluating an r_{C_2} number of index-restricted shortest connected trees from the modified network $G((N-1),(L-p_r))$. This symbol indicates that the modified network is obtained after removing the node p and all links that are emanating from the node p. Hence, the modified network will have $(N-1)$ nodes and the number of links will be given by $(L-p_r)$. Every index-restricted tree becomes a travelling salesman tour when the initial two arbitrarily selected edges are joined to that path or that IRSCN. The required minimum salesman tour becomes the one with the minimum value of these r_{C_2} travelling salesman tours.

In the above, we suggested starting the travelling salesman tour from a node with a minimum node index value. To explain this point, one may refer to the network in Figure 4 and note that the index value of node 1 is 3; nodes 2, 3, and 4 have an index value of 5; and anodes 5 and 6 have an index value of 4. These index values are the number of emanating links from that node. To minimise the computational work, we may prefer to start the salesman tour from node 1, where only 3_{c_2} number of evaluations will be required; more would be required if we started from any other node.

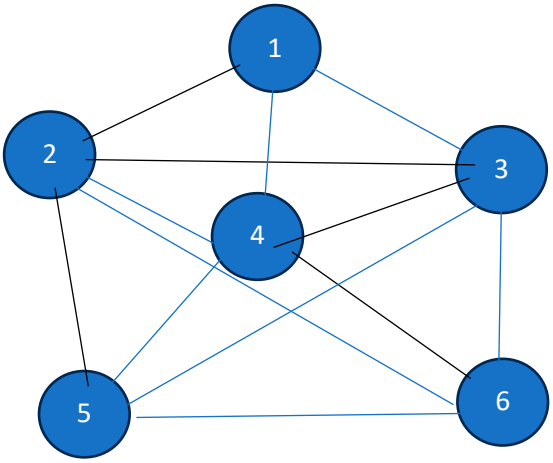


Figure 4. Given $G(6,13)$ network.

This approach was discussed in [17].

3.1.2. A Minimum Spanning Tree-Based Approach to Find the Minimum Travelling Salesman Tour (MTST) of $G(N,L)$: Approach 2 (Complexity p_r)

As discussed above, we can once again use the index-restricted shortest connected graph to find the minimum travelling salesman tour. Consider a slightly different modification of the given network. This time, we consider a modification and denote this modified network by $G(N,(L-1))$. It means there is no change in the number of nodes, and only one link connecting node p to some other node, say to node q, has been removed. Let these r links be denoted by $\{(p,q_1),(p,q_2),\ldots,(p,q_r)\}$. In this approach, it is proposed to find r number of shortest connected trees joining the nodes (p,q_1), (p,q_2), …, (p,q_r), after removing only one link at a time. For example, let us remove the link (p,q_1) and find the IRSCN joining the nodes p and q_1. After joining the link (p,q_1), the index-restricted path will become a travelling salesman tour. Similarly, we find the index-restricted shortest connected networks joining the nodes p and q_r, $r = 1, 2, \ldots, r$. After joining the link (p, q_r) and the minimum shortest connected graph, it will form a salesman tour and the minimum of these r tours will be the required minimum travelling salesman tour. Therefore, the complexity is reduced to r index-restricted shortest connected trees compared to the complexity r_{C_2}, as was the case in the previous approach.

This approach was discussed in [16].

3.1.3. A Minimum Spanning Tree-Based Approach to Establish That the Minimum Salesman Tour Is Equivalent to Index-Restricted Minimum Spanning Tree: Approach 3 (Complexity 1)

Once again, let the selected node be p. The given network $G(N, L)$ can be modified by adding node p', and the links emanating from node p are also added as links emanating from node p'; the given network is transformed to $G\{(N+1), (L+p_r)\}$. It should be noted that this time, instead of removing some nodes or links, we are adding node p and all links emanating from this node to the modified network. To maintain a difference in the selected node, its duplicate, and, similarly, the links that are emanating from the duplicate node, we have used the notation p and p'. Since node p and all links emanating from this have been duplicated, their use is possible only once as the existence of the duplicate links is virtual. Node p' is also virtual; however, when we overlap p' on p, we are talking about one node p, which is real. In the following, we explain how to take care of these virtual entities, i.e., node p' and the associated virtual links from node p'.

Since nodes p and p' are essentially the same, the index-restricted path from node p to node p' will form a salesman tour. Thus, in this case, with the help of some creative thinking, we find the index-restricted shortest connected tree joining nodes p and p', and it gives us the required minimum salesman tour.

In the above modification of the given network, the usual greedy approach is not directly applicable to finding the minimum spanning tree. The p_r number of extra links added to the node p' is not real; only one of them exists when selected to be part of the minimum spanning tree, and the other link must vanish. This means that as soon as one of these links becomes basic, the other must disappear. In other words, if the link (p, q_j) is basic, the link (p', q_j) will have an infinite length. Similarly, if the duplicate link is basic, the original link is changed to infinite value. However, if these links are non-basic, they can co-exist as non-basic links, as they have no meaningful existence in the travelling salesman tour.

Modified Network and Steps for Obtaining the Minimum Spanning Tree of $G\{(N+1), (L+p_r)\}$

Step 1: For the given network $G(N, L)$, develop the network $G\{(N+1), (L+p_r)\}$ and arrange links in increasing order. Initially, all links are non-basic, so all duplicate links will be the same as in $G(N, L)$. Links that have been duplicated are called Type 2 links and those which have not been duplicated are called Type 1 links. The number of Type 1 links is $(L - p_r)$ and that of Type 2 is 2 p_r. Type 2 links are such that at least two links have the same length value, initially.
Step 2: Set a counter K = 1. Select the link of minimum length, include it in the minimum spanning tree, and go to Step 3.
Step 3: If the selected link is Type 1 and K < n, set K = K + 1, and select the next minimum length. If K = n, go to Step 7.
Step 4: If the selected edge was Type 2 and K < n, first, set the length of the duplicate link equal to ∞ and then rearrange lengths in increasing order. Set K = K + 1 and select the next minimum. If K = n, go to Step 7.
Step 5: If the selected link forms a cycle with the spanning tree formed so far, discard it, or else include it. If this last link was Type 1, return to Step 3. If the selected link was Type 2, go to Step 6.
Step 6: For the selected edge, check the following:

6.1: If the inclusion of the edge does not lead to an isolated node in the network, discard it.
6.2: If alternatives exist, go in favour of the edge that develops better balance among the index of nodes forming the minimum spanning tree.

Return to Step 4 as the selected link is of Type 2.

Step 7: Stop as the minimum spanning tree has been obtained.

Numerical Illustration

Consider the following network $G(6,13)$ shown in Figure 4 the link weights are given in Table 3.

Table 3. Link distances for the network in Figure 4.

F\T	1	2	3	4	5	6
1	-	12	10	10	-	-
2	12	-	15	11	11	16
3	10	15	-	7	14	12
4	10	11	7	-	10	11
5	-	11	14	10	-	9
6	-	16	12	11	9	-

Now, consider the modified network that will be $G((6+1),(13+3))$ or $G(7,16)$ as shown in Figure 5, where the node 1 has been duplicated as 1′.

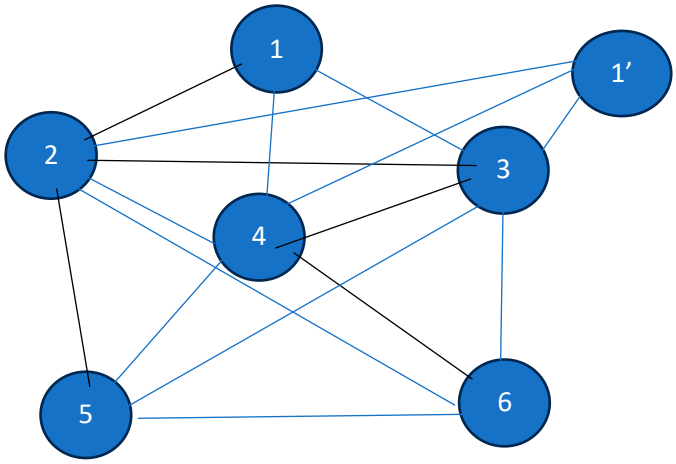


Figure 5. Modified network $G(7,16)$.

Type 1 links are (2,3), (2,4), (2,5), (2,6), (3,4), (3,5), (3,6), (4,5), (4,6), and (5,6).

Type 2 links are (1,2), (1,3), (1,4), (1′,2), (1′,3), and (1′,4).

Arranging links in increasing order, we have 7 (3,4)T1, 9(5,6)T1, 10(4,5)T1, 10(1,3)T2, 10(1,4)T2, 10(1′,3)T2, 10(1′,4)T2, 11(2,4)T1, 11(2,5)T1, 11(4,6)T1, 11(4,6)T1, 12 (3,6)T1, 12(1,2)T2, 12(1′,2)T2, 15 (2,3)T1, and 16(2,6)T1.

Determination of MST: Link 7(3,4) T1 included, link 9(5,6) T1 included, 10(4,5) T1 included, 10(1,3) T2 included; however, now the link (1′,3) is changed to infinity, so the list of increasing order is changed as given below:

10(1,4) T2, 10(1′,4) T2, 11(2,4) T1, 11(2,5) T1, 11(4,6) T1, 11(4,6) T1, 12 (3,6) T1, 12(1,2) T2, 12(1′,2) T2, 15 (2,3) T1, and 16(2,6) T1, ∞(1′,3) T2.

Note that the greedy approach will identify the next link (1,4) of length 10, but it makes a circuit, and if we go in favour of (1′,4), it also makes a circuit 1-3-4-1′, as nodes 1 and 1′ are the same. It is also excluded. The next link (2,4) is excluded, and the link (2,5) is included. The link (4,6) is excluded. Finally, we have Figure 6.

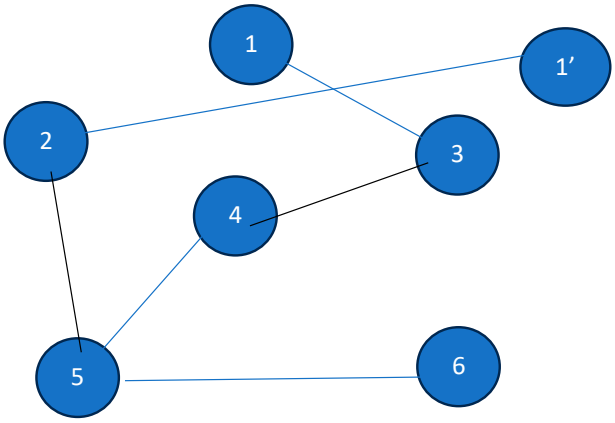


Figure 6. Spanning trees by greedy approach.

Looking at Figure 6, we find that nodes 5 and 6 are imbalanced as the node index at node 5 is 3 and the node index at node 6 is 1. If 1 is added to all nodes emanating from node 5, the weight associated with the link (4,5) will become equal to the weight of the link (4,6); therefore, it will result in an alternative. It creates a possibility of removing the link (4,5) and replacing it by the link (4,6). This is shown in Figure 7.

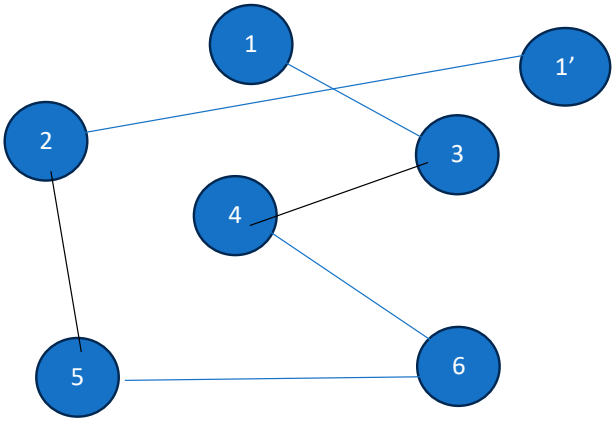


Figure 7. The path 1 -> 3 -> 4 -> 6 -> 5 -> 2 -> 1′.

This path in the network $G(7, 16)$ has an interpretation of an index-restricted minimum spanning tree joining nodes 1 and 1′ under the node index restriction and a salesman tour interpretation for the network $G(6, 13)$. It is given in Figure 8. The cost for this tour will be 12(1,2) + 11(2,5) +9(5,6) + 11(6,4) + 7(4,3) + 10((3,1) = 60 (1-> 2 -> 5 -> 6 -> 4 -> 3 -> 1).

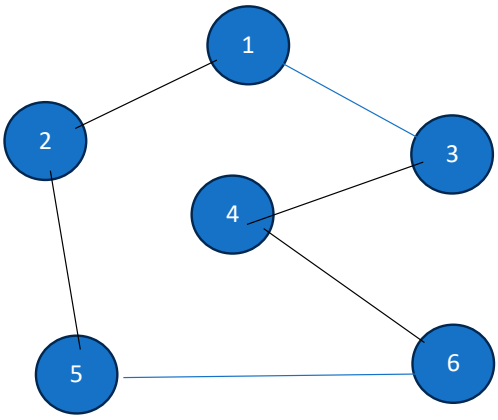


Figure 8. The spanning tree and the minimum travelling salesman tour.

3.2. Shortest Path in a Non-Directed Network under the Condition of Passing through K Specified Nodes

Kalaba [19] suggested a constrained routing problem that finds a path from the origin node to the destination node, passing through K number of specified nodes. Using the dynamic programming technique, this problem was attempted in [20]. When the specified set is empty, the problem is reduced to an ordinary routing problem, and when the set of specified nodes contains all nodes and returns to the node we started from, it becomes the salesman tour problem. Due to the combinatorial nature of the problem, it is difficult to solve this problem by the dynamic programming technique; however, when this problem is attempted by using the shortest spanning tree under the node index restriction, the computational load becomes manageable. All nodes are included in the SCN, and the combinatorial aspect, which makes the problem computationally demanding for the dynamic programming approach, is changed. For the specified nodes, we simply apply an index restriction by making sure that the specified node is visited on route. It is achieved by imposing an index restriction of 2 or more and making sure that the restriction is satisfied. See, for example [21], where authors considered an industrial application of service calls, and the problem was developed in mathematical form as a path through K specified links. All these routing problems, where the path is restricted to pass through a set of specified nodes or links, are computationally demanding, but when these problems are attempted as an index-restricted SCN, the computational complexity is reduced. In fact, it is easy to find the index-restricted SCN under any condition, such as

$$1 \leq n_i \leq k_i, \text{ where } k_i \text{ is a given number between } 1 \leq k_i \leq (n-1).$$

4. Ordered Optimum

The best solution of K^{th} $(K \geq 2)$ in general is an ordered optimisation problem. The K^{th} best, $(K \geq 2)$, is defined as the best solution after excluding $(K-1)$ ordered optimal solutions. There is no well-defined theory for the determination of ordered optimal solutions to optimisation problems, except for [22], where the authors developed a characteristic equation that identifies ordered optimum solutions for a pure linear integer programming problem. Here, we will consider the mathematics of OR problems to find the second-best solution, using the concept of an alternative optimal solution. Many suggestions have been proposed in the past, but they can be computationally demanding. These problems arise in various real-life situations, and in general, the best or optimal solution can be identified as all existing methods have defined conditions for the optimality of the solution, and when those conditions are satisfied, one can easily conclude that the optimal solution has been obtained. However, the existence of the optimality conditions for the K^{th} best solution, $K \geq 2$, has not been defined; rather, the K^{th} optimal is defined as the best, excluding a $(K-1)$ number of best solutions. A need for the K^{th} best solution arises in many different situations (see [2] for an example), where the K^{th} best is required for a linear program to analyse the extreme point mathematical programming problem.

The K^{th} best solution is required for different reasons in different situations. In this section, we approach the problem of ordered optimum, which can be computationally demanding, again, by developing an alternative solution and using that solution to identify the 2nd-best solution. Two problems have been used to illustrate the determination of the 2nd-best solution: (1) the 2nd-shortest route for a shortest route problem, and (2) the assignment problem. For the assignment problem, Murthy [3] proposed a method for ranking all the assignments in order of increasing costs. In an nXn assignment problem, Murthy [3] obtained the second-best solution by solving n assignment problems by assigning high cost to one of the optimal assignments at a time, and the best of all these n solutions was the second-best solution for that problem. His idea is that the second best will differ from the best in at least one assignment, and he evaluates all such possibilities.

In this paper, the idea is to increase the length of the current optimal solution by a certain amount to establish two or more alternative optimal solutions. One of these

alternative solutions is the existing optimal solution and the other is a newly created alternative optimal solution. Since the value of the optimal solution is increased in a controlled way, these two alternative optimal solutions have an interesting interpretation of the best and the 2nd-best solutions with respect to the original cost information. We illustrate the idea by taking the shortest route problem in a network and an assignment problem to find the best and 2nd-best solutions.

4.1. The 2nd-Best Route for the Conventional Routing Problem in a Directed Network

Numerical Illustration to Find the 2nd-Best Shortest Route

Let us apply the idea of developing alternative solutions to the network routing problem considered in Figure 9, which was transformed into Figure 10 to find the optimal solution.

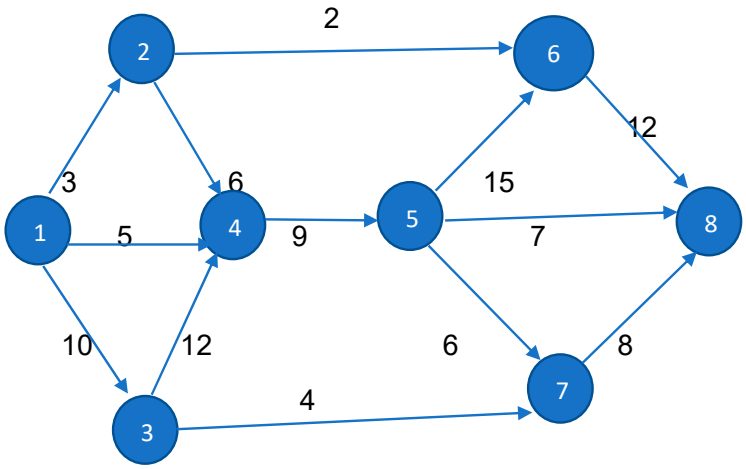


Figure 9. Directed network.

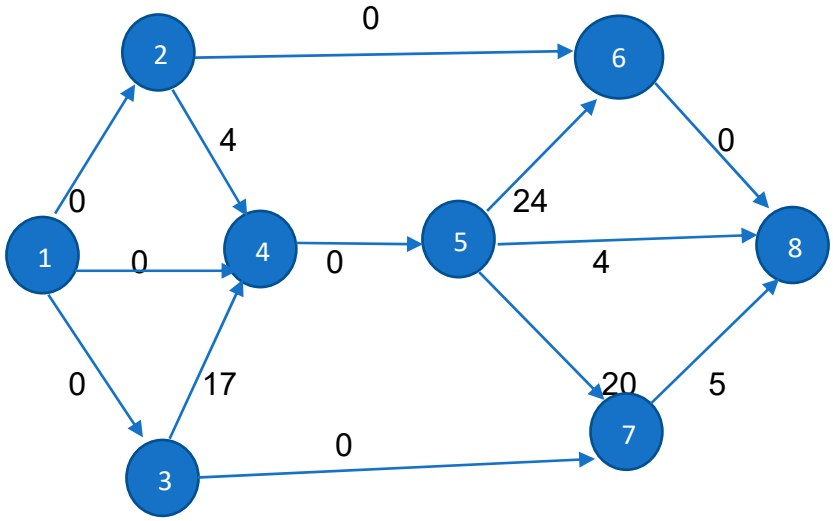


Figure 10. The final reconstructed network after 7 iterations.

The optimal solution for the network in Figure 9 was identified as 17 units, and the path was 1 -> 2 -> 6 -> 8.

We use the minimum-weight labelling approach developed in [23]. The transformed network after seven iterations of the minimum-weight label approach is shown in Figure 10.

The labels generated while applying the minimum-weight labelling approach are summarised in Table 4.

Table 4. Labels obtained by the minimum-weight label method.

Node	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Label	(1,0)	(1,3)	(1,10)	(1,5)	(4,14)	(2,5)	(3,14)	(6,17)

We arrive at node 8 from node 6, arrive at node 6 from node 2, and arrive at node 2 from node 1, resulting in the optimal path being 1 -> 2 -> 6 -> 8 and the minimum cost being 17.

The minimum non-zero cost is 4 in Figure 10, for two links. Changing the weight associated with the link (2,4) does not generate a second path, but changing the weight associated with the link (5,8) will generate a second path. We can generate two alternative paths when the link (6,8) is increased by 4 units. When the problem is solved again, we have two paths, which give the best and the 2nd best.

The best path is, as was given above, a cost of 17, and the route is 1 to 2 to 6 to 8.

The second-best path will be a cost of 21, and the route will be 1 to 4 to 5 to 8.

Increasing the link weight of (2,6) or (6,8) by 4 units will give rise to two alternative solutions.

The main idea is to increase the length of the current optimal solution by 4 units, and if we can establish two alternative optimal paths with a total length of 21 units after increasing the length of the current optimal path by 4 units, the previous path is optimal and the other is a new path that has a length equal to 21. The previous path will be the best solution and the other path will be the 2nd-best solution. Therefore, from these two alternative paths, we establish the optimal and 2nd-best solutions to the shortest route problem.

This idea can be easily extended to the non-directed network by using the labelling approach discussed in [24].

4.2. Numerical Illustration for the 2nd-Best Solution to an Assignment Problem

We reconsider the assignment problem considered by Murthy [3]. The assignment problem is given in Table 5.

Table 5. Assignment cost values for Murthy's Illustration.

R\C	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9 V↓1**	**10**
1 →	7	51	52	87	38	60	74	66	0	20
2	50	12	0	64	8	53	0	46	76	42
3	27	77	0	18	22	48	44	13	0	57
4	62	0	3	8	5	6	14	0	26	39
5	0	97	0	5	13	0	41	31	62	48
6	79	68	0	0	15	12	17	47	35	43
7	76	99	48	27	34	0	0	0	28	0
8	0	20	9	27	46	15	84	19	3	24
9	56	10	45	39	0	93	67	79	19	24
10	27	0	39	53	46	24	69	46	23	1

→ and ↓ indicate one horizontal and one vertical line.

The optimal solution with a total cost of 0 is given by the following assignments, which is denoted by A (1), given below:

$$A\ (1) = \{(1,9), (2,7), (3,3), (4,8), (5,6), (6,4), (7,10), (8,1), (9,5), (10,2)\}.$$

These allocations in Table 5 are shown in red. Vertical lines are drawn to cover zero entries in columns 9, 3, 4, 1, 5, 2, and 7. Rows are covered by three horizontal lines 4, 5, and

7. For the second-best assignment, Murthy solved 9 problems by setting one assignment from the optimal solution at a time equal to ∞. For example, if we set the link (1,9) = ∞, the problem can be solved again. Similarly, Murthy solved 9 assignment problems, which were summarised by Murthy in [3] as follows:

$$M_1 = \{\overline{(1,9)}\} = 10,$$
$$M_2 = \{(1,9),\ \overline{(2,7)}\} = 14$$
$$M_3 = \{(1,9),\ (2,7),\ \overline{(3,3)}\} = Cost\ 14$$
$$M_4 = \{(1,9),\ (2,7),\ (3,3),\ \overline{(4,8)}\} = Cost\ 1$$
$$M_5 = \{(1,9),\ (2,7),\ (3,3),\ (4,8),\ \overline{(5,6)}\} = Cost\ 12$$
$$M_6 = \{(1,9),\ (2,7),\ (3,3),\ (4,8),\ (5,6),\ \overline{(6,4)}\} = Cost\ 53$$
$$M_7 = \{(1,9),\ (2,7),\ (3,3),\ (4,8),\ (5,6),\ (6,4),\ \overline{(7,10)}\} = Cost\ 45$$
$$M_8 = \{(1,9),\ (2,7),\ (3,3),\ (4,8),\ (5,6),\ (6,4),\ (7,10),\ \overline{(8,1)}\} = Cost\ 47$$
$$M_9 = \{(1,9),\ (2,7),\ (3,3),\ (4,8),\ (5,6),\ (6,4),\ (7,10),\ (8,1)\ \overline{(9,5)}\} = Cost\ 56$$

From the above, he concluded that the second-best solution corresponds to the string M_4. This string of assignments is given by

A (2) = {(1,9), (2,7), (3,3), (4,2), (5,6), (6,4), (7,8), (8,1), (9,5), (10,10)} with total cost 1.

Now, we apply the concept of creating an alternative. From Table 5, the minimum non-zero element is 1 in the cell (10,10). In column 10 and row 10, we have an allocation in the optimal solution denoted by the string A (1). These allocations are in the cells (7,10) and (10,2). To create an alternative, the cost associated with the element (7,10) is increased by one unit or the cost associated with the element (10,2) is increased by one unit. Therefore, either column 10 or row 10 will be without a zero number, and by subtracting the minimum, we will create two zero elements in the cells (10,2) and (10,10) in row 10, which will lead to two alternative solutions. A similar case will arise with an increase in cost in the cell (7,10). The two alternative optimal solutions will lead to the best and 2nd-best solutions.

First solution: {(1,9), (2,7), (3,3), (4,2), (5,6), (6,4), (7,8), (8,1), (9.5), (10,10)}.

Second solution: {(1,9), (2,7), (3,3), (4,8), (5,6), (6,4), (7,10), (8,1), (9,5), (10,2)}.

With respect to the data in Table 5, the first solution is the second best with a total cost of 1. The second solution is the optimal solution with a total cost of 0, as obtained in [3].

The challenge remains to find the K^{th} best when $K > 2$.

5. Concluding Remarks

In this paper, we have considered a few combinatorial network optimisation problems, which were transformed into a form that became computationally manageable. The central idea behind these transformations is the development of an alternative, and by using that alternative, we obtained the desired solution. The problems addressed in this paper are the following:

(1) The travelling salesman problem;

(2) The path through K specified nodes;

(3) The ordered optimal solution, i.e., the second-best solution for the shortest route and assignment problem.

All approaches use the alternative solution, which has been created in each situation differently and used to reduce the computational load. The combinatorial aspect of the travelling salesman and the path through specified nodes was reduced to a manageable form by the development of the index-restricted minimum spanning tree.

The second-best solution for the shortest route and the assignment problem was obtained by increasing the optimal solution to a controlled and strategic position to obtain two optimal solutions, one of which was the optimal and the other was the second-best solution.

Our approach also raised a few questions that deserve further investigation.

1. Is the minimum travelling salesman tour problem really NP Hard?
2. Is the constrained routing problem computationally demanding?

3. Ordered optimisations require further investigation to find the K^{th} best when $K \geq 3$ for the two cases discussed in this paper, and for other situations.

This paper has been confined to some classical problems due to the authors' long association with them; however, it would be desirable to investigate the application of these ideas to routing problems considered by [25–27].

The approaches reviewed in this paper need further investigation to establish the computational efficiency of various situations discussed in this paper. This is under investigation and will be reported when ready.

Author Contributions: Conceptualization, S.K. and E.M.; Software, S.K. and E.M.; Formal analysis, E.M.; Investigation, S.K.; Data curation, E.M.; Writing—original draft, S.K. and E.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in the study are included in the article.

Acknowledgments: The authors express their sincere thanks to all anonymous reviewers for their constructive suggestions and comments.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Nemhauser, G.; Wolsey, L.A. *Integer and Combinatorial Optimization*; John Wiley & Sons: Hoboken, NJ, USA, 1988. [CrossRef]
2. Kumar, S.; Wagner, D. Some algorithms for solving extreme point mathematical programming problem. *N. Z. Oper. Res. J.* **1979**, *7*, 128–149.
3. Murty, K.G. An algorithm for ranking all the assignments in order of increasing cost. *Oper. Res.* **1968**, *16*, 682–687. [CrossRef]
4. Bellman, R.; Dreyfus, S. *Applied Dynamic Programming*; Princeton University Press: Princeton, NJ, USA, 1962.
5. Hastings, N.J. *Dynamic Programming with Management Applications*; OR Series; American Mathematical Society: Butterworth, Malaysia; London, UK, 1973.
6. Saksena, J.P. *Dynamic Programming*; Siddarth Publications: Delhi, India, 2006.
7. Golden, B.; Raghavan, S.; Wasil, E. *The Vehicle Routing Problem: Latest Advances and New Challenges*; Springer Science + Business Media: Berlin/Heidelberg, Germany, 2008; ISBN 978-0-087-77777-1. [CrossRef]
8. Marrekchi, E.; Besbes, W.; Dhouib, D.; Demir, E. A review of recent advances in Operations Research Literature on the green routing problems and its variations. *Ann. Oper. Res.* **2021**, *304*, 529–574. [CrossRef]
9. Anupam, G. Deterministic MST, Advanced Algorithms, CMU. Available online: http://www.cs.cmu.edu/~anupamg/advalgos15/lectures/lecture01.pdf (accessed on 1 April 2024).
10. Kruskal, J.B. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proc. Am. Math. Soc.* **1956**, *7*, 48–50. [CrossRef]
11. Prim, R. Shortest connected network and some generalizations. *Bell Syst. Tech. J.* **1957**, *36*, 1389–1401.
12. Munapo, E.; Kumar, S.; Lesaoana, M.; Nyamugure, P. A minimum spanning tree with index <=2. *Aust. Soc. Oper. Res. Bull.* **2016**, *34*, 1–14.
13. Hoffman, K.L.; Pedberg, M.; Rinaldi, G. *The Travelling Salesman Problem*; Kluer Academic Publishers: Amsterdam, The Nertherlands, 2001. [CrossRef]
14. Korte, B.; Vygen, J. *Combinatorial Optimization: Theory and Applications*; Chapter 21; Springer: Berlin/Heidelberg, Germany, 2006; pp. 501–535.
15. Laporte, G. The travelling salesman problem: An overview of exact and approximate algorithms. *Eur. J. Oper. Res.* **1992**, *59*, 231–247. [CrossRef]
16. Kumar, S.; Munapo, E.; Lesaoana, M.; Nyamugure, P. *Is the Travelling Salesman Problem Actually NP Hard? Chapter 3 in Engineering and Technology: Recent Innovations and Research*; Mathani, A., Ed.; International Research Publication House: Delhi, India, 2017; pp. 37–58, ISBN 978-93-86138-06-4.
17. Kumar, S.; Munapo, E.; Lesaoana, M.; Nyamugure, P. A minimum spanning tree based approach to the travelling salesman problem. *Opsearch* **2018**, *55*, 150–164.

18. Kumar, S.; Munapo, E.; Siguake, C.; Al-Rabeeah, M. The minimum spanning tree with node index <=2 is equivalent to the minimum travelling salesman tour. In *Mathematics in Engineering Sciences: Novel Theories, Technologies and Applications*; Chapter 8; Mathematical Engineering, Manufacturing, and Management Sciences Series; Ram, M., Ed.; CRC Press: Boca Raton, FL, USA, 2020. Available online: https://www.crcpress.com/Mathematical-Engineering-Manufacturing-and-Management-Sciences/book-series/CRCMEMMS (accessed on 1 April 2024).
19. Kalaba, R. On some communication combinatorial network analysis. *Proc. Symp. Appl. Math.* **1960**, *10*. [CrossRef]
20. Saksena, J.P.; Kumar, S. The routing problem with K specified nodes. *Oper. Res.* **1966**, *14*, 909–913. [CrossRef]
21. Kumar, S.; Munapo, E.; Nyamugure, P.; Tawanda, T. Path through K specified links: A heuristic using the minimum spanning tree approach. *Graph. Era Univ. J.* **2023**, *11*, 221–238.
22. Kumar, S.; Munapo, E.; Jones, B.C. An integer equation controlled descending path to a pure integer program. *Indian J. Math.* **2007**, *49*, 211–237.
23. Munapo, E.; Jones, B.C.; Kumar, S. A minimum weight label method and its application to CPM Network. *ORiON S. Afr. J. Oper. Res.* **2008**, *24*, 37–48. [CrossRef]
24. Kumar, S.; Munapo, E.; Ncube, O.; Sigauke, C. A minimum weight labelling method for determination of the shortest path in a non-directed network. *Int. J. Syst. Assur. Eng. Manag.* **2013**, *4*, 13–18.
25. Karimi-Managham, M.; Mohammadi, M.; Meyer, A.M.; Karimi-Managham, E.; Talbi, G. Machine learning at the service of meta-heuristics for solving combinatorial optimization problems: A state-of-the-art. *Eur. J. Oper. Res.* **2022**, *296*, 393–422. [CrossRef]
26. Labidi, H.; Azzouna, N.; Hassine, K.; Goudir, M.S. An improved genetic algorithm for solving multi-objective vehicle routing problem with environmental considerations. *Procedia Comput. Sci.* **2023**, *225*, 3866–3875. [CrossRef]
27. Sabeenian, R.S.; Kumar, V.; Ravi, G.; Pudasaini, A. An Application for Load Sharing in Trucks. *Int. J. Recent Technol. Eng. (IJRTE)* **2019**, *8*, 3203–3207. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

AppliedMath

MDPI

Article

The Mechanics Underpinning Non-Deterministic Computation in Cortical Neural Networks

Elizabeth A. Stoll

Western Institute for Advanced Study, Denver, CO 80202, USA; stoll@westerninstitute.org

Abstract: Cortical neurons integrate upstream signals and random electrical noise to gate signaling outcomes, leading to statistically random patterns of activity. Yet classically, the neuron is modeled as a binary computational unit, encoding Shannon entropy. Here, the neuronal membrane potential is modeled as a function of inherently probabilistic ion behavior. In this new model, each neuron computes the probability of transitioning from an off-state to an on-state, thereby encoding von Neumann entropy. Component pure states are integrated into a physical quantity of information, and the derivative of this high-dimensional probability distribution yields eigenvalues across the multi-scale quantum system. In accordance with the Hellman–Feynman theorem, the resolution of the system state is paired with a spontaneous shift in charge distribution, so this defined system state instantly becomes the past as a new probability distribution emerges. This mechanistic model produces testable predictions regarding the wavelength of free energy released upon information compression and the temporal relationship of these events to physiological outcomes. Overall, this model demonstrates how cortical neurons might achieve non-deterministic signaling outcomes through a computational process of noisy coincidence detection.

Keywords: cortical neuron; neural computation; thermodynamic computation; probabilistic coding; two-state quantum systems; quantum information; von Neumann entropy

check for updates

Citation: Stoll, E.A. The Mechanics Underpinning Non-Deterministic Computation in Cortical Neural Networks. *AppliedMath* **2024**, *4*, 806–827. https://doi.org/10.3390/appliedmath4030043

Academic Editors: Thomas Woolley and Gaige Wang

Received: 25 November 2023
Revised: 25 May 2024
Accepted: 30 May 2024
Published: 26 June 2024

Copyright: © 2024 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Cortical neurons have highly unpredictable signaling outcomes [1], in contrast with spinal neurons, whose signaling outcomes are easily predicted by analyzing upstream inputs [2]. Cortical neurons allow random electrical noise to gate signaling outcomes, with spontaneous subthreshold fluctuations in membrane potential significantly contributing to the likelihood of an action potential [3,4]. Indeed, these cells actively maintain a 'cortical up-state', hovering near the action potential threshold and allowing stochastic events to drive signaling outcomes [5]. This process of noisy coincidence detection results in probabilistic firing patterns, which can be modeled with Bayesian statistics [6], random connectivity models [7], fanofactor analysis of inter-spike variability [8], or by modifying the Hodgkin–Huxley equations to account for electrical noise [9–11]. Notably, the most reliable models of cortical neurons are composed of stochastic differential equations, which provide an analog membrane potential based on ion channel opening probability, and lead to spiking behavior upon threshold activation [12]. This method models 'noise' at the synapse as a Ornstein–Uhlenbeck process [13] or by applying the Fokker–Planck equations [14].

Given the success of these inherently probabilistic computational models, it is worth revisiting whether state transitions in cortical neurons should be modeled under classical assumptions. Indeed, the random movement of just a few ions can push a cortical neuron over action potential threshold [15,16], interspike intervals at the single-unit level are statistically random [3,4], and population coding at the neural network level is statistically random [6–8]. Quantum-scale events materially affect macro-scale outcomes at the level of the computational unit [9–11], and probabilistic models effectively replicate cortical neural network activity [12–14]. Furthermore, classical approaches cannot explain the

extraordinary energy efficiency of the brain, while a model of quantum computation can [17]. To understand the mechanics underlying a computational process where quantum uncertainty is preserved *in situ*, it will be useful to develop a mathematical framework for a multi-scale quantum system.

Classically, a spiking neuron is viewed as a binary unit, always in an on-state or an off-state. Here, the neuron is modeled as a two-state quantum system, with some probability of switching from an off-state to an on-state. This neuronal state is dependent on the inherently probabilistic position and momentum of each ion in the vicinity. In this multi-scale quantum model, the voltage of a cortical neuron is a function of all component pure states, and an optimal system state in the present context is selected from a probability distribution, during a process of quantum information generation and compression. This computational process yields eigenvalues for all state vectors and immediately restores uncertainty across the system, prompting the next computational cycle. This theoretical model offers a mechanism by which cortical neurons might achieve non-deterministic signaling outcomes through a process of ambient-temperature quantum computation and yields specific predictions that can be tested in the laboratory.

2. Methods

2.1. Modeling the Cortical Neuron as a Two-State Quantum System

During up-state, cortical neurons linger at their action potential threshold, allowing both upstream signals and random electrical noise to prompt a signaling outcome. So, while a neuron is classically interpreted as a binary logic gate in an 'on' or 'off' state, coded as 1 or 0, it could also be described as having some probability of converting to an 'on' state or remaining in an 'off' state.

In this new approach, a cortical neuron integrates upstream signals with random electrical noise, defining its voltage state as a function of time, as the system is perturbed. The neuron starts in off-state ϕ, not firing an action potential, and over time t, it reaches another state χ. And so, over some period of time, from t_0 to t, the state of the neuron evolves from ϕ to χ. The timepath taken from one state to another is given by:

$$\langle \chi | U(t, t_0) | \phi \rangle \, . \tag{1}$$

The probability of a state change can be represented in some basis:

$$\sum \langle \chi | k \rangle \langle k | U(t, t_0) | j \rangle \langle j | \phi \rangle, \tag{2}$$

such that U is completely described by base states k and j, which represent the initial off-state of the neuron and the state of the neuron after the time step, respectively:

$$\langle k | U(t, t_0) | j \rangle \, , \tag{3}$$

The time interval can be understood as being $t = t_0 + \Delta t$, so identifying the state of the neuron χ at time t can be understood as taking a path from one state to another:

$$\langle \chi | \psi(t_0 + \Delta t) \rangle \; = \; \langle \chi | U(t_0 + \Delta t, t_0) | \psi(t_0) \rangle \, . \tag{4}$$

If $\Delta t = 0$, there can be no state change. In this case:

$$| \psi(t_0 + \Delta t) \rangle \; = \; | \psi(t_0) \rangle \, . \tag{5}$$

In any other case, the state of the neuron at time t is given by the orthonormal base states k and j, with probability amplitudes:

$$C_k(t_0 + \Delta t) \; = \; \langle k | \psi(t_0) \rangle \tag{6}$$

And:

$$C_j(t_0 + \Delta t) \;=\; \langle j|\psi(t_0)\rangle. \tag{7}$$

The neuronal state $|\psi\rangle$ at time t can therefore be described as a normed state vector ψ, in a superposition of two orthonormal base states k and j, with probability amplitudes C_k and C_j. The sum of the squared moduli of all probability amplitudes is equal to 1:

$$|C_k|^2 \;+\; |C_j|^2 \;=\; 1. \tag{8}$$

Since the neuron starts the time evolution in state $|\psi(t_0)\rangle = j$, its probable state at time t is given by:

$$\langle k|\psi(t_0 + \Delta t)\rangle \;=\; \langle k|U(t_0 + \Delta t, t_0)|\psi(t_0)\rangle\,. \tag{9}$$

This equation can also be written in expanded form as the sum of all transition probabilities:

$$\langle k|\psi(t)\rangle \;=\; \textstyle\sum_j \langle k|U(t_0 + \Delta t, t_0)|j\rangle\langle j|\psi(t_0)\rangle\,. \tag{10}$$

For the state vector $\psi(t)$, the probability of a state change at time t is described by the U-matrix, $U_{kj}(t)$:

$$U_{kj}(t) \;=\; \langle k|U(t_0 + \Delta t, t_0)|j\rangle\,. \tag{11}$$

And so, all probability amplitudes are dependent on the amount of time that has passed, Δt:

$$C_k(t_0 + \Delta t) \;=\; \textstyle\sum_j U_{kj}(t)C_j(t_0)\,. \tag{12}$$

The neuron undergoes perturbations over time Δt. If $\Delta t = 0$, there can be no state change and $k = j$. If $\Delta t > 0$, there is some probability of a state change, where $k \neq j$. As such, the two-state quantum system is described by the Kronecker delta δ_{kj}:

$$\delta_{kj} \;=\; \langle k|j\rangle \;=\; \begin{cases} 1, & \text{if } k = j \\ 0, & \text{if } k \neq j \end{cases} \tag{13}$$

Each of the coefficients of the U-matrix U_{kj} differ from δ_{kj} by some amount proportional to Δt, such that:

$$U_{kj}(t) \;=\; \delta_{kj} + W_{kj}\Delta t\,. \tag{14}$$

where $W_{kj} = -(i/\hbar)\hat{H}_{kj}(t)$, with the Hamiltonian operator term $\hat{H}_{kj}(t)$ representing the time derivatives of each of the coefficients of U_{kj}:

$$U_{kj}(t) \;=\; \delta_{kj} - \frac{i}{\hbar}\hat{H}_{kj}(t)\Delta t\,. \tag{15}$$

The probability amplitude $C_k(t)$ at time t is therefore given by:

$$C_k(t_0 + \Delta t) \;=\; \sum_j [\delta_{kj} - \frac{i}{\hbar}\hat{H}_{kj}(t)\Delta t]C_j(t_0)\,. \tag{16}$$

Since the sum of $\delta_{kj}C_j(t_0) = \psi(t_0)$, the latter equation simplifies to:

$$C_k(t_0 + \Delta t) - C_k(t_0) \;=\; -\frac{i}{\hbar}\sum_j [\hat{H}_{kj}(t)\Delta t]\psi(t_0)\,. \tag{17}$$

In dividing both sides of the equation by Δt, it becomes apparent that any state change is a sum of all possible perturbations that affect the system under investigation, from its starting condition at time t_0, including random events. Again, $C_k(t)$ is the probability

amplitude $\langle i|\psi(t_0)\rangle$ of finding the state vector ψ in one of the base states $k = j$ or $k \neq j$ at time t. And so, the time derivative of this probability function yields the path taken:

$$\frac{C_k(t_0 + \Delta t) - C_k(t_0)}{\Delta t} \;=\; -\frac{i}{\hbar}\sum_j \hat{H}_{kj}(t)\psi(t_0)\,. \tag{18}$$

The system is therefore described by the time-dependent Schrödinger equation, with any state change related to a change in energy distribution:

$$i\hbar\frac{d\psi}{dt} \;=\; \hat{H}\psi\,. \tag{19}$$

Where i is the imaginary unit, $\hbar = h/2\pi$ is the reduced Planck constant, $\hat{H}$ is the Hamiltonian operator, which corresponds to the total energy of the system, and ψ is the eigenvector describing the quantum system after time t. This equation describes the eigenstate of a system as a function of the amount of time t that has passed and the amount of energy available for redistribution across the system, given by the Hamiltonian $\hat{H}$. This equation allows a system state to be defined as a distribution of probabilities representing the possible paths taken since the system state was last defined. By combining Equations (18) and (19), we find the probability of a state change is equal to the sum of all time derivatives of the normed state vector:

$$\frac{C_k(t_0 + \Delta t) - C_k(t_0)}{\Delta t} \;=\; \sum\frac{d\psi}{dt}\,. \tag{20}$$

2.2. Modeling the Cortical Neuron Membrane Potential as a Function of Component Pure States

The dynamic membrane potential of a cortical neuron has been previously modeled with the Fokker-Planck equation [14], with eigenfunctions ϕ_i and coefficients α_i forming a complete set on which the membrane potential distribution $P(V,t)$ can be expanded:

$$P(V,t) \;=\; \sum_i \alpha_i\phi_i(V)e^{-\lambda_i t/\tau_m}\,. \tag{21}$$

Critically, all eigenvalues are real and strictly negative, with eigenfunctions obeying $\mathcal{L}(\phi_i) = -\lambda_i\phi_i$ and $\lambda_i > 0$. A key component of this model is τ_m, the membrane time constant of the neuron, given by:

$$\tau_m \;=\; \frac{g_L}{c_m} \;+\; v_e a_e \;+\; v_i a_i\,, \tag{22}$$

where v_e and v_i are the excitatory and inhibitory synaptic inputs, respectively; a_e and a_i are the excitatory and inhibitory synaptic strengths, respectively; c_m is the membrane capacitance; and g_L is the membrane conductance in the context of a leak current, given by:

$$g_L \;=\; \frac{I_L}{V - V_L}\,. \tag{23}$$

The total current is then classically provided by the sum:

$$I_n \;=\; g_L(t)(V - V_L) + g_e(t)(V - E_e) + g_i(t)(V - E_i)\,, \tag{24}$$

where E_e and E_i are the reversal potentials of excitatory and inhibitory synapses, respectively. In *spinal neurons*, excitatory and inhibitory synaptic inputs dominate any changes to the membrane potential, and the leak current does not contribute materially to signaling outcomes. Meanwhile in *cortical neurons*, excitatory and inhibitory inputs are balanced,

such that leak currents dominate. To model the contribution of individual ions to the leak current, we can model the current density probabilistically:

$$\mathcal{I}_L \;=\; \frac{q}{2m}(\psi^* s\psi - \psi s\psi^*)\,, \tag{25}$$

where m is the mass of each point charge q. In this model, the neuronal state $|\psi_n\rangle$ evolves over time, with the signaling outcome at time t a function of all ion states. Notably, the state of each ion in the system $|\psi_i\rangle$ also evolves with time. The time-dependent Schrödinger equation for the position of each ion is:

$$i\hbar\frac{d}{dt}|\psi(r,t)\rangle \;=\; \hat{H}|\psi(r,t)\rangle\,. \tag{26}$$

And the time-dependent Schrödinger equation for the momentum of each ion is:

$$i\hbar\frac{d}{dt}|\psi(s,t)\rangle \;=\; \hat{H}|\psi(s,t)\rangle\,. \tag{27}$$

The position of each ion is given by a distribution of probability amplitudes along the x, y, and z axes. And so, the momentum of each ion s is provided by the derivative of the wavefunction with respect to its position r:

$$i\hbar\frac{d\psi}{dr} \;=\; i\hbar\left(\vec{x}\frac{\partial}{\partial x} + \vec{y}\frac{\partial}{\partial y} + \vec{z}\frac{\partial}{\partial z}\right) \;=\; i\hbar\nabla\,. \tag{28}$$

Since mass, energy, and electrical charge are conserved, we can employ the continuity equation:

$$\frac{\partial\rho}{\partial t} = -\nabla_r \cdot \mathcal{I}_L\,, \tag{29}$$

where $\rho = |\psi(r,t)|^2$. Therefore:

$$\frac{\partial|\psi(r,t)|^2}{\partial t} \;=\; -\nabla_r \cdot [\frac{q}{2m}(\psi^* s\psi - \psi s\psi^*)]\,. \tag{30}$$

Since $\nabla_r = -s/i\hbar$:

$$\frac{\partial|\psi(r,t)|^2}{\partial t} \;=\; \psi^*\frac{qs^2}{2mi\hbar}\psi - \psi\frac{qs^2}{2mi\hbar}\psi^*\,. \tag{31}$$

The time-dependent Schrödinger equation to describe the Hamiltonian of the system, with respect to momentum, is:

$$i\hbar\frac{\partial\psi}{\partial t} \;=\; (\frac{s^2}{2m})\psi\,. \tag{32}$$

And its complex conjugate is:

$$-i\hbar\frac{\partial\psi^*}{\partial t} \;=\; (\frac{s^2}{2m})\psi^*\,. \tag{33}$$

Therefore:

$$\frac{\partial\rho}{\partial t} \;=\; \frac{\partial|\psi(r,t)|^2}{\partial t} \;=\; \psi^*\frac{q\partial\psi}{\partial t} + \psi\frac{q\partial\psi^*}{\partial t}\,. \tag{34}$$

And so, the sum of all synaptic leak currents for a given neuron at time t is:

$$\mathcal{I}_L \;=\; \int\frac{q\partial\rho/\partial t}{\partial\psi/\partial r}\,. \tag{35}$$

when excitatory and inhibitory inputs are balanced, during a cortical up-state, the current density of a neuron is effectively described by the sum of all leak currents:

$$I_n = \int \frac{q\partial\rho/\partial t}{\partial\psi/\partial r}. \tag{36}$$

Therefore, the probability of the neuron crossing action potential threshold (V_{th}) is given by the relationship between current and voltage, with any negative eigenvalues for $\psi(r,t)$ prompting the neuron to return to its resting potential:

$$V - V_{th} = \frac{1}{g_L}\int \frac{q\partial\rho/\partial t}{\partial\psi/\partial r}, \tag{37}$$

2.3. Modeling the Information Encoded by the Multi-State System, in Terms of von Neumann Entropy

During cortical up-states, neurons actively maintain a resting potential right near the voltage threshold for firing an action potential [5], permitting stochastic ion leak to gate a state change in the computational unit [3,4]. At time t, the neuron either has reached voltage threshold, or it has not; there is some probability of either outcome. The neuron is a two-state quantum system, dependent on the inherently probabilistic position and momentum of each ion in the vicinity. To model the system at either the neuron level or the ion level, ρ_x is defined as the probability of an object transitioning from one to another pure state:

$$\rho_x := |\psi_x\rangle\langle\psi_x|. \tag{38}$$

The density matrix ρ is composed of an ensemble of mutually orthogonal pure states ρ_x, with each having some probability of occurring p_x:

$$\rho = \sum p_x\rho_x = \sum p_x|\psi_x\rangle\langle\psi_x|, \tag{39}$$

Because the cortical neuron is sensitive to quantum-level events, it remains suspended in a state of uncertainty—physically encoding 'information' (the von Neumann entropy of the system, or the mixed sum of all component pure states). The quantity of information encoded by the neuron is calculated by tracing the volume of probability amplitudes across a high-dimensional density matrix:

$$S(\rho) = -Tr(\rho\ln\rho). \tag{40}$$

Here, the system macro-state is the mixed sum of all component pure micro-states, or the mixed sum of all outer products multiplied by their transition probabilities. The inner product $\langle a_y|\psi\rangle$ provides the probability that the state vector $|\psi_x\rangle$ assigns to the eigenvector $\langle a_y|$. As such, the probability of measuring a certain eigenvalue a_y equals:

$$p(a_y) = \langle a_y|\rho|a_y\rangle. \tag{41}$$

The object is in some state at time t_0, and it has some probability of being found in another state at time t, after being transiently defined as a distribution of possible states. The probability of finding a system in any one eigenstate can be calculated by applying the Born rule, which equates the inner product of the state vector and its expectation value to the probability of transition to a particular actual state [18]. This rule states that any measurement of the observable has a probability $p(\psi_a)$ of being equal to an exact value a, at time t for the state vector ψ:

$$p(\psi_a(t)) = |\langle a|\psi(t)\rangle|^2. \tag{42}$$

The square of the absolute value of this wavefunction is a real number, but the wavefunction itself is a complex-valued probability amplitude, which exists along an axis

orthogonal to all real eigenvalues. Every eigenvalue a defined along this axis is dependent on the amount of time that has passed. If sufficient constraints are present, each component pure state can be defined by resolving the Hamiltonian at time t. If sufficient constraints are *not* present, some component pure states will not be resolved, and some net amount of entropy will be produced by the system during that computational cycle.

2.4. Generating a Distribution of Possible System States from Quantum Uncertainty

To understand the mechanics underlying this process, a calculation of possible system states across the neural network can be conceptualized by modeling the Poisson distribution of current density for each ion in the system, in relation to each region of neural membrane, thereby generating a distribution of possible paths. But because these events do not occur with equal probability or independently of the previous system state, it is more appropriate to model the probability distribution of position and momentum vectors as $r(x)$ and $s(x)$, respectively, each defined across three spatial axes, with a weight function $w(x) > 0$:

$$(r, s) \;=\; \int_{-\infty}^{\infty} r'(x)s(x)w(x)dx \;. \tag{43}$$

In this Sturm–Liouville system, the weight function $w(x)$ mathematically represents a quantum harmonic oscillator. $w(x)$ embodies the quantum uncertainty of the ion state and naturally generates a Hilbert space for each position and momentum eigenvector. If this operation is symmetric or Hermitian, such that:

$$\begin{aligned}(r, Ls) \;&=\; \int_{-\infty}^{\infty} L(r'(x))s(x)w(x)dx \\ &=\; \int_{-\infty}^{\infty} r'(x)L(s(x))w(x)dx \\ &=\; (Lr, s)\;,\end{aligned} \tag{44}$$

then all polynomials will form complete sets in Hilbert space, with real eigenvalues and orthonormal eigenfunctions, providing solutions to the second-order linear differentiation equation given by:

$$Lu \;=\; \lambda u\;, \tag{45}$$

In which, λ is a constant and L is a Hermitian operator defined by the real functions of x, α, β, and γ:

$$L \;=\; \alpha(x)\frac{d^2}{dx^2} \;+\; \beta(x)\frac{d}{dx} \;+\; \gamma(x)\;. \tag{46}$$

The distribution of outcomes for $r(x)$ and $s(x)$ is given by ρ, the mixed sum of mutually orthogonal states ρ_x, each occurring with some probability p_x:

$$\rho(r, s) \;=\; \sum (p_{r(x)}\rho_{r(x)})(p_{s(x)}\rho_{s(x)})\;. \tag{47}$$

The evolution of the system state ρ over time t is given by the Liouville–von Neumann equation:

$$i\hbar\frac{d}{dt}\rho \;=\; i\hbar\sum p_x(|\psi_x\rangle\langle\psi_x| + |\psi_x\rangle\langle\psi_x|)\;, \tag{48}$$

with eigenstates provided by the Hamiltonian operator:

$$i\hbar\frac{d}{dt}|\psi_x\rangle \;=\; \hat{H}|\psi_x\rangle\;. \tag{49}$$

and its Hermitian conjugate:

$$-i\hbar\frac{d}{dt}\langle\psi_x| \;=\; \langle\psi_x|\hat{H}\;. \tag{50}$$

Substituting the operator and the conjugate into the Liouville–von Neumann equation yields:

$$i\hbar\frac{d}{dt}\rho \;=\; i\hbar\sum(\hat{H}\rho_x - \rho_x\hat{H}) \;=\; [\hat{H},\rho_x]\,. \tag{51}$$

Because time itself is intrinsically uncertain, the time evolution of the system state is provided by a time-shift operator U:

$$U \;=\; e^{-i\hat{H}t}\,. \tag{52}$$

A unitary transformation, driven by this time-shift operator, allows the density matrix to evolve from $\rho_x(t_0)$ to $\rho_x(t)$:

$$\rho_x(t_0) \;\rightarrow\; \rho_x(t) \;=\; U\rho_x(t_0)U^{\dagger}\,. \tag{53}$$

Undergoing this unitary transformation yields eigenvalues for each component pure state, resolving the position and momentum of each ion, and the voltage state of each neuron. Those eigenvalues are only resolved as the Hamiltonian is resolved, as the dimensionality of the density matrix is reduced, as the wavefunction collapses, as linear correlations are extracted, or as a single system state is selected from a distribution of possible system states. These are all equivalent processes, representing the compression of 'information' or von Neumann entropy. And importantly, this multi-scale system has a single timepoint t, at which both ion states and neuron states are resolved (although some uncertainty or entropy may remain at the completion of each computational cycle).

2.5. Reducing the Probability Distribution into a Single Observable System State

Because the position and momentum of any ion at time t depends on whether it has interacted with other ions or electrical fields during the time evolution, the interdependency of eigenvectors must be taken into account. By calculating the positions and momenta of all ions within a temperature-defined system in relation to each other, this approach ensures that ions do not turn out to occupy the same position, spin, and energy state, which would render them identical. This outcome would cause the universe to lose mass, and is therefore forbidden by the Pauli exclusion principle [19]. This conservation principle can be applied to any thermodynamic system, so the total amount of energy available to the system is no more than the amount of energy stored in the system plus the net amount of energy that has entered the system over time t. In the central nervous system, which traps thermal energy to drive computational work, the distribution of system states is represented by the Hamiltonian operator:

$$i\hbar\frac{d}{dt}|\rho(t)\rangle \;=\; \hat{H}|\rho(t)\rangle\,. \tag{54}$$

In a far-from-equilibrium system that is thermodynamically coupled to its surrounding environment, one cannot assume the Brownian motion of ions occurs at the classical limit, and so one must apply the quantum mechanical formulation of the Fokker–Planck equations, known as the Caldeira–Leggett model [20–22]. Here, the Hamiltonian operator accounts for the position r and momentum s of each ion with mass m and charge q, moving in an potential $V(r)$, where ω is a quantum harmonic oscillator and c is a coupling constant:

$$\hat{H}(r,s) \;=\; \frac{s^2}{2m} + V(r) + \frac{1}{2}\sum\left(\frac{(s_t - cr_t)^2}{m} + m\omega^2 r_t^2\right). \tag{55}$$

Since the state of each ion depends (in part) on the state of all other ions, the whole multi-scale system must be considered together, with ion positions dependent on membrane voltages and vice-versa. This combined wavefunction relates each state vector ψ to the passage of time t and the Hamiltonian operator $\hat{H}$. The spectrum of the Hamiltonian operator is the set of possible outcomes at time t. The Hamiltonian operator $\hat{H}$ is related to

the Lagrangian function of position r, its time derivative $\dot{r}$ and time t. It is calculated by taking the Legendre transform, in order to minimize the action necessary to effect change:

$$\hat{H}(r_i, s_i, t) \;=\; \sum_{i=1}^{n} s_i \dot{r}_i \;-\; L(r_i, \dot{r}_i, t)\,. \tag{56}$$

The vector spaces represented by $\dot{r}$ and $\dot{s}$ are generated by the sum of all perturbances and quantum oscillations, which are mathematically represented by the weight function within the Sturm–Liouville equation. As a result, the derivative of the Hamiltonian operator is related to any changes in position and momentum and time, and therefore can be calculated by taking the partial derivatives of these eigenvectors:

$$d\hat{H} \;=\; \frac{\partial \hat{H}}{\partial r_i}\, dr_i \;+\; \frac{\partial \hat{H}}{\partial s_i}\, ds_i \;+\; \frac{\partial \hat{H}}{\partial t}\, dt\,. \tag{57}$$

The phase space distribution $\rho(r, s)$ describes the probability of a particular system state being selected from the total phase space volume $d^n r\, d^n s$. To describe this phase space volume, the Liouville equation yields the evolution of $\rho(r, s)$ over time t:

$$\frac{\partial \rho}{\partial t} \;+\; \sum_{i=1}^{n} \left(\frac{\partial \rho}{\partial r_i}\, \dot{r}_i \;+\; \frac{\partial \rho}{\partial s_i}\, \dot{s}_i \right) \;=\; \frac{d\rho}{dt} \;=\; 0\,. \tag{58}$$

As a result of this geometrical constraint, the state vectors $(\rho, \rho\dot{r}, \rho\dot{s})$ are conserved across the system, and the vector field $(\dot{r}, \dot{s})$ has zero divergence. This permits temperature and energy to be conserved across the system as well. The continuity equation is given by:

$$\frac{\partial \rho}{\partial t} \;+\; \sum_{i=1}^{n} \left(\frac{\partial (\rho \dot{r}_i)}{\partial r_i} \;+\; \frac{\partial (\rho \dot{s}_i)}{\partial s_i} \right) \;=\; 0\,. \tag{59}$$

Because $\partial\rho \;/\; \partial t$ equals zero, the continuity in the probability density can be termed as:

$$\rho \sum_{i=1}^{n} \left(\frac{\partial \dot{r}_i}{\partial r_i} \;+\; \frac{\partial \dot{s}_i}{\partial s_i} \right) \;=\; 0\,. \tag{60}$$

And because the vector spaces $\dot{r}$ and $\dot{s}$ are defined as:

$$\dot{r}_i \;=\; \frac{\partial \hat{H}}{\partial s_i}\,, \tag{61}$$

and:

$$\dot{s}_i \;=\; \frac{\partial \hat{H}}{\partial r_i}\,, \tag{62}$$

the continuity laws persist while taking the second derivative of the Hamiltonian operator. This process identifies the observable boundary of the total volume of possible system states:

$$\rho \sum_{i=1}^{n} \left(\frac{\partial^2 \hat{H}}{\partial r_i\, \partial s_i} \;-\; \frac{\partial^2 \hat{H}}{\partial s_i\, \partial r_i} \right) \;=\; 0\,. \tag{63}$$

The relationship between the density matrix ρ and the Hamiltonian operator describing the distribution of energy across the system becomes apparent in this equation, as does the relationship between position and momentum vectors and the Hamiltonian operator. Once a trace is taken across the density matrices representing $r(x)$ and $s(x)$, the weight function $w(x)$ underpinning the probability distribution can be solved in relation to the other parameters. And so, during a system-wide computation, eigenvalues are selected and eigenstates $\psi r(x)$ and $\psi s(x)$ are transiently resolved. It should be noted that, in this formulation of temporally proceeding events, the operators drive the computational cycle,

rather than the vector states. That is, the Hamiltonian operator may evolve over time (permitting changes in position and momentum) while the vector states themselves remain time-independent. Any events occurring within a time evolution may therefore contribute to the outcome of a computation. This includes any quantum oscillations and any classical input currents occurring during that time evolution.

2.6. Restoring Uncertainty after the System State Is Transiently Defined

By relating any probabilistic changes in the position and momentum of each ion to the distribution of energy across the system, the Hamiltonian operator effectively portrays all observable outcomes. The unitary change of basis guiding this computational process is defined by the phase factor $\exp(i\hat{H}t/\hbar)$. Therefore, any reversal to the direction of time causes positive energies to become negative—a result that is disallowed by the first law of thermodynamics, unless the spin is also reversed with the introduction of an electric dipole moment [23]. An electric dipole moment is a spontaneous shift in the energy state of an electron in the presence of an external electric field; this event is expected to occur here, as the probability distribution is reduced and eigenvalues are calculated.

Initially, as the system was perturbed by its external environment, all component pure states (and the Hamiltonian operator) were integrated to create a volume of probability. Now, we can take the derivative of that volume of probability (and the Hamiltonian operator) to define eigenvalues on the boundary region of that high-dimensional probability distribution. Doing so reduces the distribution of system states into a single actualized system state, with a non-deterministic outcome:

$$\frac{dE}{d\lambda} = \langle \psi_\lambda | \frac{d\hat{H}}{d\lambda} | \psi_\lambda \rangle . \tag{64}$$

This equation is the Hellman–Feynman theorem, and it lies at the core of quantum electrodynamics [24,25]. Here, any change in the position or momentum of an ion is proportional to the change in the Hamiltonian, because the Hamiltonian corresponds to all potential and kinetic energies in the system. It is useful to note that the total amount of energy in the system is not what is uncertain, but rather how this energy is distributed. The measure of possible system states, or energetic configurations, is directly related to the total entropy encoded by the system at the beginning of the computational cycle [26]. The system state, upon completion of the computational cycle, is related to any parameter λ that contributes to the Hamiltonian operator, such as any shift in atomic position, momentum, orbital, or electrical field strength. And so, the derivative of the total energy in the system E is related to the inner product of the state vectors ψ and the derivative of the Hamiltonian $\hat{H}$, both with respect to the parameter λ.

In developing this solution to the Schrödinger equation, Richard Feynman discovered that any wavefunction collapse has a discrete effect on the component atoms. This event is accompanied by a shift in the charge distribution across each atom in the system, in relation to their newly- defined distance from each other [24,25]. For two atoms interacting at a separation D, the induced dipole moment for each atom is $1/D^7$. The smaller the distance between atoms, the larger the dipole moment, and the larger the boost to angular momentum. As a result, any perturbation to the system—for example, an ion channel opening or a shift in the local electrical field—changes the state of the system, in a manner related to the original unperturbed state and the derivative of the Hamiltonian. The assignment of atomic locations, relative to other atoms in the system, is paired with an alteration in the organization of electrons around the nucleus. This event prompts van der Waals forces between neighboring atoms [27,28].

In summary: By integrating complex-valued probability amplitudes, the system generates a quantity of quantum information. By taking the derivative of this volume, the system identifies eigenvalues on the observable boundary region of that high-dimensional vector space. As eigenvalues are selected, all other eigenstates are eliminated. At that point, the wavefunction collapses and the information held by the system is abruptly compressed.

In accordance with the Landauer principle, any compression of information is paired with a release of free energy [29–32]. This free energy can be used to perform work in the system, e.g., restoring the resting membrane potential as uncertainty within the neuron's receptive field is reduced. Critically, this bidirectional exchange of free energy and entropy ensures that the total energy of the system is conserved over the course of the computation, while respecting the first and second laws of thermodynamics [26].

By moving from a prior state to a posterior state, the system actualizes a solution to a computational problem: What is the optimal system state to encode the surrounding environment? However, this defined system state is transient, because it is paired with an immediate restoration of uncertainty, through an alteration in the charge distribution around each atom. The dipole moment induced by wavefunction collapse then prompts new atomic interactions—and for successive measurements with discrete results, which do not destroy the entanglement of the particle system, each measurement with value a establishes the basis for a new state, which then undergoes subsequent time evolution, in accordance with the von Neumann projection postulate [33]. And so, immediately after a wavefunction is resolved, the system again begins to evolve over time, forming a new probabilistic system state, and undergoing another computational cycle.

2.7. Converting Probabilistic System States to Temporally Irreversible Signaling Outcomes

Taking the derivative of the Hamiltonian involves taking the derivative of all component pure states, with respect to all perturbations since the last detection event. This process compresses information, yielding eigenvalues on the boundary region of the high-dimensional probability distribution. During this computational process, the complex-valued distribution of possible system states is reduced, and component pure states throughout the system (e.g., ion position and momentum) are transiently defined. But this defined multi-scale system state immediately becomes the past, as the charge distribution across each atom is altered and a new probabilistic system state emerges.

Since the expectation value for the energy of a given atom $\langle a_i|\rho|a_i\rangle$ is proportional to the expectation value for its spin $\langle\psi|k|\psi\rangle$, the energy shift due to the induced dipole moment causes a sign shift in both values, and time symmetry is broken. Essentially, as ground states lose degeneracy, the resulting dipole moments should alter the attraction between sodium ions outside the neuron and atoms comprising the lipid bilayer of the neural membrane. The resulting van der Waals forces are expected to cause a permitted violation of time-reversal symmetry, thereby effecting causation within the system as information is compressed, eigenvalues are identified, and leak currents are observed.

Yet a system will only demonstrate violations of time-reversal symmetry if quantum uncertainty contributes to thermal fluctuation–dissipation dynamics. Only if coincident upstream signaling events and random oscillations trigger a change in membrane resistance, within the temporal parameters of ion dissipation and ion pump rectification kinetics, will inherently probabilistic events contribute to gating a neuronal state change.

2.8. Conditions under Which Quantum Fluctuations Contribute to Dissipation Dynamics

A physical structure that actively generates an electrochemical resting potential will generate entropy and heat. The neural membrane provides resistance, and therefore some quantity of free energy will be dissipated into entropy as electrical interactions occur. However, in a heat-trapping system, this energy is not necessarily irreversibly lost; it can be used to do work [34]. The Callen–Welton fluctuation–dissipation theorem asserts that thermal fluctuations drive a response function affecting the impedance of the structure, given by $\chi(t)$:

$$\chi(t) = -\beta \frac{A(t)\theta(t)}{dt}. \tag{65}$$

where $\beta = k_B T$, $\theta(t)$ is the Heaviside step function, and $A(t)$ is the expectation value of $O_k(t)$, an observable that is subject to thermal fluctuations in a dynamical system. If decoherence timescales within the system are longer than the timescales of ion dissipation

and ionization dynamics, quantum uncertainty can materially contribute to ion behavior at the neuronal membrane [35]. That is, if a time-dependent change in the voltage of a *neuron* relies on any quantum fluctuations in the position or momentum of nearby *ions*, then the ensemble average of each observable (a measure of fluctuation, given by the Hermitian operator $[O_k(t), O_j(t_0)]$) will be related to the response function (a measure of dissipation, given by $\chi(t - t_0)$, as a function of time. This relationship is given by the Kubo fluctuation–dissipation formula [36,37]:

$$\chi(t - t_0) \; = \; i\theta(t - t_0) \, \langle \, [\, O_k(t), O_j(t_0) \,] \, \rangle \, . \tag{66}$$

This response function, $\chi(t)$, can be written as a function of oscillatory events:

$$\chi(t) \; = \; \int_{-\infty}^{\infty} \frac{d\omega}{2\pi} \, e^{i\omega t} \chi(\omega) \, . \tag{67}$$

If $t < 0$, then $i\omega t$ will be negative and $e^{i\omega t}$ will be zero, so the entire response function $\chi(t)$ will be zero. As such, the quantum states underpinning this spectral function can only causally contribute to dissipation dynamics as time moves in a forward direction. The Fourier transform of the response function provides for dissipation and fluctuation dynamics in the frequency domain, given by $\chi(\omega)$:

$$\chi(\omega) \; = \; -i \int_0^{\infty} Tr \, (\, e^{\beta \hat{H}} [\, O_k(t), O_j(t_0)] \,) \, e^{i\omega t} dt \, . \tag{68}$$

This time-dependent function provides the density of possible states for a particle system, which emerge in the presence of a perturbation or changing electrical field. And so $[O_k(t), O_j(t_0)]$ is simply a description of how the density matrix, the wavefunction, or the Hamiltonian operator changes over some time evolution, compared with the probability that all energy states would rearrange in that same way under time reversal. As such, the response function $\chi(\omega)$ can be written in expanded form:

$$\begin{aligned} \chi(\omega) \; = \; -i \int_0^{\infty} e^{i\omega t} dt \sum_{mn} e^{-E_m \beta} [\, \langle m|O_k|n\rangle \langle n|O_j|m\rangle e^{i(E_m - E_n)t} \\ - \, \langle m|O_j|n\rangle \langle n|O_k|m\rangle e^{i(E_n - E_m)t}] \; . \end{aligned} \tag{69}$$

If the uncertainty in the position and momentum of an ion is sustained in the presence of a constantly changing electrical field, then quantum fluctuations may contribute to ion behavior, affecting the state of ions interacting with the electrochemical potential of the neural membrane. In this case, quantum fluctuations may contribute to the probability of a state change in the macro-scale computational unit, from an off-state to an on-state. However, the state of each neuron at the moment the Hamiltonian operator is resolved will govern its response. Neurons in a cortical up-state, which allow stochastic events to gate a signaling outcome, may be nudged toward action potential threshold or away from it as uncertainty is abruptly reduced. Indeed, *only* neurons in cortical up-state, allowing random noise to gate a signaling outcome, will be nudged. Meanwhile, neurons receiving suprathreshold stimulation from a resting state will exhibit deterministic firing patterns.

2.9. Assumptions of the Model

2.9.1. Neurons Are Functionally Isolated but Remain Sensitive to External Perturbations

This model assumes that neurons are functionally isolated, but remain sensitive to external perturbations. The architecture of a deep layered neural network permits perturbations from outside the system itself, while the physiology of individual computational units within the deeper layers permits isolated computations to occur.

With regard to the network architecture: Neurons in the thalamus, which receive signals from the periphery, exhibit deterministic firing patterns that faithfully represent mappings within their receptive fields [38,39]. These neurons synapse onto deeper layers

of the neural network, with signals cascading onto primary sensory neurons and higher-order cortical neurons, which increasingly integrate multiple inputs with noise to produce probabilistic outcomes [40,41]. Neurons in the motor cortex and basal ganglia integrate various inputs to initiate voluntary movement and inhibit other movement [42,43]. These neurons synapse onto downstream motor regions in the thalamus, reticular formation, and the spinal cord, with these signals converging onto peripheral neurons with deterministic outcomes that instruct the extension and flexion of specific muscle groups [44,45]. Essentially, higher-order neurons within the deepest layers of cortex receive messages from within the system itself, with these perturbations representing the outside world.

With regard to the neuronal physiology: In cortical neurons, excitatory and inhibitory inputs are actively balanced, so that thermal fluctuations dominate any exit from equilibrium potential [5]. While a suprathreshold stimulus can certainly lead to spiking behavior in neurons of primary sensory cortex [46,47], random electrical noise drives spiking behavior in higher-order cortical neurons [3,4]. In these areas, signaling outcomes are more probabilistic [48,49]. For this reason, modeling the contribution of noise to signaling outcomes is a valid approach. Notably, in peripheral neurons and superficial layers of the cortical circuitry, random noise is insufficient to affect signaling outcomes, and the computational units themselves are highly robust to noise, so thermal fluctuations do not materially contribute to neuronal outcomes, and these terms drop from the equation. However, a neuron in cortical up-state, hovering near action potential threshold and allowing random electrical noise to gate signaling outcomes, is highly sensitive to noise. In this case, the noise must be mechanistically taken into account to effectively model the probability of the neuron transitioning from off-state to on-state.

2.9.2. Uncertainty in the State of Individual Ions Affects the Voltage State of the Neuron

This model assumes that a shift in the state of very few ions can affect the likelihood of a neuron firing an action potential. When neurons are in a cortical up-state, a 1 mV increase in membrane potential can easily push a neuron over threshold [15]. This change in voltage corresponds to 1 pA of current, or only 10 sodium ions moving through a channel over the course of a microsecond [16]. Since the random motion of just a few sodium ions can push a cortical neuron over action potential threshold, and these cells have significant leak currents [50], quantum-scale events appear to affect macro-scale outcomes at the level of the computational unit.

Neurons in cortical up-state are expected to encode maximal amounts of information, with high probability amplitudes for each outcome $k = j$ and $k \neq j$. Yet for ions, the probability amplitudes across both position and momentum space will be zero almost everywhere. For ions located outside the brain, the probability that an ion will be located within a cortical neuron in the next instant is negligible. And for ions already located within the brain, the probability amplitudes are largely centered around the most recent position and momentum values, except in the context of nearby ion channel opening or a shift in the local electrical field. An unlikely outcome for an ion might be a positional change corresponding to leak across a neural membrane, with this event affecting the voltage state of the neuron.

Here, each ion is described by a wavefunction, or a distribution of eigenstates. During the unitary transformation, any non-distinguishable states (e.g., linear correlations between pure states) are eliminated, thereby compressing the von Neumann entropy of the system. As eigenvalues for each ion are assigned, the voltage of the neuronal membrane is defined, and the cell either reaches action potential threshold or not. If uncertainty is reduced during the computation, and eigenvalues are assigned, then information is compressed, free energy is released, and the neuron restores its resting potential. If information cannot be compressed, and no eigenvalues can be assigned, then free energy is not released, and the neuron fires an action potential, thereby encoding 'uncertainty' in its receptive field.

2.9.3. The Estimated Decoherence Timescales Meet the Criteria for a Quantum System

The model for decoherence in neural systems, focusing on interactions between sodium ions and the neuronal membrane potential, was originally provided by Max Tegmark [35]. This model was recalculated using the newest coulomb scattering data, which are presented in a sister paper, along with the matrix algebra formulation of quantum information generation and compression [51]. These estimates demonstrate that decoherence timescales at the neuronal membrane are indeed longer than ion dissipation rates, and therefore meet the criteria for a quantum system—but only for cortical neurons in up-state.

A cortical neuron in up-state is effectively isolated, because excitatory and inhibitory inputs are balanced, so random thermal fluctuations drive signaling outcomes. While there are upstream inputs, these upstream inputs do not drive signaling outcomes in cortical neurons—thermal fluctuations do. For this reason, each computational unit is modeled as a two-state quantum system, with some probability of transitioning from off-state to on-state. Of course, the neuron is not truly isolated; it exists within a circuit, and so a suprathreshold stimulus can still kick it over action potential threshold and cause it to fire. The concept of modeling the contribution of quantum-level noise to the state of the computational unit is always valid; however, in many circuits, the decoherence timescales are too short, or the random noise is insufficient to affect signaling outcomes, or the computational units themselves are robust to noise, so thermal fluctuations do not materially contribute to neuronal outcomes, and these terms naturally fall out as neurons obey classical dynamics.

Only in a cortical up-state is the uncertainty in both neuronal membrane potential and extracellular ion position expected to be sustained for sufficiently long timescales for linear correlations to be reduced—for the wavefunction to collapse, for information to be compressed, for free energy to be released, for atomic momentum to be altered—and for voltage changes in the computational unit to be observed, at the completion of the computation.

3. Results

3.1. The Expected Wavelength of Spontaneous Free Energy Release during Information Compression

If cortical neural networks engage in ambient-temperature quantum computation, then these far-from-equilibrium thermodynamic systems must bidirectionally exchange free energy for information, with any free energy expended on information generation during the initial stage of the thermodynamic computing cycle being partially recovered during the information compression stage. As such, discrete quantities of free energy should be released, local to any reduction in uncertainty, as an optimal system state is selected from some probability distribution. Any thermal fluctuation should correspond to a shift in the atomic dipole moment, boosting the angular momentum of individual ions. These events should therefore be observable. To evaluate this hypothesis, we can calculate the expected effects of quantum fluctuation–dissipation dynamics in the mammalian central nervous system.

If the perturbation to the state of an ion during some time evolution relies on any random changes to the spin or energy state of a component electron, then the ensemble average of each observable (a measure of fluctuation, given by $\langle \hat{x}(t), \hat{x}(0) \rangle$) will be related to the response function (a measure of dissipation, given by $\chi(t)$) in the frequency domain [36]. This relationship essentially models how the oscillatory behavior of a quantum system, along an imaginary axis, affects the thermal dynamics of the observable system:

$$\chi(t) = i\theta(t - t_0) \langle \hat{x}(t), \hat{x}(0) \rangle, \tag{70}$$

where $\theta(t)$ is the Heaviside step function and the response $\chi(t)$ provides an expectation value for $x(t)$, which is a time-dependent 'observable' subject to thermal fluctuation in a dynamical system. The time-dependent equation is given by:

$$f(t) = \int \frac{d\omega}{2\pi} e^{i\omega t} f(\omega), \tag{71}$$

and its Fourier transform is given by:

$$f(\omega) = \int f(t)\, e^{-i\omega t} dt\,. \tag{72}$$

If the eigenvalues are to be real, the sum of the real part and the imaginary part of the response function over some time evolution $\chi(t - t_0)$ must also be real. The full response function is given by:

$$\chi(\omega) = \mathrm{Re}\chi(\omega) \;+\; \mathrm{Im}\chi(\omega)\,. \tag{73}$$

The real part of the response function is given by:

$$\begin{aligned} \mathrm{Re}\chi(\omega) &= \frac{1}{2}\left[\langle \chi(\omega) - \chi^{\dagger}(\omega)\rangle\right] \\ &= \frac{1}{2}\int_{-\infty}^{\infty} dt\, e^{i\omega t}\left[\langle \chi(t) - \chi(-t)\rangle\right]. \end{aligned} \tag{74}$$

And the imaginary part of the response function is given by:

$$\begin{aligned} \mathrm{Im}\chi(\omega) &= -\frac{i}{2}\left[\langle \chi(\omega) - \chi^{\dagger}(\omega)\rangle\right] \\ &= -\frac{i}{2}\int_{-\infty}^{\infty} dt\, e^{i\omega t}\left[\langle \chi(t) - \chi(-t)\rangle\right]. \end{aligned} \tag{75}$$

The energy of quantum fluctuations $E(\omega)$ is related to the frequency ω:

$$E(\omega) = \frac{1}{2}\,\hbar\omega\, coth\,\frac{1}{2}\,\beta\hbar\omega, \tag{76}$$

where $\beta = 1/k_B T$. The classical power spectrum is related to the complex-valued quantum spectral density in such a way that quantum noise can contribute to the local thermal density under certain conditions:

$$\int_{-\infty}^{\infty} \langle \hat{x}(t) - \hat{x}(t_0)\rangle\, e^{i\omega t}\, dt \;=\; \frac{1}{\beta E(\omega)} \int_{-\infty}^{\infty} \langle \hat{x}(t) - \hat{x}(t_0)\rangle\, e^{i\omega t}\, dt. \tag{77}$$

If $E > k_B T$, with a high temperature, a broad distribution of electrons across energy states, and low occupancy of energy states, then the quantum contribution to ion behavior is negligible, and the behavior of ions will reduce to Boltzmann–Maxwell statistics. If $E < k_B T$, with electrons at low frequencies obeying the Rayleigh–Jeans law, then the quantum contribution to ion behavior is also negligible, and again the behavior of ions will reduce to classical Boltzmann–Maxwell statistics. Only in cases of high particle density, when the energy held by an ion is greater than its chemical potential, and $E > k_B T$, will quantum fluctuations contribute to ion behavior. This is predicted to be the case in the mammalian brain.

If quantum fluctuations do indeed contribute to ion behavior in biological systems, then E must be greater than $k_B T$. Since:

$$k_B T = (1.38 \times 10^{-23}\,\mathrm{J/K})(310\,\mathrm{K}) \;=\; 4.28 \times 10^{-21}\,\mathrm{J}, \tag{78}$$

and:

$$f = \frac{E}{h} = \frac{4.28 \times 10^{-21}\ \mathrm{kg\,m^2/s^2}}{6.63 \times 10^{-34}\ \mathrm{kg\,m^2/s}} = 6.46 \times 10^{12}\mathrm{Hz}, \quad (79)$$

it is expected that high-energy particles of $E > k_B T$ will be observed in the central nervous system at 37◦ C (310K). Specifically, these high-energy particles should have a frequency of $f >$ 6.46 x 10^{12} Hz, a wavelength of $\lambda <$ 46 microns, or an energy of $E >$ 0.0267 eV, within the infrared light spectrum. Spontaneous emissions of photons in this range have indeed been observed in mammalian brain tissue [52–55] and infrared stimulation of the brain has been shown to have a functional effect on neural activity [56–58]. Further studies are needed to measure the exact wavelengths of these photon emissions and temporally correlate these events with neuronal signaling outcomes.

In summary, it is predicted that photons, specifically in the infrared range of the electromagnetic spectrum, should be released upon information compression in the mammalian brain. Since the first law of thermodynamics states that energy cannot be created nor destroyed, any system capable of reducing entropy to achieve a non-deterministic computation must release free energy upon information compression. If quantum computing does occur in cortical neural networks, then spontaneous thermal fluctuations should be observed, locally to any reduction in uncertainty. These thermal fluctuations are expected to drive synchronous firing of a statistically random ensemble of neurons across the network.

Therefore, in this approach, the reduction in thermodynamic entropy is paired with both the selection of an optimal system state from a large probability distribution (one that is thermodynamically favored to correlate with the surrounding environment) and the release of thermal free energy (which is used to physically instantiate the solution to that computational problem).

3.2. Specific Predictions of This Model

If cortical neural networks are indeed quantum computing systems rather than classical computing systems, then evidence of quantum information generation and compression should be observed in the neocortex. As such, this theory makes specific predictions with regard to the energy efficiency of the brain [17]; with regard to coulomb scattering and decoherence timescales at the neuronal membrane [51]; and with regard to the expected effects of electromagnetic stimulation and pharmacological intervention in cortical neural networks [59]. Some additional specific predictions of the theoretical framework, prompted by the present model, include the following.

3.2.1. Thermal Free Energy Is Spontaneously Released during Computation as Information Is Compressed

Infrared particles with wavelengths of $\lambda <$ 46 microns or $f > 6.46 \times 10^{12}$ Hz should spontaneously appear at the neural membrane during cortical information processing. This prediction must be tested with sensitive infrared detection devices rather than classical electrodes or imaging systems; the spontaneous release of infrared-wavelength particles should be observed in the brain as uncertainty is resolved into signaling outcomes. A quantitative increase in these particles should be observed, for example, upon perceptual recognition of a highly uncertain visual or auditory stimulus, with a strong temporal correspondence to P300 event-related potentials in the cerebral cortex. By contrast, this spontaneous thermal free energy release should not occur in the case of an epileptic seizure—when constitutive ion channel activation, rather than information processing, leads to highly synchronized neural activity across the cerebral cortex. Of course, spontaneous emissions of photons in this range have been observed in mammalian brain tissue [52–55], and infrared stimulation of the brain has been shown to have a functional effect on neural activity [56–58], but further studies are needed to measure the exact wavelengths of these photon emissions and evaluate whether these events are temporally correlated with neuronal signaling outcomes. If the brain does cyclically generate and compress entropy, the system should demonstrate much higher energy efficiency than expected under classical conditions (Table 1).

Table 1. A comparison of explanatory power: the energy efficiency of the neural network.

Classical models		
Proposed Mechanism	Predicted Observation	Evidence For/Against
The energy efficiency of the system is the result of optimal synaptic weighting, optimal ion channel distribution, and other molecular mechanisms.	Net production of physical entropy in the human brain is compatible with classical assumptions, with ATP turnover producing some amount of entropy.	The computational cost for each spike is an astounding 0.1 W: In the context of known caloric intake, this energy requirement is "off by a factor of 10^8." [60]
Present model		
Proposed Mechanism	Predicted Observation	Evidence For/Against
The energy efficiency of the system is the result of information compression, with neural outcomes prompted by the extraction of correlations, consistencies, or 'predictive value'.	Net production of physical entropy in the human brain is far too low to retain the assumptions of a classical system, with the amount of work done per calorie showing near-perfect use.	"The energy efficiency of the human brain is consistent with this model of non-deterministic computation." [17] "This computational process maximizes free energy availability." [26]

3.2.2. The Spontaneous Release of Thermal Free Energy during Information Compression Prompts Synchronized Firing across the Neural Network

This model describes both a computational process and a thermodynamic process, since information compression is both the selection of an optimal system state from a large probability distribution and the reduction of entropy. This system-wide thermocomputational event, resolving the uncertainty in component pure states, is predicted to lead to a spontaneous release of free energy. Neurons that reduce uncertainty during the computation should therefore recover free energy and restore their resting potential; neurons that retain uncertainty in their receptive field should distribute free energy to entropy and fire a signal. This system-wide computation should therefore result in the synchronous firing of a statistically random ensemble of neurons across the network. Synchronous neural activity is indeed observed at a range of frequencies in cortical neural networks, and is considered a correlate of higher-order cognitive processes [61,62]. Critically, the high frequency oscillations which are observed during perceptual tasks cannot be modeled by coupling and recruitment under classical assumptions and timescales [63,64]. Here, information compression events are predicted to prompt spontaneous synchronized activity across sparsely distributed neurons. This coordinated activity is predicted to occur in cells that allow random noise to gate signaling outcomes (e.g., cortical neural circuits) but not in cells that act entirely deterministically (e.g., spinal reflex circuits). While this oscillatory activity, occurring at a range of frequencies, has been observed in the mammalian brain, additional studies could explore the potential correlation between probabilistic coding and network-level activity in avian and cephalopod species. If synchronous activity is caused by classical methods of signal propagation, rather than being the result of a system-wide non-deterministic computation, then both classical simulations of cortical neural networks and spinal reflex circuits should readily demonstrate fast and slow oscillations. If instead, synchronous activity across the network is caused by information compression paired with free energy release, then synchronous firing should be eliminated by absorption of the predicted wavelengths and should be prompted by introduction of these wavelengths. In short, if a classical model is correct, then classical mechanisms should readily generate oscillations at a range of nested frequencies, and if the present model is correct, then cyclical information generation and compression should readily generate these oscillations (Table 2).

Table 2. A comparison of explanatory power: the synchronous firing of sparse neuronal ensembles.

Classical models		
Proposed Mechanism	Predicted Observation	Evidence For/Against
The observed synchronous firing at a range of nested frequencies is the result of information encoding, with neural signaling outcomes prompted by a common stimulus.	A combination of gap junctions, chemical synapses, ephaptic coupling, changes in ion concentration, and optimization of neural connectivity over time prompts synchronous firing.	These events are not readily simulated: "It is difficult, however, to identify the exact contribution of each mechanism to a specific type of oscillation." [65] The problem is "non-trivial." [66]
Present model		
Proposed Mechanism	Predicted Observation	Evidence For/Against
The observed synchronous firing at a range of nested frequencies is the result of information compression, with neural signaling outcomes prompted by the extraction of correlations.	An infrared photon pulse drops the membrane potential of some neurons, while other neurons in cortical up-state fire, resulting in synchronous firing but not ictal activity across the network.	The predictions of this model should be tested: specifically, spontaneous infrared photon release is expected to be temporally correlated with neural oscillations, but not with ictal activity.

4. Discussion

The neuron is classically viewed as a transistor, always in either an on-state or an off-state. Here, the cortical neuron is modeled as a qubit, with some probability of transitioning from an off-state to an on-state over some time evolution. In this new approach, a state change in the computational unit relies on inherently probabilistic events at the ion level.

Noise plays a critical role in cortical neural networks and other complex dynamical systems, facilitating information transmission through a mechanism of stochastic resonance [67,68]. Random electrical noise prompts phase transitions in individual computational units, leading to the highly variable interspike intervals that are observed in cortical neurons [69]. Random electrical noise also leads to synchronized activity and large-scale oscillations at the systems level [70]. Noise-induced resonant activation and noise-enhanced stability is observed in both neural networks [71] and memristic architectures [72].

The result of a noisy computational process, at the network level, is 'sparse populational coding'—the synchronous firing of a statistically random ensemble of sparsely distributed neurons across the network [73,74]. This neuronal activity corresponds with the realization of a multi-sensory percept [62,63,75,76]. Notably, sparse populational coding and percept realization are observed in neural networks that retain sensitivity to random electrical noise (such as the mammalian neocortex), but not in neural networks that are robust to random electrical noise (such as spinal reflex circuits).

Many previous models have used classical methods to describe the interspike interval or the synchronous firing of statistically random ensembles of neurons at the network level [9–14,67–72], yet none of these methods provide mechanistic insight into how individual cortical neurons gate signaling outcomes by allowing thermoelectric noise to affect neuronal signaling outcomes. The present approach demonstrates how noisy coding at the synapse is mechanistically linked to probabilistic signaling outcomes. This method makes specific predictions that can be tested in the laboratory, to evaluate whether the brain is a classical system (as typically assumed) or a multi-scale quantum system (as presented here).

Hameroff and Penrose have also proposed the brain to be a quantum system, with orchestrated objective reduction of the system state corresponding to conscious experience [77]. Indeed, the collapse of alternative eigenstates is a key property of cortical neural networks in both OrchOR theory and the present model. However, OrchOR theory focuses on the role of quantum decoherence on microtubule dynamics, while the present report models the contribution of random electronic noise to the cortical neuron membrane potential. The former approach does not connect with established mechanisms of neuronal information processing, while the latter approach is highly compatible with the known contributions of Brownian motion to cortical neuron signaling outcomes.

While the idea of the brain as a quantum system has often been considered highly controversial, researchers in computational neuroscience have recently adopted the concept to drive significant advances in neural network modeling and machine learning [78–80], building on previous efforts to incorporate noise into the Hodgkin and Huxley model [9–11,81]. However, these approaches, when implemented in classical computing architecture, do not permit physical information compression and free energy release. A network of computational units with probabilistic gating behavior and similar decoherence timescales to the mammalian brain would have to be built to achieve this computational process in engineered neural network architecture.

In this model, inherently probabilistic component states are integrated to populate a Hamiltonian operator. The Hamiltonian operator is then differentiated with respect to all perturbations to the system. The redistribution of energy that results from this computational process assigns eigenvalues for the position and momentum of each ion in the system at a single point in time. This process defines the voltage state of each neuron, as quantum-level events affect outcomes at the level of the computational unit. This study demonstrates how neurons in a cortical up-state retain a state of uncertainty, physically generating and compressing information to achieve non-deterministic signaling outcomes. This model is explicitly a form of ambient-temperature quantum computation. There are three additional ways to describe this computational process, provided in sister reports:

In accordance with the laws of thermodynamics, free energy must be expended to create information and this free energy is partially recovered upon information compression. In this model, probabilistic component pure states can be represented algebraically by a density matrix [51]. The density matrix undergoes a unitary change of basis, as the system state is perturbed by its surrounding environment over some time evolution. The diagonalization of the density matrix yields a zero determinant, leading to observables on the boundary region of that high-dimensional probability distribution.

In accordance with the laws of holography, these probabilistic component pure states can also be represented geometrically as complex-valued waves or wavefunctions [59]. These complex-valued probability amplitudes constructively and destructively interfere on the sensitive charge-detecting polymer surface of the neural membrane. As a result of this physical interference between probability amplitudes, the wavefunction collapses and eigenvalues are actualized on the boundary region of the high-dimensional probability distribution. This process of physically encoding information on the polymer neural membrane surface generates a holographic projection of the encoded content.

In accordance with the laws of thermodynamics, entropy must always increase and the total energy of the system must be conserved. Under classical assumptions, most free energy in the cerebral cortex should be dissipated to entropy, as an optimal system state is selected from a large probability distribution and all other possible system states are lost. Yet the observed energy efficiency of the brain suggests this is not the case. By modeling the system as cyclically generating and compressing quantum information, the extraordinary energy efficiency of the brain can be explained without violating the first or second law of thermodynamics [17]. This novel approach also forges a deep connection between predictive processing and the thermodynamic limits of quantum computation [26]. Achieving sparse populational coding, through a process of inherently probabilistic computation, may lead to more generally intelligent and energy-efficient neural networks.

In summary, the present model of Hamiltonian mechanics is complemented by these models of matrix mechanics, wave mechanics, and wetware-instantiated thermodynamic computation, which demonstrate the same process of information generation and compression. Indeed, cortical neurons may be better described as qubits, encoding von Neumann entropy, rather than classical bits, encoding Shannon entropy. This theoretical framework for ambient-temperature quantum computation may not only provide useful insight into the operation of biological systems, but also drive advances in engineered intelligence.

Funding: The author received support for this work from the Western Institute for Advanced Study, with generous donations from Jason Palmer, Bala Parthasarathy, and Dave Parker.

Data Availability Statement: All methods, materials, and data needed to replicate this study are included in the manuscript.

Conflicts of Interest: The author declares no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Softky, W.R.; Koch, C. The highly irregular firing of cortical cells is inconsistent with temporal integration of random EPSPs. *J. Neurosci.* **1993**, *13*, 334–350. [CrossRef] [PubMed]
2. Powers, R.K.; Binder, M.D. Effective synaptic current and motoneuron firing rate modulation. *J. Neurophysiol.* **1995**, *74*, 793–801. [CrossRef] [PubMed]
3. Stern, E.A.; Kincaid, A.E.; Wilson, C.J. Spontaneous subthreshold membrane poten-tial fluctuations and action potential variability of rat corticostriatal and striatal neurons in vivo. *J. Neurophysiol.* **1997**, *77*, 1697–1715. [CrossRef] [PubMed]
4. Dorval, A.D.; White, J.A. Channel noise is essential for perithreshold oscillations in entorhinal stellate neurons. *J. Neurosci.* **2005**, *25*, 10025–10028. [CrossRef] [PubMed]
5. Haider, B.; Duque, A.; Hasenstaub, A.R.; McCormick, D.A. Neocortical network activity in vivo is generated through a dynamic balance of excitation and inhibition. *J. Neurosci.* **2006**, *26*, 4535–4545. [CrossRef] [PubMed]
6. Beck, J.M.; Ma, W.J.; Kiani, R.; Hanks, T.; Churchland, A.K.; Roitman, J.; Shadlen, M.N.; Latham, P.E.; Pouget, A. Probabilistic population codes for Bayesian decision making. *Neuron* **2008**, *60*, 1142–1152. [CrossRef] [PubMed]
7. Maoz, O.; Tkačik, G.; Esteki, M.S.; Kiani, R.; Schneidman, E. Learning probabilistic neural representations with randomly connected circuits. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 25066–25073. [CrossRef] [PubMed]
8. Fayaz, S.; Fakharian, M.A.; Ghazizadeh, A. Stimulus presentation can enhance spik-ing irregularity across subcortical and cortical regions. *PLoS Comput. Biol.* **2022**, *18*, e1010256. [CrossRef] [PubMed]
9. Rinzel, J.; Miller, R.N. Numerical calculation of stable and unstable periodic solu-tions to the Hodgkin-Huxley equations. *Math. Biosci.* **1980**, *49*, 27–59. [CrossRef]
10. Rowat, P. Interspike interval statistics in the stochastic Hodgkin-Huxley model: Coex-istence of gamma frequency bursts and highly irregular firing. *Neural Comput.* **2007**, *19*, 1215–1250. [CrossRef]
11. Austin, T.D. The emergence of the deterministic Hodgkin-Huxley equations as a limit from the underlying stochastic ion channel mechanism. *Ann. Appl. Prob.* **2008**, *18*, 1279–1325. [CrossRef]
12. Montbrió, E.; Pazó, D. Exact mean-field theory explains the dual role of electrical synapses in collective synchronization. *Phys. Rev. Lett.* **2020**, *125*, 248101. [CrossRef] [PubMed]
13. Goldman, J.S.; Kusch, L.; Aquilue, D.; Yalcinkaya, B.H.; Depannemaecker, D.; Ancourt, K.; Nghiem, T.-A.E.; Jirsa, V.; Destexhe, A. A comprehensive neural simulation of slow-wave sleep and highly responsive wakefulness dynamics. *Front. Comput. Neurosci.* **2021**, *16*, 1058957. [CrossRef] [PubMed]
14. Ostojic, S.J. Interspike interval distributions of spiking neurons driven by fluctuating inputs. *J. Neurophysiol.* **2011**, *106*, 361–373. [CrossRef] [PubMed]
15. Armstrong, C.M.; Hille, B. Voltage-gated ion channels and electrical excitability. *Neuron* **1998**, *20*, 371–380. [CrossRef] [PubMed]
16. Chung, S.-H.; Hoyles, M.; Allen, T.; Kuyucak, S. Study of ionic currents across a model membrane channel using brownian dynamics. *Biophys. J.* **1998**, *75*, 793–809. [CrossRef]
17. Stoll, E.A. A thermodynamical model of non-deterministic computation in cortical neural networks. *Phys. Biol.* **2024**, *21*, 016003. [CrossRef]
18. Born, M. Statistical Interpretation of Quantum Mechanics. *Science* **1955**, *122*, 675–679. [CrossRef] [PubMed]
19. Pauli, W. Über den Zusammenhang des Abschlusses der Elektronengruppen im Atom mit der Komplexstruktur der Spektren. *Z. Phys.* **1925**, *31*, 765–783. [CrossRef]
20. Caldeira, A.O.; Leggett, A.J. Path integral approach to quantum Brownian motion. *Phys. A* **1983**, *121*, 587–616. [CrossRef]
21. Caldeira, A.O.; Leggett, A.J. Quantum tunnelling in a dissipative system. *Ann. Phys.* **1983**, *149*, 374–456. [CrossRef]
22. Chang, L.-D.; Waxman, D. Quantum Fokker-Planck equation. *J. Phys. C* **1985**, *18*, 5873–5879. [CrossRef]
23. Khriplovich, I.B.; Lamoreaux, S.K. *CP Violation Without Strangeness: Electric Dipole Moments of Particles, Atoms, and Molecules*; Springer: Berlin, Germany, 2012
24. Feynman, R.P. Forces in Molecules. *Phys. Rev.* **1939**, *56*, 340. [CrossRef]
25. Esteve, J.G.; Falceto, F.; Garcia Canal, C. Generalization of the Hellmann-Feynman theorem. *Phys. Lett. A* **2010**, *374*, 819–822. [CrossRef]
26. Stoll, E.A. An energy-efficient process of non-deterministic computation drives the emergence of predictive models and exploratory behavior. *Front. Cognit.* **2024**, *2*, 1171273 [CrossRef]
27. Bethe, H.A. The electromagnetic shift of energy levels. *Phys. Rev.* **1947**, *72*, 339. [CrossRef]
28. Holstein, B.A. The van der waals interaction. *Am. J. Phys.* **2001**, *69*, 441–449. [CrossRef]

29. Landauer, R. Irreversibility and heat generation in the computing process. *IBM J. Res. Dev.* **1961**, *5*, 183–191. [CrossRef]
30. Berut, A.; Arakelyan, A.; Petrosyan, A.; Ciliberto, S.; Dillenschneider, R.; Lutz, E. Experimental verification of Landauer's principle linking information and thermodynamics. *Nature* **2012**, *483*, 187–189. [CrossRef]
31. Jun, Y.; Gavrilov, M.; Beckhoefer, J. High-precision test of Landauer's principle in a feedback trap. *Phys. Rev. Lett.* **2014**, *113*, 190601. [CrossRef]
32. Yan, L.L.; Xiong, T.P.; Rehan, K.; Zhou, F.; Liang, D.F.; Chen, L.; Zhang, J.Q.; Yang, W.L.; Ma, Z.H.; Feng, M. Single-atom demonstration of the quantum Landauer principle. *Phys. Rev. Lett.* **2018**, *120*, 210601. [CrossRef] [PubMed]
33. Von Neumann, J. *Mathematical Foundations of Quantum Mechanics*; Springer: Berlin, Germany, 1932; Volume 205, p. 116277.
34. Callen, H.B.; Welton, T.A. Irreversibility and generalized noise. *Phys. Rev.* **1951**, *83*, 34. [CrossRef]
35. Tegmark, M. Why the brain is probably not a quantum computer. *Inf. Sci.* **2000**, *128*, 155–179. [CrossRef]
36. Kubo, R. The fluctuation-dissipation theorem. *Rep. Prog. Phys.* **1966**, *29*, 255–284. [CrossRef]
37. Martyushev, L.M.; Nazarova, A.S.; Seleznev, V.D. On the problem of minimum en-tropy production in the nonequilibrium stationary state. *J. Phys. A* **1966**, *40*, 371–380. [CrossRef]
38. Taube, J.S. Interspike interval analyses reveal irregular firing patterns at short, but not long, intervals in rat head direction cells. *J. Neurophysiol.* **2010**, *104*, 1635–1648. [CrossRef] [PubMed]
39. Naito, T.; Sadakane, O.; Okamoto, M.; Sato, H. Orientation tuning of surround sup-pression in lateral geniculate nucleus and primary visual cortex of cat. *Neuroscience* **2007**, *149*, 962–975. [CrossRef] [PubMed]
40. Averbeck, B.B.; Latham, P.E.; Pouget, A. Neural correlations, population coding and computation. *Nat. Rev. Neurosci.* **2006**, *7*, 358–366. [CrossRef]
41. Shadlen, M.N.; Newsome, W.T. The variable discharge of cortical neurons: Implica-tions for connectivity, computation, and information coding. *J. Neurosci.* **1998**, *18*, 3870–3896. [CrossRef]
42. Lee, D.; Port, N.L.; Kruse, W.; Georgopoulos, A.P. Variability and correlated noise in the discharge of neurons in motor and parietal areas of the primate cortex. *J. Neurosci.* **1998**, *18*, 1161–1170. [CrossRef]
43. Cui, H.; Anderson, R.A. Different representations of potential and selected motor plans by distinct parietal areas. *J. Neurosci.* **2011**, *31*, 18130–18136. [CrossRef]
44. Jackson, A.; Mavoori, J.; Fetz, E.E. Correlations between the same motor cortex cells and arm muscles during a trained task, free behavior, and natural sleep in the macaque monkey. *J. Neurophysiol.* **2007**, *97*, 360–374. [CrossRef]
45. Russo, A.A.; Bittner, S.R.; Perkins, S.M.; Seely, J.S.; London, B.M.; Lara, A.H.; Miri ,A.; Marshall, N.J.; Kohn, A.; Jessell, T.M.; et al. Motor cortex embeds muscle-like commands in an untangled population response. *Neuron* **2018**, *97*, 953–966. [CrossRef]
46. Hubel, D.H.; Wiesel, T.N. Receptive fields of single neurones in the cat's striate cor-tex. *J. Physiol.* **1959**, *148*, 574–591. [CrossRef]
47. Mendonça, P.R.; Vargas-Caballero, M.; Erdélyi, F.; Szabó, G.; Paulsen, O.; Robinson, H.P. Stochastic and deterministic dynamics of intrinsically irregular firing in cortical inhibitory interneurons. *eLife* **2016**, *5*, e16475. [CrossRef]
48. Wimmer, K.; Compte, A.; Roxin, A.; Peixoto, D.; Renart, A.; de la Rocha, J. Sensory integration dynamics in a hierarchical network explains choice probabilities in cortical area MT. *Nat. Commun.* **2015**, *6*, 6177 [CrossRef] [PubMed]
49. Melloni, L.; Schwiedrzik, C.M.; Müller, N.; Rodriguez, E.; Singer, W. Expectations change the signatures and timing of electro-physiological correlates of perceptual awareness. *J. Neurosci.* **2011**, *31*, 1386–1396. [CrossRef]
50. Huang, S.; Hong, S.; De Schutter, E. Non-linear leak currents affect mammalian neuron physiology. *Front. Cell. Neurosci.* **2015**, *9*, 432.
51. Stoll, E.A. Random electrical noise drives non-deterministic computation in cortical neural networks. *bioRxiv* **2022**. [CrossRef]
52. Isojima, Y.; Isoshima, T.; Nagai, K.; Kikuchi, K.; Nakagawa, H. Ultraweak biochem-iluminescence detected from rat hippocampal slices. *Neuroreport* **1995**, *6*, 658–660. [CrossRef]
53. Kobayashi, M.; Takeda, M.; Sato, T.; Yamazaki, Y.; Kaneko, K.; Ito, K.; Kato, H.; Inaba, H. In vivo imaging of spontaneous ultraweak photon emission from a rat's brain correlated with cere-bral energy metabolism. *Neurosci. Res.* **1999**, *34*, 103–113. [CrossRef] [PubMed]
54. Kataoka, Y.; Cui, Y.; Yamagata, A.; Niigaki, M.; Hirohata, T.; Oishi, N.; Watanabe, Y. Activity-dependent neural tissue oxidation emits intrinsic ultraweak photons. *Biochem. Biophys. Res. Commun.* **2001**, *285*, 1007–1011. [CrossRef] [PubMed]
55. Tang, R.; Dai, J. Spatiotemporal imaging of glutamate-induced biophotonic activities and transmission in neural circuits. *PLoS ONE* **2014**, *9*, e85643. [CrossRef] [PubMed]
56. Amaroli, A.; Marcoli, M.; Venturini, A.; Passalacqua, M.; Agnati, L.F.; Signore, A.; Raffetto, M.; Maura, G.; Benedicenti, S.; Cervetto, C. Near-infrared laser photons induce glutamate release from cerebro-cortical nerve terminals. *J. Biophotonics* **2018**, *11*, e201800102. [CrossRef] [PubMed]
57. Naeser, M.A.; Ho, M.D.; Martin, P.I.; Hamblin, M.R.; Koo, B.B. Increased functional connectivity within intrinsic neural networks in chronic stroke following treatment with red/near-infrared transcranial photobiomodulation. *Photobiomodul. Photomed. Laser Surg.* **2020**, *38*, 115–131. [PubMed]
58. Tan, X.; Rajguru, S.; Young, H.; Xia, N.; Stock, S.R.; Xiao, X.; Richter, C.P. Radiant en-ergy required for infrared neural stimulation. *Sci. Rep.* **2015**, *5*, 13273. [CrossRef] [PubMed]
59. Stoll, E.A. Modeling electron interference at the neuronal membrane yields a holo-graphic projection of representative information content. *bioRxiv* **2022**. [CrossRef]

60. Levy, W.B.; Calvert, V.G. Communication consumes 35 times more energy than computation in human cortex, but both costs are needed to predict synapse number. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2008173118. [CrossRef]
61. Buzsaki, G.; Draguhn, A. Neuronal oscillations in cortical networks. *Science* **2004**, *304*, 1926–1929. [CrossRef]
62. Engel, A.K.; Singer, W. Temporal binding and the neural correlates of sensory awareness. *Trends Cogn. Sci.* **2001**, *5*, 16–25. [CrossRef]
63. Stacey, W.C.; Krieger, A.; Litt, B. Network recruitment to coherent oscillations in a hippocampal computer model. *J. Neurophysiol.* **2011**, *105*, 1464–1481. [CrossRef]
64. Whittington, M.A.; Cunningham, M.O.; LeBeau, F.E.N.; Racca, C.; Traub, R.D. Mul-tiple origins of the cortical gamma rhythm. *Dev. Neurobiol.* **2010**, *71*, 92–106. [CrossRef]
65. Timofeev, I.; Bazhenov, M.; Seigneur, J.; Sejnowski, T. Neuronal synchronization and thalamocortical rhythms in sleep, wake, and epilepsy. In *Jasper's Basic Mechanisms of the Epilepsies*, 4th ed.; Oxford University Press: Oxford, UK, 2012.
66. Gansel, K.S. Neural synchrony in cortical networks: Mechanisms and implications for neural information processing and coding. *Front. Integr. Neurosci.* **2022**, *16*, 900715. [CrossRef]
67. Astumian, R.D.; Moss, F. The constructive role of noise in fluctuation driven transport and stochastic resonance. *Chaos* **1998**, *8*, 533–538. [CrossRef] [PubMed]
68. Lucarini, V. Stochastic resonance in non-equilibrium systems. *Phys. Rev. E* **2019**, *100*, 062124. [CrossRef]
69. Lindner, B.; Schimansky-Geier, L. Analytical approach to the stochastic Fitz-Hugh-Nagumo system and coherence resonance. *Phys. Rev. E* **1999**, *60*, 7270–7276. [CrossRef]
70. Lindner, B.; Garcia-Ojalvo, J.; Neiman, A.; Schimansky-Geier, L. Effects of noise in excitable systems. *Phys. Rep.* **2004**, *392*, 321–424. [CrossRef]
71. Valenti, D.; Augello, G.; Spagnolo, B. Dynamics of a FitzHugh-Nagumo system sub-jected to autocorrelated noise. *Eur. Phys. J. B* **2008**, *65*, 443–451. [CrossRef]
72. Surazhevsky, I.A.; Demim, V.A.; Ilyasov, A.I.; Emelyanov, A.V.; Nikiruy, K.E.; Rylkov, V.V.; Shchanikov, S.A.; Bordanov, I.A.; Gerasimova, S.A.; Guseinov, D.V.; et al. Noise-assisted persistence and re-covery of memory state in a memristive spiking neuromorphic network. *Chaos* **2021**, *146*, 110890.
73. Geisler, C.; Brunel, N.; Wang, X.J. Contributions of intrinsic membrane dynamics to fast network oscillations with irregular neuronal discharges. *J. Physiol.* **2005**, *94*, 4344–4361.
74. Brunel, N. Dynamics of sparsely-connected networks of excitatory and inhibitory spik-ing neurons. *J. Comput. Neurosci.* **2000**, *8*, 183–208. [CrossRef]
75. Csibra, G.; Davis, G.; Spratling, M.W.; Johnson, M.H. Gamma oscillations and ob-ject processing in the infant brain. *Science* **2000**, *290*, 1582–1585. [CrossRef] [PubMed]
76. Herrmann, C.S.; Knight, R.T. Mechanisms of human attention: Event-related poten-tials and oscillations. *Neurosci. Biobehav. Rev.* **2001**, *25*, 465–476. [CrossRef] [PubMed]
77. Hameroff, S.R.; Penrose, R. Consciousness in the universe: A review of the 'OrchOR' theory. *Phys. Life Rev.* **2014**, *11*, 39–78. [CrossRef] [PubMed]
78. Khalid, M.; Wu, J.; Ali, T.M.; Ameen, T.; Altaher, A.S.; Moustafa, A.A.; Zhu, Q.; Xiong, R. Cortico-Hippocampal computational modeling using quantum-inspired neural networks. *Front. Comput. Neurosci.* **2020**, *14*, 80. [CrossRef] [PubMed]
79. Wang, Z.; Xu, M.; Zhang, Y. Quantum pulse coupled neural network. *Neural Netw.* **2022**, *152*, 105–117. [CrossRef] [PubMed]
80. Jeswal, S.K.; Chakraverty, S. Recent developments and applications in quantum neu-ral network: A review. *Arch. Comput. Methods Eng.* **2019**, *26*, 793–807. [CrossRef]
81. Adair, R.K. Noise and stochastic resonance in voltage-gated ion channels. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 12099–12104. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

AppliedMath

MDPI

Article

Network Goodness Calculus Propositions

Marina Bershadsky [1,†], Božidar Ivanković [2,*,†] and Marko Pušić [2,†]

[1] Department of Computer Science, Shamoon College of Engineering (SCE), 56 Bialik, Be'er Sheva 8410802, Israel; marinab@ac.sce.ac.il
[2] Hrvatsko Zagorje Krapina University of Applied Sciences, Šetalište Hrvatskog Narodnog Preporoda 6, 49000 Krapina, Croatia; marko.pusic@vhzk.hr
* Correspondence: bozidar.ivankovic@vhzk.hr; Tel.: +385-98-407-189
† These authors contributed equally to this work.

Abstract: We coin the term "network goodness" for a value we define for a network embedded in a given environment as a metric that describes the suitability of that network for meeting a demand. Three formulas are proposed to calculate the metric from three variable values. The first variable considers parts of the environment gravitated by the network. For these parts of the environment, we define a value that measures user costs refusing them the use of the network. Last but not least, the network maintenance costs are considered. The results are obtained after focusing on infrastructure and transport networks, but can be used for other types of networks as well.

Keywords: network evaluation; network goodness; demand center; network inertia

1. Introduction

A network is designed to meet the demands of a specific environment. There are three essential factors to consider when evaluating a network in any environment:

- The coverage range of the network within the environment;
- The availability of the network for users, as reflected in the costs incurred by the user;
- The maintenance costs associated with the network.

If the demand exceeds network capacity, some areas may not be serviced, service may become more expensive, or maintenance expenses may increase. Also, several networks may be embedded in the same environment to serve an environment's demand.

In order to evaluate the function of a network, in this paper, we defined the "network goodness" function and propose formulas for calculating goodness values and comparing two or more infrastructure networks embedded in the same environment. We also thoroughly prove all theorems, propositions, lemmas, and corollaries, since no similar approach has, to date, been conducted in the available literature.

Currently, the predominant term used for evaluating network function is "Quality of Service" (QoS), and many millions of papers have been published on the subject over the years. A recently published work by Kanellopoulos et al. in [1] offers an extensive survey of network architectures and protocols. However, the authors only present a qualitative comparison of network requirements for some Smart City Applications (SCAs) and do not attribute an overall score that would enable ranking amongst the alternatives. Karakus and Durresi wrote an exhaustive survey paper [2] for comprehensively examining QoS in Software-Defined Networking (SDN)/OpenFlow networks. The authors provide a detailed overview of the SDN architecture and divide 43 different methods according to two approaches for maintaining the ability of a network to provide the required services for a selected traffic scenario. While it is clear that the soft Quality and Outcomes Framework (QoF) outperforms the hard QoF approach, no quantitative metric is defined.

check for updates

Citation: Bershadsky, M.; Ivanković, B.; Pušić, M. Network Goodness Calculus Propositions. *AppliedMath* **2024**, *4*, 828–842. https://doi.org/10.3390/appliedmath4030044

Academic Editor: Takayuki Hibi

Received: 12 January 2024
Revised: 10 June 2024
Accepted: 20 June 2024
Published: 2 July 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

In other works, network embeddings are discussed in terms of solving complex problems once a network is defined. In [3], Social Network Analysis (SNA) is presented as a powerful technique for building and maintaining social relationships within the network. In social networks, users interact and communicate with one another and also post personalized content. Novel Dynamic Attributed Network Embedding (DANE) is experimentally studied in [4] to advance learning tasks such as node classification, network clustering, and link prediction. Network embedding leverages node proximity to learn a low-dimensional node vector representation for each node within the existing social network, and hence studies how social relationships are established and evolve. However, the SNA values are also only given qualitatively, as DANE is evaluated. In [5], Salkhi et al. consider networks as a means of evaluating structural isolation and node integration, assessing their accessibility and navigability within the graph, and predicting potential structural changes that may impact the network's evolution. However, again, no quantitative value is provided to measure either isolation or integration in the given scenario.

The application area we choose to showcase our approach is public transport networks. Typically, traffic infrastructure networks emerge naturally as demand increases in an environment. However, an existing network cannot meet the environment's needs in time. In the last section, we presented several alternatives for public transportation routes that could accommodate a transportation mode with sufficient volume and speed to make public transport as attractive as possible. Despite searching the available literature, we could not find an instrument for quantitatively selecting heuristically drawn variants.

Bogdanova et al. investigate the transport facilities of a comfortable urban environment in [6], where the "network comfort" is measured by the number of users with less than 500 (or 800) meters to travel among the total population. The security rating of the public transport network is calculated by dividing the number of citizens injured in public transport by the total number of passengers using it. Transportation systems have become increasingly vital to modern economies, as Colon et al. describe in [7]. In [8]. Guliot proposes a decision tool that offers several alternative scenarios for reconfiguring transit services. Each scenario includes a set of existing lines to be closed and the corresponding delay profile caused by the closure of each line. Another study, by Wang et al. [9], addresses the unbalanced urban development by improving transportation efficiency, reducing traffic congestion and accidents, and coordinating transport transport. Kuo et al. in [10] emphasize the importance of readily available public transportation that provides real-time demand, offers convenient and high quality service, and operates in an energy-efficient way. In [11], Nes et al. present an optimization model for selecting routes, assigning frequencies, and determining the number of passengers on the existing network by generating a set of possible routes.

The rest of this paper is organized as follows. In Section 2, we introduce our terminology and outline the methodology, including lemmas and theorems and their proofs that underscore our approach. In Section 3, we present results and discuss them. In the following section, we offer an example of our proposed approach (Section 4), and in the final section, we offer conclusions (Section 5).

2. Terminology and Basic Theory

In this section, we define the terms we then use in theorems and lemmas. We draw our terminology from the fields of set theory, graph theory, and analytic geometrical calculus. First, we assume that any environment is within some "ground": a state, a continent, or the Earth overall. The ground is a two-dimensional plane.

Definition 1 (Network environment and size)**.** *Network environment $\mathcal{U}$ is part of the ground defined uniquely by its administrative borders or in any other unique way that determines which locations are within and which outside the ground. The area of $\mathcal{U}$ is denoted $\mu(\mathcal{U})$, and is measured in squares.*

The network environment is divided into zones comprising demand points where demand is concentrated. Different zones do not share the same demand point(s): mathematically, they are mutually disjointed.

Definition 2 (Demand zone and demand point)**.** *Demand zone $\mathcal{T}_i, i = 1, \ldots, n$ is part of the network environment: $\mathcal{T}_i \subset \mathcal{U}$. Further, for all $i, j, \mathcal{U} = \bigcup_{i=1}^{n} \mathcal{T}_i$ and $\mathcal{T}_i \cap \mathcal{T}_j = \emptyset$ for $i \neq j$.*

Demand point $P_{ij} \in \mathcal{T}_i, j = 1, \ldots, m(i)$ is a point with positive associated value $Q(P_{ij})$, which is the demand concentrated at P_{ij}.

The number of demand points is $m(i)$, and it varies from zone to zone. The formal definition is given below.

Definition 3. *Suppose that for each $\mathcal{T}_i$ there exists $m(i) \in \mathbb{N}$, and a finite set $\bar{\mathbb{N}} = \{1, 2, \ldots, m(i)\}$ with injective function $p_i : \bar{\mathbb{N}} \to \mathcal{T}_i$, denoted $p_i(j) = P_{ij}$, and $p_i(j) \neq p_i(k)$ when $j \neq k$. Then, $Q_i : p_i(\bar{\mathbb{N}}) \to \mathbb{R}^+$ is the demand quantity given in demand point P_{ij} denoted $Q_i(P_{ij}) = Q_{ij}$.*

To calculate the network goodness, we require coordinates, a pair of numbers that uniquely determines the position of a point on the plane.

Definition 4 (Coordinated network environment)**.** *An environment is coordinated if any point can be uniquely defined by a pair of real numbers (x, y).*

The coordinates are necessary for algebraic calculations, and we guarantee this by proving the following Lemma.

Lemma 1. *Any network environment can be coordinated.*

Proof of Lemma 1. We set a well-known location in the environment as a pair $(0, 0)$. We set the west–east direction through $(0, 0)$ as the abscissa and the south–north direction as the ordinate axis. Then, any point $P \in \mathcal{U}$ is uniquely defined by pair $P \equiv (x, y)$. Usually, it is $P_{ij} = (x_{ij}, y_{ij})$. □

Our objective is to evaluate a network, and for this we need to algebraically define a "segment" between two points.

Definition 5. *A segment between two points, A and B, with coordinates (x_A, y_A) and (x_B, y_B), respectively, is a set consisting of all the points on the straight line connecting the two points.*

$$\overline{AB} = \{P : P = (1 - t)A + tB, t \in [0, 1]\}. \tag{1}$$

In graph terminology [12], a graph is a pair $G = (N, E)$ defined by a given set of nodes N and a set of edges E given as a relation, $E \subseteq N \times N$. This means that E is a set of pairs taken from the set of all the possible pairs of all the nodes given in N. In our paper, a network is a connected graph $G = (N, E)$, where E is a set of segments and N is a set of their endpoints. Because $E \subseteq N \times N$, a network is defined correctly, and we prove it formally in the following theorem, where $E = N \times N$.

Theorem 1. *Any set of segments defines a graph.*

Proof of Theorem 4. For a given set of segments $\{\overline{A_i B_i},\ i = 1, 2, \ldots s\}$, we denote every segment $\overline{A_i B_i} = (A_i, B_i)$ as an element in the set of edges $E = \{(A_i, B_i)\ i = 1, 2, \ldots s\}\}$, with $|E| = s$. Further, each segment $\overline{A_i B_i}$ gives two points into the set of nodes $N = \{A_i, B_i\ i = 1, 2, \ldots s\}\}$ with $|N| \leq 2s$ elements, since different segments may have one endpoint in

common. Then, the given segments set generates $G = (N, E)$ with $E = N \times N \subseteq N \times N$, in the well-known graph definition sense. □

A path is a sequence of chained segments. Path segments $\{\overline{A_iB_i},\ i = 1, \ldots, p\}$ are considered chained if one segment endpoint matches the starting point of the next segment or the ending point of the path: $B_i = A_{i+1},\ i = 1, \ldots, p-1$. In Figure 3, a path is illustrated. We need the path term for the following definition.

Definition 6 (Connected graph). *A graph is connected if any pair of nodes has a path that begins at one and ends at the other node.*

The concept of a connected graph is necessary to define a network in our paper, as not every graph is a network. Definition 6 leads us to define the network we evaluate in our paper, which is given below.

Definition 7 (Network). *A network is a connected graph consisting of a set of segments.*

In the following theorem, we prove that each segment defined by Equation (1) corresponds to a unique equation derived from the segment's boundary points.

Theorem 2. *We let $A = (x_A, y_A)$ and $B = (x_B, y_B)$ be points given in an environment covered by a coordinate grid. Then, segment $\overline{AB}$ is defined as set*

$$\overline{AB} = \{(x, y) \in [x_A, x_B] \times [y_A, y_B] : (y_B - y_A)x + (x_A - x_B)y + x_B y_A - x_A y_B = 0\}. \quad (2)$$

Proof of Theorem 2. From Lemma 1 and Equation (1), we obtain $x - x_A = t(x_B - x_A)$, $y - y_A = t(y_B - y_A)$, and $\dfrac{x - x_A}{x_B - x_A} = \dfrac{y - y_A}{y_B - y_A}$. Because $t \in [0, 1]$, we have $(x, y) \in [x_A, x_B] \times [y_A, y_B]$. Now, $(y_B - y_A)(x - x_A) - (x_B - x_A)(y - y_A) = 0$ produces Equation (2).

Further, from Equation (2), it emerges that $\dfrac{x - x_A}{x_B - x_A} = \dfrac{y - y_A}{y_B - y_A} = t$. If $(x, y) \in [x_A, x_B] \times [y_A, y_B]$, then $t \in [0, 1]$, and Equation (1) results immediately. □

Figure 1 illustrates the positions of different points outside the segment concerning it. Having proven Equation (2), Theorem 2 enables us to calculate the square distance between a point and a segment. In the following theorem, we derive the square distance formulas for the position of each point depicted in Figure 1.

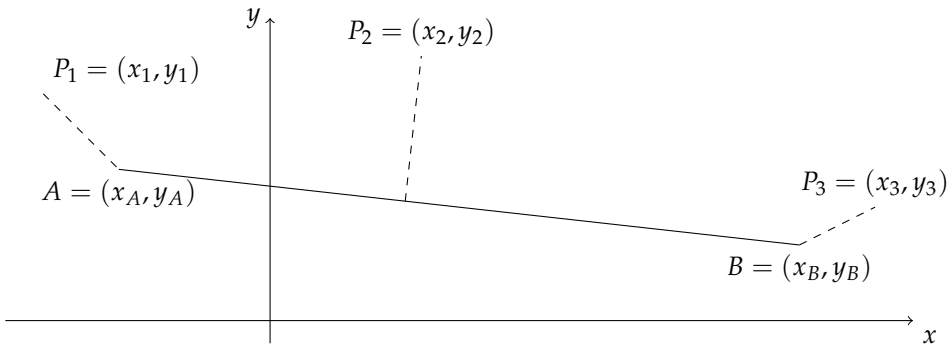


Figure 1. Distances from the segment related to demand zones positions.

Theorem 3. *The square distance between point $P = (x, y) \in \{P_1, P_2, P_3\}$ and segment $\overline{AB}$ is given by*

$$d^2(P, \overline{AB}) = (x - x_A)^2 + (y - y_A)^2, \text{ if } (x - x_A)(x_B - x_A) + (y - y_A)(y_B - y_A) < 0 \quad (3)$$

$$d^2(P, \overline{AB}) = (x - x_B)^2 + (y - y_B)^2, \text{ if } (x - x_B)(x_A - x_B) + (y - y_B)(y_A - y_B) < 0 \quad (4)$$

or

$$d^2(P, \overline{AB}) = \frac{[(y_B - y_A)x + (x_A - x_B)y + x_B y_A - x_A y_B]^2}{(x_A - x_B)^2 + (y_B - y_A)^2} \tag{5}$$

in any other case.

Equation (5) is obtained after we define $\overrightarrow{AB}$ as the linear combination of vectors $\vec{i} = (1,0) - (0,0)$ and $\vec{j} = (0,1) - (0,0)$, representing the unit steps in the east and north directions, respectively. Then, vector $\overrightarrow{AB}$ is given by

$$\overrightarrow{AB} = (x_B - x_A)\vec{i} + (y_B - y_A)\vec{j}, \tag{6}$$

where $A = (x_A, y_A)$ and $B = (x_B, y_B)$ are already defined. The cosine of angle $\angle(PAB)$, with the vertex at A, is given by formula

$$\cos(\angle PAB) = \frac{\overrightarrow{AP} \cdot \overrightarrow{AB}}{\sqrt{\overrightarrow{AP}^2 \cdot \overrightarrow{AB}^2}}. \tag{7}$$

Here, we note that $\vec{i} \cdot \vec{j} = 0$, and $\vec{i}^2 = \vec{j}^2 = 1$. Furthermore, $\cos(\angle PAB) < 0$ for $90^\circ \leq \angle PAB \leq 180^\circ$, and $\cos(\angle PAB) > 0$ for $0^\circ \leq \angle PAB \leq 90^\circ$.

In addition, we take cross-product $\vec{i} \times \vec{j}$ with $\vec{i} \times \vec{j} = -\vec{j} \times \vec{i}$ and area $|\vec{i} \times \vec{j}| = 1$. Namely, $\vec{i} \times \vec{j}$ is a vector perpendicular to $\vec{i}$ and $\vec{j}$, so it does not belong to the environment ground. Then, the Theorem 3 proof follows. More details about vector products can be found in [13].

Proof of Theorem 3. In the case where $P = P_1$, we have $\overrightarrow{AP} \cdot \overrightarrow{AB} < 0$, and then Equation (3) is valid. If $P = P_3$, then $\overrightarrow{BP} \cdot \overrightarrow{BA} < 0$, and Equation (4) is valid.

In other cases, we equate the areas of triangle $\Delta(PAB)$ calculated using two approaches,

$$\frac{1}{2} \cdot d(P, \overline{AB}) \cdot |\overline{AB}| = \frac{1}{2} |\overrightarrow{AP} \times \overrightarrow{BP}|, \tag{8}$$

where the vector product is calculated in a three-dimensional space after environment $\mathcal{U}$ is embedded in the sense $P = (x, y, 0)$. The absolute value on the right-hand side in Equation (8) equals

$$\begin{aligned} & |[(x - x_A)\vec{i} + (y - y_A)\vec{j}] \times [(x - x_B)\vec{i} + (y - y_B)\vec{j}]| \\ = \; & |x(y_A - y_B) + y(x_B - x_A) + x_A y_B - x_B y_A| \cdot |\vec{i} \times \vec{j}| \\ = \; & |(y_B - y_A)x + (x_A - x_B)y + x_B y_A - x_A y_B|, \end{aligned} \tag{9}$$

because area $|\vec{i} \times \vec{j}| = 1$. On the left-hand side of Equation (8), we have

$$d(P, \overline{AB}) \cdot \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2}. \tag{10}$$

Substituting Equations (9) and (10) into Equation (8), we obtain (5) after tidying up. □

In Definition 2, we define a demand point as a point with a positive number assigned to it that represents the demand at that point. It is important to note that not every point has a demand value. According to Definition 2, a demand zone contains only a finite number of demand points. If a point in the demand zone has no demand, it is not a demand point.

3. Results and Discussion

In this section, we explain the basics of our theory, starting with the zone demand center. Each demand zone has a unique demand center point. This point is calculated similarly to the center of gravity calculation in physics. Here, the calculation is given as a theorem that follows.

Theorem 4 (Zone demand center). *Each demand zone $\mathcal{T}_i$ has a unique demand center point P_i with coordinates $P_i = (x_{\mathcal{T}_i}, y_{\mathcal{T}_i})$, calculated using formula*

$$\begin{aligned} x_{\mathcal{T}_i} &= \frac{Q_{i1} \cdot x_{i1} + Q_{i2} \cdot x_{i2} + \cdots + Q_{im(i)} \cdot x_{im(i)}}{Q(\mathcal{T}_i)} \\ y_{\mathcal{T}_i} &= \frac{Q_{i1} \cdot y_{i1} + Q_{i2} \cdot y_{i2} + \cdots + Q_{im(i)} \cdot y_{im(i)}}{Q(\mathcal{T}_i)}, \end{aligned} \tag{11}$$

where $Q(\mathcal{T}_i) = Q_{i1} + Q_{i2} + \cdots + Q_{im(i)}$ is the total zone demand concentrated at P_i.

The number of demand points within $\mathcal{T}_i$ depends on $\mathcal{T}_i$, and thus it is denoted $m(i)$.

Proof of Theorem 4. We prove the x-coordinate calculation from Equation (11) taking points $P_1 := P_{i1}$ and $P_3 := P_{i3}$ from Figure 1. If we suppose that mass $Q_{i1} > Q_{i3}$, then x_{i1}, x_{i3}, and their center of gravity x_{i13} are positioned on the abscissa axis as shown in Figure 2.



Figure 2. Calculating the two-demand-point demand center.

The center x_{i13} of the "teeter" in Figure 2 must satisfy the following "balance" law:

$$Q(P_{i1}) \cdot (x_{i13} - x_{i1}) = Q(P_{i3}) \cdot (x_{i3} - x_{i13}). \tag{12}$$

According to Equation (12), the center of gravity abscissa position is

$$x_{i13} = \frac{Q_{i1}x_{i1} + Q_{i1}x_{i1}}{Q_{i1} + Q_{i1}}. \tag{13}$$

Next, we extend the calculation to include $P_2 := P_{i2}$. Supposing $Q(P_{i2}) \geq Q_{i1} + Q_{i3}$, we obtain the center of gravity between (13) with $Q_{i1} + Q_{i3}$ and $P_2 := P_{i2}$ with $Q(P_{i2})$:

$$x_{i132} = \frac{Q_{i1}x_{i1} + Q_{i3}x_{i3} + Q_{i2}x_{i2}}{Q_{i1} + Q_{i3} + Q_{i2}}. \tag{14}$$

Proceeding, we obtain a penultimate center of gravity:

$$x_{i1\ldots[m(i)-1]} = \frac{\sum_{j=1}^{m(i)-1} Q_{ij}x_{ij}}{\sum_{j=1}^{m(i)-1} Q_{ij}}. \tag{15}$$

Together with the last demand point, we reach Equation (11) (this procedure is commonly termed "a pondering average"). □

The square distance between the demand zone and the network is the minimum square distance between the demand zone center and a network segment. The following theorem guarantees existence and uniqueness.

Theorem 5. *For a given network $\bigcup_{j=1}^{s} \overline{A_jB_j}$, and for each P_i, where $i = 1, \ldots, n$, there exists*

$$d_i^2 = d^2(P_i) = \min_j \{d^2(P_i, \overline{A_jB_j})\}, \text{ for each } i = 1, \ldots, n. \tag{16}$$

In the proof below, we show induction as a suggestion for program applications. In the Example section, we use only the Excel tool to calculate the minimum distance, although the number of segments and zones is large.

Proof of Theorem 5. Because the network is finite, the minimum distance always exists. Next, we perform induction by index j, $1 \leq j \leq s$.

For induction basic $j = 1$, we have $d_i^2 = d^2(P_i, \overline{A_1B_1}) = \min\limits_{j\leq 1}\{d^2(P_i, \overline{A_jB_j})\}$. For $j = 2$, we take $d_i^2 = \min\left\{d^2(P_i, \overline{A_2B_2}), \min\limits_{j\leq 1}\{d^2(P_i, \overline{A_jB_j})\}\right\} = \min\limits_{j\leq 2}\{d^2(P_i, \overline{A_jB_j})\}$.

For step $j = s-1$, we suppose that for graph $\bigcup\limits_{j=1}^{s-1} \overline{A_jB_j}$ there exists $d_i^2 = \min_{j\leq s-1}\{d^2(P_i, \overline{A_jB_j})\}$.

If $j = s$, we separate $\bigcup\limits_{j=1}^{s} \overline{A_jB_j} = \bigcup\limits_{j=1}^{s-1} \overline{A_jB_j} \cup \overline{A_sB_s}$, and then we have

$$\begin{aligned} d_i^2 &= \min\left\{\min_{j\leq s-1}\{d^2(P_i, \overline{A_jB_j})\}, d_i^2(P_i, \overline{A_sB_s})\right\} \\ &= \min_{j\leq s}\{d^2(P_i, \overline{A_jB_j})\}, \end{aligned}$$

for each $i = 1, \ldots, n$. It has been proven that a square distance exists for each zone $\mathcal{F}_i$ with a demand center at point P_i. It is unique because if the minimum exists, it is always only one value. □

Next, we introduce the "inertia" of the demand zone. We remind that in the overall concept of network goodness, a higher score indicates better network service within the environment.

To determine the network goodness, we calculate the inertia of the demand zone that "gravitates" to the network. The inertia of a demand zone is determined by multiplying the zone's demand with the minimum square distance between the demand zone center and a network segment. We formulate the inertia of a demand zone using the following definition.

Definition 8 (Demand zone inertia)**.** *The inertia of the ith demand zone is product*

$$M_i = Q(\mathcal{T}_i) \cdot d_i^2,\ i = 1, \ldots, n, \tag{17}$$

where d_i^2 is calculated from Equation (16) and $Q(\mathcal{T}_i)$ is calculated in Theorem 4.

It should be noted that not all zones where there is demand for network service are considered when calculating the system's inertia. To use the network service, their users must connect to the network independently.

In an infrastructure network, two types of distance influence a user's choice of network for a service. A short distance attracts the user to choose the proposed network over other options. For instance, if a bus stop is close to someone's apartment, they would only use the bus instead of their vehicle. Conversely, a greater distance is a limit beyond which the user will use the given network indirectly. In the following definition, we formally define both types of distances: gravitating and attracting distances and zones.

Definition 9 (Gravitating and attracting zones and distances)**.** *Demand zone $\mathcal{T}_i$, where $i = 1, \ldots, n$, **gravitates** to the network iff $d_i < D$ for gravitating distance D, which is a value determined in advance.*

*Demand zone $\mathcal{T}_j$, with $j = 1, \ldots, n$, is **attracted** by the network iff $d_j < d$ for an attracting distance d, which is a value determined in advance, too.*

Further discussion of d and D related to public transport networks can be found in [14]. The value of D represents an acceptable distance of up to two kilometers that one can walk to reach the network. At the same time, d is the distance of a few hundred meters that a user is willing to walk instead of using any different mode of transportation.

The total inertia of a network is the sum of the inertia of all gravitating demand zones. The quality of a network is determined by how low its network inertia is. A simple example is illustrated in Figure 3.

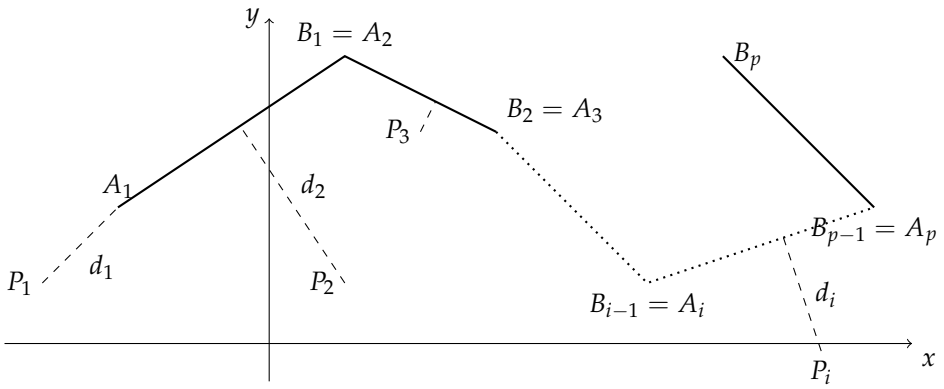


Figure 3. Calculating the moment zone inertia along the path in a network.

Definition 10 (Network inertia and gravitating zones number)**.** *The network inertia is calculated by adding up the inertia of all the gravitating zones in the network.*

$$M = \sum_{i=1}^{k} M_i = \sum_{i=1}^{k} Q(\mathcal{T}_i) d_i^2, \tag{18}$$

There are k demand zones that gravitate towards the network, each represented by $\mathcal{T}_i$.

The total length of a network is the sum of its segment lengths.

Definition 11 (Network length)**.** *For a given network* $\bigcup_{j=1}^{s} \overline{A_j B_j}$*, the length is*

$$l = \sum_{j=1}^{s} |A_j B_j| = \sum_{j=1}^{s} \sqrt{(x_{B_j} - x_{A_j})^2 + (y_{B_j} - y_{A_j})^2}. \tag{19}$$

When calculating the goodness of a network, we take into account three essential factors: the number of gravitating zones (k), the inertia (M), and the length (l). A network is considered good if it has as many zones connected and its inertia and length are as small as possible.

Next, we provide a weak definition for a network goodness function formula necessary for proving the propositions as our main results.

Definition 12. *A network goodness function of three variables is properly defined by a formula* $g(k, M, l)$ *if it increases as k increases and decreases as l and M decrease.*

When properly defined according to Definition 12, the value of the network goodness function provides a quantitative network evaluation in the sense that two or more networks embedded in the same environment are ranked such that a network with a higher goodness value is the "better".

We propose three formulas that increase with the number of gravitating zones and decrease with the length and inertia of the network. They are presented in the order they

were developed and have the same proof because increasing and decreasing are governed by positive and negative partial derivatives, respectively.

Proposition 1. *If k zones are attracted towards the network, and the network has an inertia of M with a length of l, then the network goodness function is given by formula*

$$g_1(k, M, l) = \frac{k^2 + k^3}{l^2 + M}. \tag{20}$$

Proposition 1 outlines how Equation (20) quantifies the goodness of the network by Definition 12. Next, we discuss the powers attributed to l and k in (20). We note that the dimension of l^2 in the denominator is the same of that of M, that is, km^2. The nominator highlights the number of gravitating zones, comparing it to the denominator's value degrees. Namely, l^2 is in degree two, but M from (18) is in total degree three since it is obtained by multiplying demand with square distance. In accordance with Definition 12, Equation (20) defines a correct goodness function in three variables: k, M, and l.

The partial derivatives of Equation (20) are given by

$$\frac{\partial g_1(k, M, l)}{\partial k} = \frac{2k + 3k^2}{l^2 + M} = \frac{2 + 3k}{k + k^3} g_1(k, M, l), \tag{21}$$

$$\frac{\partial g_1(k, M, l)}{\partial M} = -\frac{k^2 + k^3}{(l^2 + M)^2} = \frac{-1}{l^2 + M} g_1(k, M, l), \tag{22}$$

$$\frac{\partial g_1(k, M, l)}{\partial l} = -2l\frac{k^2 + k^3}{(l^2 + M)^2} = \frac{-2l}{l^2 + M} g_1(k, M, l). \tag{23}$$

Because the network length is usually longer than half a kilometer, from Equations (21)–(23), the next inequality holds:

$$\frac{\partial g_1(k, M, l)}{\partial k} > 0 > \frac{\partial g_1(k, M, l)}{\partial M} > \frac{\partial g_1(k, M, l)}{\partial l}. \tag{24}$$

Proposition 2. *Given k, M, and l from Proposition 1, the network goodness function can be expressed using the following equation:*

$$g_2(k, M, l) = \frac{k^2}{l^2} + \frac{k^3}{M}. \tag{25}$$

In the following, we offer the partial derivatives of the function in (25):

$$\frac{\partial g_2(k, M, l)}{\partial k} = \frac{2k}{l^2} + \frac{3k^2}{M}$$

$$\frac{\partial g_2(k, M, l)}{\partial M} = -\frac{k^3}{M^2}$$

$$\frac{\partial g_2(k, M, l)}{\partial l} = -\frac{2k^2}{l^3}.$$

These lead to the following inequalities:

$$\frac{\partial g}{\partial k} \geq 0 \geq \frac{\partial g}{\partial l}, \frac{\partial g}{\partial M}. \tag{26}$$

Proposition 3. *We let k, M, and l be given as in the previous propositions. We let d and D be positive values from Definition 9 and let $Q(\mathcal{T}_i)$ be calculated as in Theorem 4.*

In the given environment, n represents the total number of demand zones that the network aims to serve. Then, the network goodness function is defined by equation

$$g(k, M, l) = \frac{k^2 \cdot \mu(\mathcal{U})}{2 \cdot n^2 \cdot D \cdot l} + \frac{k^3 \cdot d^2 \cdot \sum_{i=1}^{k} Q(\mathcal{T}_i)}{2 \cdot n^3 \cdot M}. \tag{27}$$

Equation (27) is dimensionless. This is particularly valuable when all demand zones gravitate attractively toward the network in the sense of Definition 9. Taking into account the values of n, d, D and the total demand value $\sum_{i=1}^{n} Q(\mathcal{T}_i)$, the following result validates the use of Equation (25) in the network goodness function.

Corollary 1. *If $n = k$, $\mu(\mathcal{U}) = D \cdot l$, and $d_i = d$ for each $\mathcal{T}_i \subset \mathcal{U}$, then the value calculated by Equation (27) is equal to one, that is,* $1 = 100\%$.

Proof of Corollary 1. If the distances of all the zones to the network are $d_i = d$, and if all zones gravitate to the network with $k = n$, then (18) offers $M = d^2 \cdot \sum_{i=1}^{n} Q(\mathcal{T}_i)$. If $d_i \leq d$, Equation (27) yields a result that is greater than 100%, and the network is too good or overcapacitated. □

If conditions in Corollary 1 are fulfilled, then we say that the network is available with distance D and attractive with distance d. Next, we prove all the propositions.

The proofs of Propositions 1–3. In the sense of Definition 12, the proofs of all the propositions are the same since the basis requests are satisfied. Following (26), all values g_1, g_2, and g increase with k and decrease with l and M, as required. □

In Equation (27), the power in n^2 is used because in our Example section $n^2 \geq \mu(\mathcal{U})$ when $n \leq \mu(\mathcal{U})$. Cube n^3 is used because in our example, $n^3 \geq \sum_{i=1}^{n} Q(\mathcal{T}_i)$ when $n^2 \leq \sum_{i=1}^{n} Q(\mathcal{T}_i)$.

To avoid confusion that may arise from the appearance of square and cube exponents in Equations from (20) to (27), we add the following theorem that uses ceiling function $\lceil x \rceil$ of an actual number x to determine the lowest integer greater than or equal to x.

Theorem 6. *We let values k, l, M be as in Proposition 1 and let n, d, D and $Q(\mathcal{T}_i)$ be as in Proposition 3. We define α and β as follows:*

$$\alpha = \lceil \log_n \mu(\mathcal{U}) \rceil \text{ and } \beta = \lceil \log_n \sum_{i=1}^{n} Q(\mathcal{T}_i) \rceil. \tag{28}$$

If $\alpha, \beta > 0$, then the value of function

$$\mathcal{G}(k, M, l) = \frac{k^\alpha \cdot \mu(\mathcal{U})}{2 \cdot n^\alpha \cdot D \cdot l} + \frac{k^\beta \cdot d^2 \cdot \sum_{i=1}^{n} Q(\mathcal{T}_i)}{2 \cdot n^\beta \cdot M} \tag{29}$$

depends increasingly upon k and decreasingly upon l and M. Depending on arbitrary values for d and D, $\mathcal{G}$ may be lesser or greater than 100%.

Proof. Again, partial derivatives $\frac{\partial \mathcal{G}}{\partial k} > 0 > \frac{\partial \mathcal{G}}{\partial l}, \frac{\partial \mathcal{G}}{\partial M}$. □

Remark 1. *Formula (27) is a special case of (29) when $\alpha = 2$ and $\beta = 3$.*

In the next section, we propose a Public City Transport (PCT) corridor network. The method combines environmental physics and social observations, behavior statistics, and heuristics, and applies the proposed theory to evaluate several network variants.

4. Example

Urban transport is a crucial pillar supporting citizen mobility in cities experiencing population growth and territorial expansion. Private transport tends to congest urban traffic until it cannot connect locations within the urban area.

It is necessary to invest in new modes of public transport in order to attract inhabitants to use it thanks to its availability, speed, and capacity. New modes need new corridors. Below is an example of the proposed procedure for the heuristic design and evaluation of public city transport corridor networks proposals. In the following, we demonstrate the process step by step.

The initial aim was to propose a network for the public transportation mode with large capacity and velocity. Usually, networks are embedded using the spanning tree or the least square method, but we proposed a network after several heuristically drawn variants and chose the best one. In the previous section, we proposed several functions for calculating the goodness of a network embedded in an environment for serving the users' traffic demands. In this section, we applied them to compare several network variants obtained heuristically following exhausting statistical investigations.

The primary prerequisite was to consider the natural obstacles in the environment. In Figure 4, four red barriers in the City of Zagreb are depicted, from north to south: the mountain, the railway, the river, and the highway.



Figure 4. Zagreb City County natural environment.

The zone partition was adopted from the former administrative local communities division and is shown in Figure 5. Demand points are buildings, schools, supermarkets, and other places where people can appear. Demand is the number of inhabitants or users taken from various reports. Demand centers are calculated by (11) after the coordinate system was established, according to Lemma 1. It is also necessary to calculate the moment inertia.

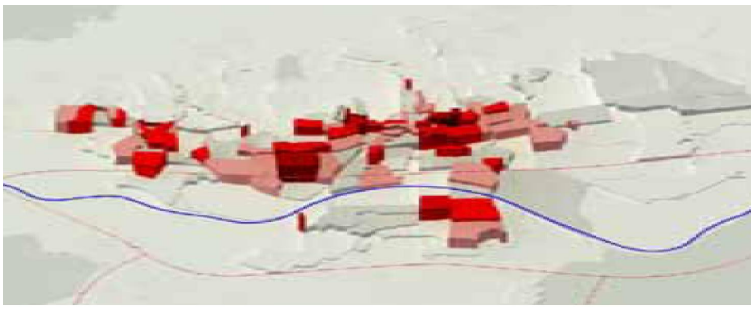


Figure 5. Zagreb City local population centers. The darker the red color the larger the population at that location. Source: ZagrebPlan [15].

Once the environment was partitioned (see Figure 5), and taking into consideration the barriers shown in Figure 4, we sketched corridors composed of segments (see Equation (1)). The segments were embedded according to the main traffic flow directions. Flow directions were obtained after statistical monitoring and are presented in Figure 6.

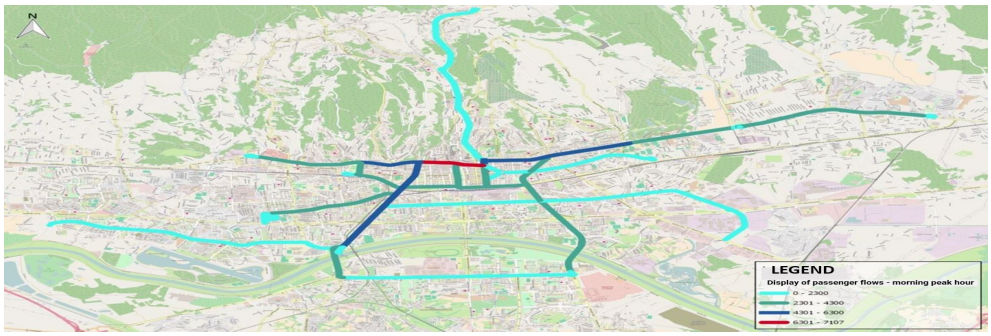


Figure 6. Zagreb City passenger flows.

The heuristic design process was guided through the nine corridor network variants presented in Figures 7–9. The network inertia for each of them was calculated according to Equations (17) and (18).

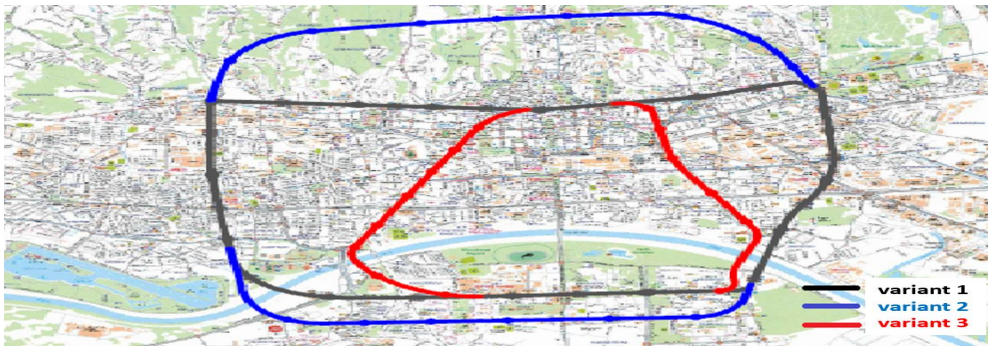


Figure 7. The first proposed corridor network variants.

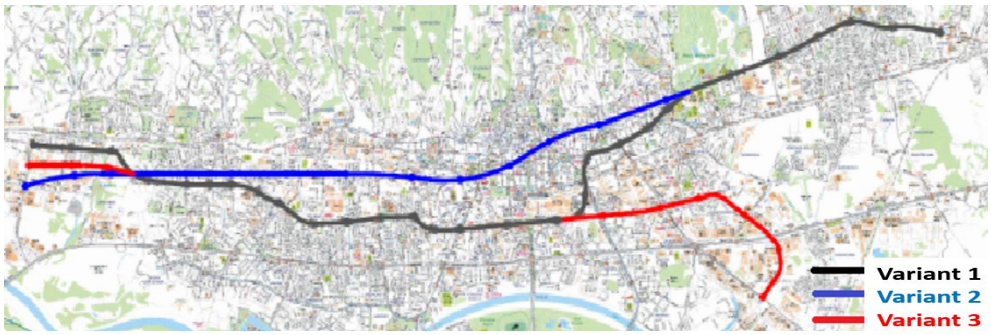


Figure 8. The second proposed corridor network variants.

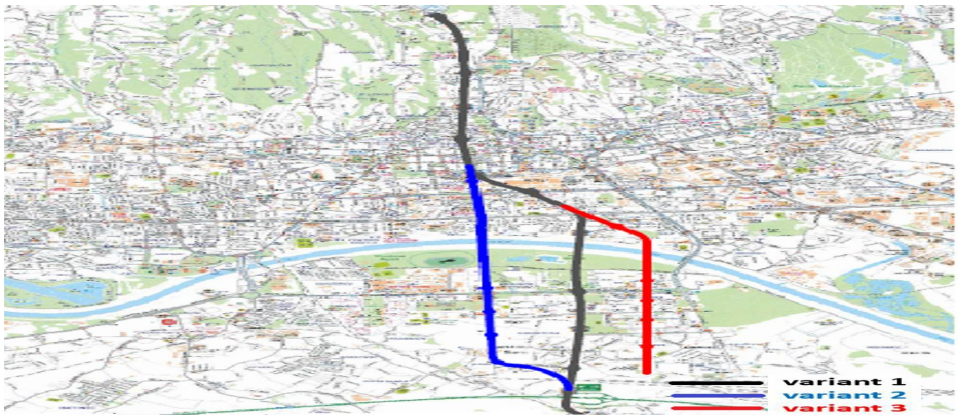


Figure 9. The third proposed corridor network.

The variants of the corridor network in Figure 7 are drawn in a circle, relying on the flows from Figure 6. Only zones up to $D = 4$ km are observed as the gravitating zones for inertia calculation.

The variants in Figure 9 are drawn according to the north–south directions in Figure 6. New, modified corridors imply consideration of Figure 5.

The corridors in Figure 8 stretch in the west–east direction following the flows from Figure 6. Then, these directions are connected with lines in the north–south direction. The process continues by considering the west–south direction in Figure 6. A good public traffic network attracts citizens to use the public traffic network rather than private transport devices if it is created to reduce the amount of walking necessary, as expressed by network inertia (see Equation (18)).

The formulas proposed in Propositions 1–3 were used to quantitatively compare different corridor networks, considering network availability through the number of gravitating demand zones and network costs roughly proportional to network length.

In the city of Zagreb, the environment area size is $\mu(\mathcal{U}) = 250$ km^2, the traffic demand is $\sum_{i=1}^{n} Q(\mathcal{T}_i) = 680{,}000$, and the total environment parts number is $n = 160$. Values $d = 200$ m and $D = 4$ km were taken following the recommendations in [14].

In Figures 7–9, three kinds of variants are proposed, according to the passengers' flow directions from Figure 6. In Table 1, goodness values for each variant are listed, and the best variant's goodness values are considered. Then, we collect the final network in Figure 10 by taking the first variants from Figures 7 and 8, and the third variant from Figure 9.

Table 1. Network goodness of the networks using Equations (20), (25) and (27).

Figure	Variant	M (km^2)	l (km)	k	g_1/km^2	g_2/km^2	g/%
Figure 7	1	27,000	22.3	47	3.86	8.29	13.33%
	2	17,000	27.5	36	2.70	4.46	6.65%
	3	9000	16.4	21	1.05	2.67	3.61%
Figure 8	1	17,000	19.6	36	1.05	6.12	8.95%
	2	13,000	19.2	33	2.76	5.72	7.82%
	3	11,000	18.7	28	2.00	4.24	5.76%
Figure 9	1	13,000	11.8	21	2.77	3.88	4.78%
	2	4000	13.7	18	2.00	3.18	3.36%
	3	10,000	9.8	20	0.83	4.96	5.23%
Figure 10	final	58,000	53.7	213	159	182	81.52%
Figure 11	tram	25,000	116	93	21	107	19.80%

For comparison, we calculated values for the temporary Zagreb Electrician Tramway network given in Figure 11.

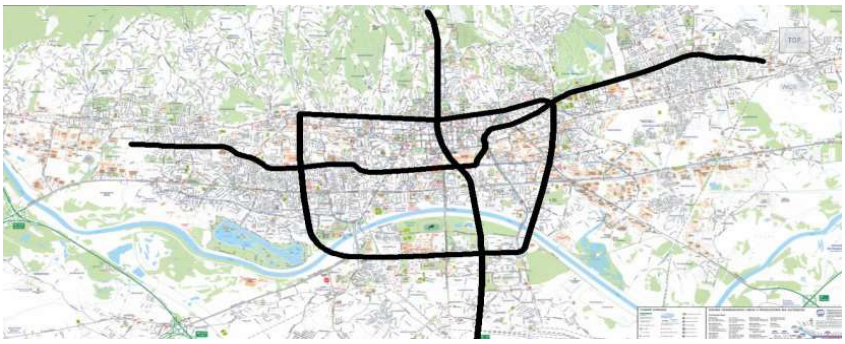


Figure 10. The final proposed corridor network variant.

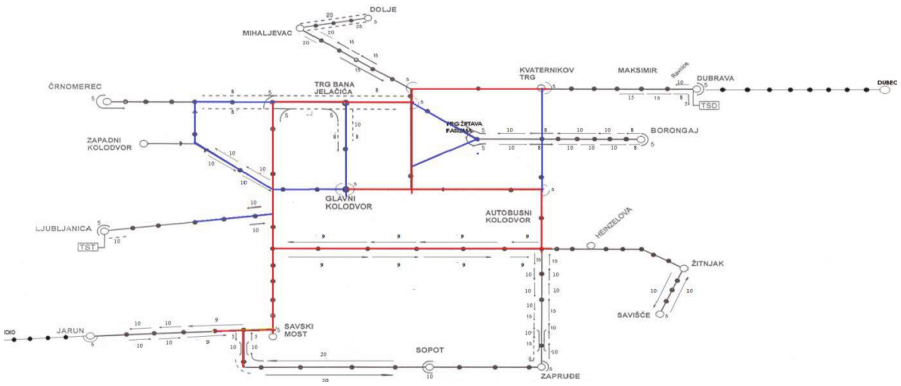


Figure 11. ZET network that is existing. Source: Time table of the ZET branch (2020).

5. Conclusions

We offer mathematical expressions that quantify the goodness of different networks based on (i) coverage, (ii) availability, and (iii) maintenance expenses. Although the model is developed heuristically, we show it can be a reliable tool for assessing various networks suggested for embedding in an environment.

The ongoing discussion on using heuristics to create mathematical formulas involves using rules of thumb, intuition, and common sense to find a solution that is good enough and rapid. Heuristics from experience and intuition are sometimes better than exact methods or exhaustive simulation procedures. Moreover, if there is nothing similar in the literature available to evaluate a network quantitatively with a unique value, then the heuristic approach offers a good solution.

However, it is crucial to evaluate the results of heuristics. The first issue we tackle is selecting the best network variant among heuristically obtained variants. We propose several variants to solve the problem in Section 4 and justify our approach for choosing the best option.

Finally, we plan to develop this work further in the future.

Author Contributions: Conceptualization and validation, M.B.; methodology and formal analysis, B.I.; investigation, resources, and original draft preparation, M.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: Data sharing not applicable.

Acknowledgments: We thank the Krapina University of Applied Sciences and the Sami Shamoon College of Engineering for their moral support.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kanellopoulos, D.; Sharma, V.K.; Panagiotakopoulos, T.; Kameas, A. Networking Architectures and Protocols for IoT Applications in Smart Cities: Recent Developments and Perspectives. *Electronics* **2022**, *12*, 2490. [CrossRef]
2. Karakus, M.; Durresi, A. Quality of Service (QoS) in Software Defined Networking (SDN). *A Surv. J. Netw. Comput. Appl.* **2017**, *80*, 200–218. [CrossRef]
3. Smit, L.C.; Dikken, J.; Schuurmans, M.J.; de Wit, N.J.; Bleijenberg, N. Value of social network analysis for developing and evaluating complex healthcare interventions: A scoping review. *BMJ Open* **2020**, *10*, e039681. [CrossRef] [PubMed] [PubMed Central]
4. Li, J.; Harsh, D.; Hu, X.; Tang, J.; Yi, C.; Liu, H. Attributed Network Embedding for Learning in a Dynamic Environment. In Proceedings of the CIKM'17, Singapore, 6–10 November 2017; Session 2D: Network Embedding 2. [CrossRef]
5. Salkhi Khasraghi, G.; Volchenkov, D.; Nejat, A.; Hernandez, R. University Campus as a Complex Pedestrian Dynamic Network: A Case Study of Walkability Patterns at Texas Tech University. *Mathematics***2023**, *12*, 140. [CrossRef]
6. Bogdanova, T.V.; Evdokimov, K.A. Transport component of a comfortable urban environment. *Inov. Investments* **2021**, *2021*, 163–170.
7. Colon, C.; Hallegatte, S.; Rozenberg, J. Criticality analysis of a country's transport network via an agent-based supply chain model. *Nat. Sustain.* **2021**, *4*, 209–215. [CrossRef]
8. Guillot, M.; Furno, A.; Aghezzaf, E.; El Faouzi, N. Transport network downsizing based on optimal sub-network. *Commun. Transp. Res.* **2022**, *2*, 100079. [CrossRef]
9. Wang, S.; Wang, Z. Collaborative Development and Transportation Volume Regulation Strategy for an Urban Agglomeration. *Sustainability* **2022**, *15*, 14742. [CrossRef]
10. Kuo, Y.; Leung, J.M.; Yan, Y. Public transport for smart cities: Recent innovations and future challenges. *Eur. J. Oper. Res.* **2023**, *306*, 1001–1026. [CrossRef]
11. Nes, V.R.; Hamerslag, R.; Immers, R. Design of Public Transport Networks. *Transp. Res. Rec.* **1988**, *1202*, 74–83.
12. Wilson, R.J. *Introduction to Graph Theory*, 4th ed.; Longman Group Ltd.: Harlow, UK, 1996. Available online: https://www.maths.ed.ac.uk/~v1ranick/papers/wilsongraph.pdf (accessed on 20 February 2023).
13. Lang, S. Introduction to Linear Algebra. In *Springer Books on Elementary Mathematics*; Springer: Berlin/Heidelberg, Germany, 1986; ISBN 978–1-4612-7002-7. Available online: https://www.math.nagoya-u.ac.jp/~richard/teaching/f2014/Lin_alg_Lang.pdf (accessed on 15 April 2024).
14. Šimunović,LJ.; Ćosić, M. Osobitosti Pješačkoga Prometa. In *Nemotorizirani Promet*; Hrvoje Gold; Faculty of Traffic and Transport Sciences: Zagreb, Croatia, 2015; pp. 3–9.
15. Available online: https://www.zagreb.hr/UserDocsImages/arhiva/zagrebplan-ciljeviiprioritetirazvojado2020.pdf (accessed on 12 December 2020).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

AppliedMath

MDPI

Article

Long- and Medium-Term Financial Strategies on Equities Using Dynamic Bayesian Networks

Karl Lewis [†], **Mark Anthony Caruana** [†] **and David Paul Suda** [*,†]

Department of Statistics and Operations Research, Faculty of Science, University of Malta, 2080 Msida, Malta; karl.lewis.17@um.edu.mt (K.L.); mark.caruana@um.edu.mt (M.A.C.)

* Correspondence: david.suda@um.edu.mt; Tel.: +356-9986-9884

† These authors contributed equally to this work.

Abstract: Devising a financial trading strategy that allows for long-term gains is a very common problem in finance. This paper aims to formulate a mathematically rigorous framework for the problem and compare and contrast the results obtained. The main approach considered is based on Dynamic Bayesian Networks (DBNs). Within the DBN setting, a long-term as well as a short-term trading strategy are considered and applied on twelve equities obtained from developed and developing markets. It is concluded that both the long-term and the medium-term strategies proposed in this paper outperform the benchmark buy-and-hold (B&H) trading strategy. Despite the clear advantages of the former trading strategies, the limitations of this model are discussed along with possible improvements.

Keywords: finance; dynamic Bayesian networks; trading strategies; equities

1. Introduction

In this paper, Dynamic Bayesian Networks (DBNs) are used to study the problem of obtaining and testing a financial strategy whose return is higher than the buy and hold strategy for a given equity. A Bayesian Network (BN) is a graphical and compact representation of a joint probability density function (PDF) that makes use of conditional independence and can be used to model a system under one time instance. DBNs extend the BN to more than one time period, which is to say that DBNs are a temporally driven extension of BNs.

Some benefits of using DBNs over other models, such as basic time series models, are the following. Firstly, DBNs have the capacity to incorporate what are known as hidden (or latent) variables, which either govern, or are thought to govern, the observations of the observable random variable. These hidden variables are random variables whose true value cannot be measured directly, sometimes due to instrumental inadequacies, and other times simply because the existence of these variables is hypothesized. Examples of such hidden variables are the intelligence of an individual (which cannot be measured precisely using any instrument) or the state (bull or bear) of a financial market (whose existence itself is hypothetical). Secondly, since they are an offshoot of BNs, DBNs allow for compact representations of an otherwise possibly cumbersome joint distribution. Russell and Norvig [1] argue that due to conditional independence, representing a joint distribution as a DBN may eliminate any redundant terms, which, in turn, also makes parameter estimation simpler. Murphy [2] state that another benefit of DBNs is the ease with which variations to the model at hand can be introduced. DBNs can be used to alternatively represent both basic Hidden Markov Models (HMMs), and HMMs with variations, such as HMMs with a mixture-of-Gaussian outputs, Auto-regressive HMMs, Input–Output HMMs and Hierarchical HMMs. DBNs can also be used to represent Kalman Filter Models where, contrary to HMMs, the hidden state is continuous rather than discrete. Lastly, DBNs also have the capacity to answer questions about different types of reasoning or inference, such

check for updates

Citation: Lewis, K.; Caruana, M.A.; Suda D. Long- and Medium-Term Financial Strategies on Equities Using Dynamic Bayesian Networks. *AppliedMath* **2024**, *4*, 843–855. https://doi.org/10.3390/appliedmath4030045

Academic Editor: Tommi Sottinen

Received: 28 April 2024
Revised: 21 June 2024
Accepted: 24 June 2024
Published: 3 July 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

as *diagnostic* inference (from effects to causes), *causal* inference (from causes to effects), *intercausal* inference (between causes of the same effect), and mixed inference (a mixture of any two of the above types of inference).

In this research paper, we use the price–earnings (PE) ratio in the models discussed. In the literature, various papers can be found which apply the said PE ratio to examine investment strategies. Most notably, one can mention Basu [3]. This ratio has also been used in a vast range of literature including that of Lleo [4] and Angelini [5]. Chang and Tian [6] also use BNs to model the qualitative and quantitative relationships between several variables that affect the dynamics of the S&P 500 stock index, with the aim of optimizing trading decisions, namely when to open a short position and when to invest in a long position. Damiano et al. [7] base their work on a previous study, that of Tayal [8], which employs HHMMs to analyze financial data. The results obtained by Tayal are reproduced in Damiano et al.'s paper, and further insight is given. Damiano et al. [7] conclude that probabilistic inference allows the identification of two distinct states in high-frequency data that are mainly marked by buying and selling pressure.

Historically, Paul Dagum is thought to have kick-started the development of DBNs in the early 1990s, namely, in the work by Dagum et al. [9] in 1992 and by Dagum et al. [10] in 1995. Ever since, DBNs have been used in various areas, namely, in finance and economics. However, other areas include (but are not limited to) speech recognition (Murphy [2], Zweig, Russell [11], and Nefian et al. [12]), biology(Yao et al. [13], Raval et al. [14], Murphy and Mian [15]), robotics (Prembida et al. [16], Patel et al. [17]), image processing (Delage et al. [18]), and fault detection (Cozar et al. [19]). Parameter estimation techniques of DBNs are discussed by Benhamou et al. [20], whereas Murphy [21] works on the theoretical foundations of DBNs.

The rest of this paper is structured as follows. Section 2 introduces the theoretical underpinnings of DBNs. In Section 3, DBNs are applied to twelve equities from both developed and developing markets, and the return on investment is presented. The reason for considering both developed and developing markets is to test model robustness under different economic conditions and growth patterns. Finally, in the concluding section, the key points and results obtained are highlighted.

2. Theoretical Framework

This model is based on the work by Wang [22] and exploits two behavioral finance phenomena—*behavioral volatility* and *mean reversion*. Behavioral volatility refers to the phenomenon whereby market-related events such as irrational trades executed by inexperienced traders cause the trading price of a stock to deviate away from its 'true' value. Events like these continually affect market price either until the effects caused cancel each other out, or until rational investors balance these effects out with their rational trades. This phenomenon is known as *mean reversion*. The theory of mean reversion is applicable not only to the price of a stock, but to any price-related metric, such as the price–earnings (PE) ratio (share price divided by the earnings per share). As a result, one can decide to buy (or sell) a stock only when some metric that follows the mean reversion theory is far away from the mean, knowing that the 'true' value of the metric will be returned to at some point in the future.

A *Directed Graphical Model* (DGM) is a representation of a probabilistic model that uses conditional independence assumptions through a directed graph, where the vertices (note that the terms "state", "random variable", "node" and "vertex" may be used interchangeably) of the graph represent random variables and the arcs represent conditional dependence between pairs of random variables. An arc from vertex A to vertex B represents the statement "A causes B" (Murphy [2]). DGMs are convenient due to their compactness in exhibiting some joint probability distribution—for N binary random variables, the general closed form of the joint probability distribution of these random variables may need $O(2^N)$ parameters, whereas the graphical model may give the same information using fewer parameters due to the omission of terms via conditional independence statements.

A *Bayesian Network* is a DGM that represents a set of random variables $X_1, ..., X_n$ and their conditional dependencies through a directed *acyclic* graph (DAG), denoted by $G = (\mathcal{V}, \mathcal{E})$. Each vertex in $\mathcal{V}$ represents a random variable X_i, and each directed edge in $\mathcal{E}$ represents the conditional dependence a random variable has on another. A conditional probability distribution is associated to each node X_i, and the joint probability distribution on the vertex set $\mathcal{V}$ is given by $\mathbb{P}[\mathcal{V}] = \prod_{i=1}^{n} \mathbb{P}[X_i \mid parents(X_i)]$. The generalization of a BN to multiple time slices gives rise to the definition of a DBN. A *Dynamic Bayesian Network* is defined to be a pair $(B_0, B_{\rightarrow})$, where B_0 is the BN representing the *prior probability* (or the initial state probability), that is, the probability distribution of the random variables at time 0, and $B_{\rightarrow}$ is a two-slice temporal BN which describes *transition probabilities* from time $t-1$ to time t for any node X in the vertex set $\mathcal{V}$ of the DAG $G = (\mathcal{V}, \mathcal{E})$, denoted by $\mathbb{P}[X_t|X_{t-1}]$. The joint probability distribution of the vertex set $\mathcal{V}$ over all time slices is given by:

$$\mathbb{P}[\mathcal{V}] = \mathbb{P}[\mathcal{V}_0]\,\mathbb{P}[\mathcal{V}_{1:T}] = \prod_{X_i \in \mathcal{V}} \mathbb{P}[X_{i_0} \mid parents(X_{i_0})] \prod_{t=1}^{T} \prod_{X_i \in \mathcal{V}} \mathbb{P}[X_{i_t} \mid parents(X_{i_t})],$$

where $\mathcal{V}_{t_1:t_2}$ refers to the set of all vertices indexed from time t_1 up to time t_2. A DBN can be parametrized by its transition matrix $\mathbb{A}_t = \mathbb{P}[X_t = j \mid X_{t-1} = i]$, and its prior distribution $\pi(i) = \mathbb{P}[X_1 = i]$. If the transition probabilities are assumed to be constant for all time slices ($\mathbb{A}_t = \mathbb{A}$), then they are said to be *homogeneous* and have a much more compact joint distribution function. When representing a DBN pictorially, two or three time slices are typically shown—the initial time slice and the subsequent one or two—since its structure is assumed to replicate throughout time.

An important property of DAGs is that "nodes can be ordered such that parents come before children. This is called a *topological ordering*, and it can be constructed from any DAG." (Murphy [23]). Given such an ordering, the *Ordered Markov property* (or *Local Markov property*) is defined to be the "assumption that a node only depends on its parents, not on all its predecessors in the ordering" (Murphy [23]). This assumption is in fact a generalization of the Markov property for Markov chains.

2.1. Inference for DBNs

The objective in inference is to infer the value of the latent states given the observations of the observable states, that is, inferring the marginals $\mathbb{P}[X_t = i \mid y_{1:\tau}]$. If $\tau = t$, the process is known as *filtering* (or *'now-casting'*); if $\tau > t$, then this is *smoothing*; and if $\tau < t$, then one would be performing *prediction* (or *forecasting*). A commonly used inference algorithm for DBNs is the *forward–backward algorithm*. In this algorithm, dynamic programming is implemented through two steps, known as *passes*, that run in a counter-directional manner—one runs forward in time, whilst the other runs backward. Note that it is assumed that the transition probability matrix, emission probability matrix, and prior probabilities are all known. In the *forward pass*, the value of $\alpha_t(i) := \mathbb{P}[X_t = i \mid y_{1:t}]$ is found in a recursive manner. In the *backward pass*, the value of $\beta_t(i) := \mathbb{P}[y_{t+1:T} \mid X_t = i]$ is also found recursively but moving in counter chronological order (from time T to time 2). After the forward and backward passes are complete, the value for $\gamma_t(i) := \mathbb{P}[X_t = i \mid y_{1:T}]$ can finally be obtained:

$$\gamma_t(i) = \mathbb{P}[X_t = i \mid y_{1:T}] = \frac{\mathbb{P}[y_{t+1:T} \mid X_t = i]\mathbb{P}[X_t = i \mid y_{1:t}]}{\mathbb{P}[y_{1:T}]}$$

where $\mathbb{P}[y_{1:T}] = \prod_{t=1}^{T} c_t$.

2.2. Learning DBNs

In this context, learning refers to the parameter estimation process. In parameter estimation, learning can be tackled either through a Maximum Likelihood (ML) approach or through the Maximum A Posteriori (MAP) approach. If using ML, then the data are used to obtain parameter estimates via the solution of the optimization problem:

$\theta^*_{ML} = \arg\max_\theta \Big\{\mathbb{P}[Y \mid \theta]\Big\} = \arg\max_\theta \Big\{\log \mathbb{P}[Y \mid \theta]\Big\}$, where θ is the set of parameters to be estimated. Typically, θ contains the transition matrix and parameters pertaining to the probability distribution used in the emission matrix. On the other hand, if using the MAP method, the optimization problem above is adjusted slightly to become $\theta^*_{MAP} = \arg\max_\theta \Big\{\log \mathbb{P}[Y \mid \theta] + \log \mathbb{P}[\theta]\Big\}$, where $\mathbb{P}[\theta]$ is the parameter prior distribution. The approach to solving the optimization problems above is through an adaptation of the Expectation–Maximization (EM) algorithm known as the *Baum–Welch* algorithm, proven to give a local optimum to the optimization problems above (Baum et al. [24], Dempster et al. [25]). It uses the forward–backward algorithm as a subroutine. Hence, in cases where the model parameters are unknown, the Baum–Welch algorithm is first used to estimate (or learn) the model parameters. Then, the forward–backward algorithm is used to infer the posterior marginals.

2.3. Application of Theory

The main hypothesis of the model is that the stock price of a firm is not always equal to the firm's 'true' intrinsic value. Utilizing the phenomena of behavioral volatility and mean reversion, temporary effects that cause stock metrics to deviate from their 'true' value are classified into two: short-term effects (length of a few days) and medium-term effects (length of several weeks).

It is hypothesized that the fundamental value of a company i at time t, denoted by $P^*_{i,t}$, is directly proportional to its annual earnings at time t, denoted by $E_{i,t}$, with the *fundamental PE ratio*, denoted by $PE^*_{i,t}$, acting as the constant of proportionality:

$$P^*_{i,t} = (PE^*_{i,t})(E_{i,t}).$$

On the other hand, the *observed PE ratio* (openly available on the public domain), denoted by $PE_{i,t}$, is given by $PE_{i,t} = \frac{P_{i,t}}{E_{i,t}}$ where $P_{i,t}$ denotes the actual trading price of a company i at time t.

Note that from this point onward, the index i is dropped, as the analysis on stocks is performed univariately. The model equation that results from the above is given by:

$$\begin{aligned} P_t = P^*_t(1+Z_t)(1+\varepsilon_t) \quad &\Longrightarrow \quad y_t = \ln\Big[PE^*(1+Z_t)\Big] + \ln\Big[1+\varepsilon_t\Big] \\ &\Longrightarrow \quad y_t = \ln\Big[PE^*(1+Z_t)\Big] + \varepsilon_t \end{aligned} \tag{1}$$

where $y_t = \ln\left[\frac{P_t}{E_t}\right]$, $\{Z_t\}_{t\in\mathbb{N}}$ is a discrete-time Markov chain modeling the medium-term noise effects, and $\varepsilon_t \sim \mathcal{N}(0,\sigma^2)$ is a random variable modeling the short-term noise effects. Due to the above model equation, the model used is only applicable for firms that have positive earnings throughout the period under study. Furthermore, one must note that y_t is an observable quantity (since both P_t and E_t are); however, PE^* and Z_t are not. It is assumed that both Z_t and PE^* are discrete-valued; hence, $Z_t \in \{a_1,\ldots,a_M\}$ and $PE^* \in \{b_1,\ldots,b_N\}$ where M and N represent the number of possible latent states of Z_t and PE^*, respectively.

The conditional independence assumptions used in this model are presented next:

$$\forall\, t \geq 1, \forall\, i,j \geq 1,$$

$$Z_t \perp PE^* \mid \emptyset. \tag{2}$$

$$Z_i \perp Z_j \mid Z_k \qquad \exists\, k \in \{i+1,\ldots,j-1\} \tag{3}$$

$$y_i \perp Z_j \mid Z_k, PE^* \quad \exists\, k \in \{i,\ldots,j-1\} \text{ or } k \in \{j+1,\ldots,i\}. \tag{4}$$

$$y_i \perp y_j \mid Z_k, PE^* \quad \exists\, k \in \{i,\ldots,j\}. \tag{5}$$

For $t \geq 2$ and $r, s \in \{1, \ldots, M\}$, the matrix $\mathbb{W} := [w_{rs}]_{M \times M}$ is defined to be the **transition probability matrix**, where

$$w_{rs} := \mathbb{P}[Z_t = a_s | Z_{t-1} = a_r].$$

Note that $w_{rs} \in [0,1]$ and $\sum_{s=1}^{M} w_{rs} = 1$. For $t \geq 1$, $m \in \{1, \ldots, M\}$ and $n \in \{1, \ldots, N\}$, the matrix $\mathbb{D}_t := [d_{mn}(y_t)]_{M \times N}$ is defined to be the **emission probability matrix** at time t, where

$$d_{mn}(y_t) := \mathbb{P}[y_t \mid Z_t = a_m, PE^* = b_n].$$

where $d_{mn}(y_t) \sim \mathcal{N}(\ln[b_n(1 + a_m)], \sigma^2)$. For $m \in \{1, \ldots, M\}$ and $n \in \{1, \ldots, N\}$, the vectors $\boldsymbol{u} := (u_m)_{m \in \mathbb{N}}$ and $\boldsymbol{v} := (v_n)_{n \in \mathbb{N}}$ are defined to be the **initial probability vectors** that contain the initial probability distributions, where

$$u_m := \mathbb{P}[Z_1 = a_m]; \qquad v_n := \mathbb{P}[PE^* = b_n].$$

Note that $u_m, v_n \in [0,1]$, $\sum_{m=1}^{M} u_m = 1$ and $\sum_{n=1}^{N} v_n = 1$. These prior distributions serve to incorporate any expert knowledge that the researcher may have available. A graphical representation of the model described above can be found in Figure 1.

Having laid out the principles needed for inference, the aim of the analysis now makes itself clearer—that of inferring the value of PE^* so as to estimate the fundamental price of a stock and formulate a trading strategy based on this knowledge. Furthermore, inferring the value of Z_t is also useful, as it can be used to test an alternative trading strategy. In this context, the data used will be split into a training and a test set. The set of model parameters is given by $\theta = \{\mathbb{W}, \boldsymbol{u}, \boldsymbol{v}, \sigma^2\}$. The space of possible parameters is given by the set:

$$\Theta = \left\{ \theta : u_m, v_n, w_{rs} \in [0,1], \sum_{m=1}^{M} u_m = 1, \sum_{n=1}^{N} v_n = 1, \sum_{s=1}^{M} w_{rs} = 1, \sigma^2 > 0 \right\}. \tag{6}$$

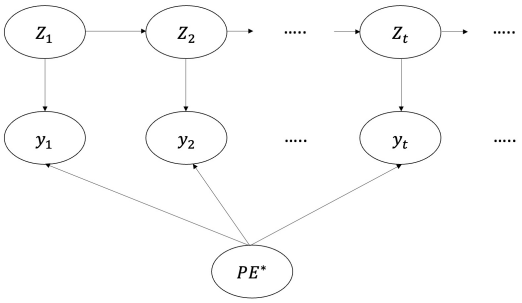


Figure 1. Representation of the model described above using a DBN, where y_t is observable and PE^* and Z_t are latent, discrete variables.

Learning and inference then follow. Since the parameters in θ are unknown, they will first need to be estimated (learning procedure). As mentioned, the algorithm used to obtain the MAP estimates is the Baum–Welch algorithm. After the unknown model parameters are estimated, the forward–backward algorithm is used on the parameter estimates to infer the constant value of PE^* for the training set and test set, and to infer the value of Z_t through smoothing for the training set, and through filtering for the test set:

- Filtering probabilities: $\mathbb{P}[Z_T, PE^* \mid y_{1:T}, \theta]$;
- Smoothing probabilities: $\mathbb{P}[Z_t, PE^* \mid y_{1:T}, \theta]$, where $t \in \{1, \ldots, T-1\}$.

2.4. Inference with Known Parameters

As per the forward pass of the forward–backward algorithm, the following definitions are made and equations derived for the filtering probabilities:

$$\forall\, t \in \{1, \ldots, T\}, m \in \{1, \ldots, M\}, n \in \{1, \ldots, N\},$$

$$\alpha_{tmn} \coloneqq \mathbb{P}[Z_t = a_m, PE^* = b_n \mid y_{1:t}, \theta]\,;$$

$$\alpha_{1mn} = \frac{d_{mn}(y_1)u_m v_n}{\sum_{m'=1}^{M}\sum_{n'=1}^{N} d_{m'n'}(y_1)u_{m'}v_{n'}}\,;$$

$$\alpha_{tmn} = \frac{d_{mn}(y_t)\sum_{i=1}^{M}\mathbb{P}[Z_{t-1} = a_i, PE^* = b_n \mid y_{1:t-1}]w_{mi}}{c_t}\,;$$

$$c_t \coloneqq \sum_{m=1}^{M}\sum_{n=1}^{N} d_{mn}(y_t)\sum_{i=1}^{M}\mathbb{P}[Z_{t-1} = a_i, PE^* = b_n \mid y_{1:t-1}]w_{mi}.$$

After defining the filtering probabilities, the smoothing probabilities are given through the estimate denoted by γ_{tmn}:

$$\forall\, t \in \{1,\ldots,T-1\}, m \in \{1,\ldots,M\}, n \in \{1,\ldots,N\},$$

$$\gamma_{tmn} \coloneqq \mathbb{P}[Z_t = a_m, PE^* = b_n \mid y_{1:T}].$$

Next, as per the backward pass of the forward–backward algorithm, the definition of β_{tmn} is to be given so as to be able to obtain the values for γ_{tmn}:

$$\beta_{tmn} \coloneqq \frac{\mathbb{P}[y_{t+1:T} \mid Z_t = a_m, PE^* = b_n]}{\mathbb{P}[y_{t+1:T} \mid y_{1:t}]} = \frac{\mathbb{P}[y_{t+1:T} \mid Z_t = a_m, PE^* = b_n]}{\prod_{t'=t+1}^{T} c_{t'}}$$

Therefore, as per the forward–backward algorithm, one has:

$$\gamma_{tmn} = \alpha_{tmn}\beta_{tmn}.$$

What remains to be derived are the expressions for $\beta_{tmn}\ \ \forall\, t \in \{1,\ldots,T-1\}$:

$$\beta_{(T-1)mn} = \frac{\beta'_{(T-1)mn}}{c_T} = \frac{\sum_{i=1}^{M} d_{in}(y_T)w_{im}}{\sum_{m=1}^{M}\sum_{n=1}^{N} c_T}.$$

For $\beta_{tmn}\ \ \forall\, t \in \{1,\ldots,T-2\}, m \in \{1,\ldots,M\}, n \in \{1,\ldots,N\}$:

$$\beta'_{tmn} = \sum_{i=1}^{M}\beta'_{(t+1)in}d_{in}(y_{t+1})w_{im}\,; \qquad \beta_{tmn} = \frac{\sum_{i=1}^{M}\beta_{(t+1)in}d_{in}(y_{t+1})w_{im}}{c_{t+1}}.$$

With expressions found for both the smoothing and filtering probabilities, the most probable values of the latent variables PE^* and Z_t are found through marginalization:

$$\widehat{PE^*} = \arg\max_{b_n}\left\{\sum_{m=1}^{M}\mathbb{P}[Z_t = a_m, PE^* = b_n \mid y_{1:T}]\right\} = \arg\max_{b_n}\left\{\sum_{m=1}^{M}\gamma_{tmn}\right\}$$

To find the estimate $\widehat{Z}_t$ for the latent state Z_t, smoothing is used on the training ,set whilst filtering is applied on the test set:

$$\widehat{Z}_t = \arg\max_{a_m}\left\{\sum_{n=1}^{N}\gamma_{tmn}\right\} \quad \forall\, t \in \{t\,:\, y_t \in \mathbb{X}_{train}\};$$

$$\widehat{Z}_t = \arg\max_{a_m}\left\{\sum_{n=1}^{N}\alpha_{tmn}\right\} \quad \forall\, t \in \{t\,:\, y_t \in \mathbb{X}_{test}\}.$$

where $\mathbb{X}_{train}$ and $\mathbb{X}_{test}$ denote the training set and test set, respectively.

2.5. Learning Unknown Parameters

The optimization problem in learning is given by $\widehat{\theta_{MAP}} = \arg\max_{\theta\in\Theta}\left\{\mathbb{P}\left[\theta \mid y_1^T\right]\right\}$ and is solved using the Baum–Welch algorithm (note that in the forthcoming expressions, θ should technically be written with a hat superscript (ˆ) since it is a set of parameter estimates):

1. Set $j = 1$.
2. Set $\theta^{(1)} - \theta^{(0)} > \delta$
3. **while** $\theta^{(j)} - \theta^{(j-1)} > \delta$ **do**

   (i) Calculate probabilities $\mathbb{P}\left[Z_t = a_m, PE^* = b_n \mid y_{1:T}, \theta^{(j)}\right]\ \forall\, t, m, n$

   (ii) Solve the constrained maximization problem

$$\widehat{\theta^{(j+1)}} = \arg\max_{\theta\in\Theta}\left\{Q\left(\theta;\theta^{(j)}\right) + \ln\mathbb{P}[\theta]\right]\}$$

   where $Q\left(\theta;\theta^{(j)}\right) = \mathbb{E}_{Z_{1:T},PE^*|y_{1:T},\theta^{(j)}}\left[\ln\mathbb{P}[y_{1:T}, Z_{1:T}, PE^* \mid \theta]\right]$.

   (iii) Increment j

An expression in closed form can be obtained for $Q\left(\theta;\theta^{(j)}\right)$ by using the smoothing probabilities:

$$\begin{aligned}
Q\left(\theta;\theta^{(j)}\right) = {} & \sum_{t=1}^{T-1}\sum_{m=1}^{M}\sum_{n=1}^{N}\alpha_{tmn}^{(j-1)}\beta_{tmn}^{(j-1)}\ln[d_{mn}(y_t)] + \sum_{m=1}^{M}\sum_{n=1}^{N}\alpha_{Tmn}^{(j-1)}\ln[d_{mn}(y_T)] \\
& + \sum_{t'=2}^{T-1}\sum_{i,m=1}^{M}\sum_{n=1}^{N}\beta_{tmn}^{(j-1)}\left(\frac{d_{mn}^{(j-1)}(y_t)}{c_t^{(j-1)}}\right)w_{mi}^{(j-1)}\alpha_{(t-1)in}^{(j-1)}\ln[w_{mi}] \\
& + \sum_{i,m=1}^{M}\sum_{n=1}^{N}\left(\frac{d_{mn}^{(j-1)}(y_T)}{c_T^{(j-1)}}\right)w_{mi}^{(j-1)}\alpha_{T-1,in}^{(j-1)}\ln[w_{mi}] \\
& + \sum_{m=1}^{M}\sum_{n=1}^{N}\alpha_{1mn}^{(j-1)}\beta_{1mn}^{(j-1)}(\ln[u_m] + \ln[v_n]).
\end{aligned}$$

In the argument of the maximization problem above, $\ln\mathbb{P}[\theta]$ is the logarithm of the prior distribution $\mathbb{P}[\theta]$, where $\mathbb{P}[\theta] = \mathbb{P}[\boldsymbol{u}]\mathbb{P}[\boldsymbol{v}]\mathbb{P}[\mathbb{W}]\mathbb{P}[\sigma^2]$ since the independence of priors is assumed. Within this prior distribution, two types of expert knowledge can be included—prior knowledge of the ballpark value of PE^* and prior knowledge of the persistence of the medium-term noise effects, which are encoded through the prior $\mathbb{P}[\boldsymbol{v}]$ and the prior $\mathbb{P}[\mathbb{W}]$, respectively.

The prior for the vector $\boldsymbol{v}$ is represented by the Dirichlet distribution:

$$f(\boldsymbol{v}) = \frac{\Gamma(k_1 + k_2 + \ldots + k_N)}{\Gamma(k_1)\Gamma(k_2)\ldots\Gamma(k_N)}\prod_{n=1}^{N} v_n^{k_n-1}. \tag{7}$$

The values k_n for $n \in \{1, \ldots, N\}$ intuitively correspond to the degree of belief an expert has on the event that b_n is the 'true' value for PE^*. For this analysis, $k_n = 1\ \forall\, n$. The prior for the matrix $\mathbb{W}$ is also derived from the Dirichlet distribution $f(\mathbb{W}) = \prod_{m=1}^{M} f(\boldsymbol{w}_m)$, where $\boldsymbol{w}_m = (w_{im})_{i=1,\ldots,M}$ and

$$f(\boldsymbol{w_m}) = \frac{\Gamma(k_{1m} + k_{2m} + \ldots + k_{Mm})}{\Gamma(k_{1m})\Gamma(k_{2m})\ldots\Gamma(k_{Mn})}\prod_{m=1}^{M} w_{im}^{k_{im}-1}.$$

Since $\mathbb{W}$ is the transition matrix for the hidden Markov chain $\{Z_t\}_{t\in\mathbb{N}}$, then the diagonal entries of $\mathbb{W}$ represent the probability that a particular state persists (stays as is in the next time point). The greater the value of Z_t (or, correspondingly, a_m), the greater that

state's persistence. In this analysis, the off-diagonal entries (the reader is suggested to refer to Wang [22] for more information on the values of the off-diagonals) are set to 0.

Taking all the above into consideration, the logarithm of the prior $\mathbb{P}[\theta]$ becomes

$$\ln \mathbb{P}[\theta] = \ln \mathbb{P}[\mathbb{W}] + \ln \mathbb{P}[v] = \sum_{m=1}^{M} \sum_{i=1}^{M} (k_{im} - 1) \ln[w_{im}] + \sum_{n=1}^{N} (k_n - 1) \ln[v_n] + s_1,$$

where $s_1 = \ln\left[\frac{\Gamma(k_1+k_2+\ldots+k_N)}{\Gamma(k_1)\Gamma(k_2)\ldots\Gamma(k_N)}\right] + \ln\left[\frac{\Gamma(k_{1m}+k_{2m}+\ldots+k_{Mm})}{\Gamma(k_{1m})\Gamma(k_{2m})\ldots\Gamma(k_{Mn})}\right]$.

The constant s_1 is only included for completeness' sake—it is rendered irrelevant when maximizing in the Baum–Welch algorithm.

With the priors set, the constrained optimization problem is now fully defined. Note that only the equality constraints in (6) will be considered, as the inequality constraints will end up being satisfied still. Therefore, the method of Lagrange multipliers can be used to solve this optimization problem. The expressions for the estimators of the four variables in question are given below:

$$\widehat{u_m^{(j)}} = \sum_{n=1}^{N} \alpha_{1mn}^{(j-1)} \beta_{1mn}^{(j-1)}; \tag{8}$$

$$\widehat{v_n^{(j)}} = \frac{\sum_{m=1}^{M} q_{1mn}^{(j)}(Z_1, PE^*) + (k_n - 1)}{1 + \sum_{n'=1}^{N}(k_{n'} - 1)} = \frac{\sum_{m=1}^{M} \alpha_{1mn}^{(j-1)} \beta_{1mn}^{(j-1)} + (k_n - 1)}{1 + \sum_{n'=1}^{N}(k_{n'} - 1)}; \tag{9}$$

$$\begin{aligned} \widehat{w_{mi}^{(j)}} &= -\frac{\sum_{t'=2}^{T}\sum_{n=1}^{N} q_{t',m,i,n}^{(j)}(Z_{t'}, Z_{t'-1}, PE^*) + (k_{mi} - 1)}{\sum_{t'=2}^{T}\sum_{m=1}^{M}\sum_{n=1}^{N} q_{t',m,i,n}^{(j)}(Z_{t'}, Z_{t'-1}, PE^*) + \sum_{m'=1}^{M}(k_{m'i} - 1)} \\ &= \frac{\omega_{t\prime,m,i,n}^{(j)}}{\sum_{m'=1}^{M} \omega_{t\prime,m,i,n}^{(j)}}, \end{aligned} \tag{10}$$

where

$$\begin{aligned} \omega_{t\prime,m,i,n}^{(j)} = &\sum_{t'=2}^{T-1}\sum_{n=1}^{N} \beta_{t'mn}^{(j-1)} \left(\frac{d_{mn}^{(j-1)}(y_{t'})}{c_{t'}^{(j-1)}}\right) w_{mi}^{(j-1)} \alpha_{t'-1,i,n}^{(j-1)} \\ &+ \sum_{n=1}^{N} \left(\frac{d_{mn}^{(j-1)}(y_T)}{c_T^{(j-1)}}\right) w_{mi}^{(j-1)} \alpha_{(T-1)in}^{(j-1)} + (k_{mi} - 1) \end{aligned}$$

$$\begin{aligned} \widehat{\sigma^{2(j-1)}} = &\frac{\sum_{t=1}^{T-1}\sum_{m=1}^{M}\sum_{n=1}^{N} \alpha_{tmn}^{(j-1)} \beta_{tmn}^{(j-1)} (y_t - \ln[b_n(1 + a_m)])^2}{\sum_{t=1}^{T-1}\sum_{m=1}^{M}\sum_{n=1}^{N} \alpha_{tmn}^{(j-1)} \beta_{tmn}^{(j-1)} + \sum_{m=1}^{M}\sum_{n=1}^{N} \alpha_{Tmn}^{(j-1)}} \\ &+ \frac{\sum_{m=1}^{M}\sum_{n=1}^{N} \alpha_{Tmn}^{(j-1)} (y_T - \ln[b_n(1 + a_m)])^2}{\sum_{t=1}^{T-1}\sum_{m=1}^{M}\sum_{n=1}^{N} \alpha_{tmn}^{(j-1)} \beta_{tmn}^{(j-1)} + \sum_{m=1}^{M}\sum_{n=1}^{N} \alpha_{Tmn}^{(j-1)}}. \end{aligned} \tag{11}$$

3. Methodology of Analysis and Results

Twelve equities were chosen to be analyzed in this paper (see Table 1)—nine from a developed market (US) and three from emerging markets (Brazil and China). The training set contains data from the 1st of January 2011 to the 31st of December 2019 whilst the test set contains data from the 1st of January 2020 to the 30st of September 2020. Using a 12-month rolling Sharpe ratio as a point of reference, all equities displayed average or above average returns (with respect to an S&P benchmark) in the test period, with the exception of the two underperforming Brazilian equities [26]. Two datasets were collected for each stock for the above-mentioned period—daily price data and quarterly earnings per share data.

Table 1. Table of results, showing the return on investment as a percentage of the sum invested. Values in brackets are negative.

Stock	Thresholds for LT Strategy				Thresholds for MT Strategy				Profit from B&H
	0.05	**0.10**	**0.15**	**0.20**	**0.03**	**0.05**	**0.07**	**0.10**	
AAPL (US)	55.30%	55.30%	55.30%	55.30%	69.51%	57.82%	48.06%	76.02%	55.30%
ADBE (US)	25.46%	37.86%	44.45%	0%	46.65%	46.65%	46.65%	46.65%	46.65%
BLK (US)	12.98%	12.98%	12.98%	12.98%	33.92%	34.66%	20.85%	20.85%	12.98%
COST (US)	22.59%	22.59%	22.59%	22.59%	50.12%	42.85%	28.21%	0%	22.59%
HD (US)	28.71%	28.71%	28.71%	28.71%	57.76%	51.31%	37.99%	47.35%	28.71%
ITUB (BZ)	(39.86%)	(29.36%)	(29.36%)	(27.57%)	9.10%	7.07%	7.07%	(9.59%)	(55.85%)
MA (US)	27.17%	29.34%	61.09%	60.41%	47.70%	32.57%	24.66%	34.39%	11.93%
MCD (US)	11.38%	11.38%	11.38%	11.38%	18.94%	18.94%	30.63%	30.63%	11.38%
NTES (CHN)	47.89%	50.84%	57.35%	39.54%	39.54%	39.54%	39.54%	39.54%	39.54%
NVDA (US)	125.90%	125.90%	125.90%	125.90%	149.00%	156.79%	125.90%	125.90%	125.90%
SAN (BZ)	(2.08%)	(14.35%)	(17.78%)	(17.78%)	(57.17%)	(57.17%)	(57.17%)	(57.17%)	(57.17%)
UNH (US)	7.95%	7.95%	15.90%	19.88%	72.41%	21.83%	38.58%	2.87%	7.95%
Average	26.12%	25.07%	32.38%	27.61%	44.79%	37.74%	32.60%	29.79%	20.83%
W/D/L	4/7/1	4/7/1	5/6/1	4/7/1	9/3/0	9/3/0	7/4/1	6/4/2	
Totals	17/27/4				31/14/3				

After the preliminary data cleaning and preparation phase, the initial values of vectors $\boldsymbol{a} = (a_m)$ and $\boldsymbol{b} = (b_n)$ are set next. Priors of the vector $\boldsymbol{v}$ and matrix $\mathbb{W}$ are set as discussed earlier; the prior for $\boldsymbol{u}$ is set to be the discrete uniform distribution, and the initial value of σ^2 is set to 5. Parameter learning is then performed before inference of the latent states $\widehat{PE^*}$ and $\widehat{Z}_t$. Simulation of the trading strategies follows. The trading strategies proposed by Wang [22] and used in this paper will be compared to the benchmark B&H strategy. For both the proposed trading strategies, let I_t denote the amount of cash available at time t; let N_t denote the units of a security held at time t; let T_{train} represent the size of the training set; and let T represent the size of the dataset (sum of sizes of training and test sets). Both the long-term and the medium-term trading strategies can be described by three possible courses of action (labeled (i), (ii) and (iii) below) at time t; that is, courses of action (i) through (ii) are common to both trading strategies:

(i) If $PE_t \leq A_t(1 - Tr)$ and $I_t > 0$, then buy the security using all the available cash I_t. Hence, $N_{t+1} = \frac{I_t}{P_T}$ and $I_{t+1} = 0$.

(ii) If $PE_t \geq A_t(1 + Tr)$ and $I_t = 0$, then sell all the units held N_t. Hence, $N_{t+1} = 0$ and $I_{t+1} = P_t N_t$.

(iii) If neither (i) or (ii) are satisfied, do not execute any trades. Hence, $N_{t+1} = N_t$ and $I_{t+1} = I_t$.

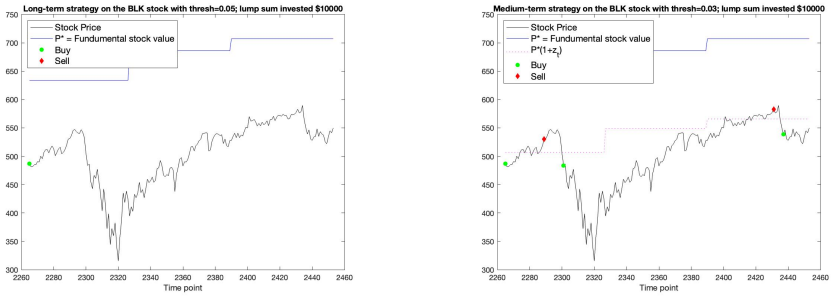
Note that A_t is considered to be a baseline and depends on the trading strategy. The threshold value $Tr \in (0, 1)$ acts as a sensitivity gauge defining how much the investor wants to allow $PE_t = \frac{P_t}{E_t}$ to deviate from the baseline A_t before triggering a particular course of action in the trading strategies. This is clear from how the courses of action are defined. The total profit at the end of the trading period is given by $I_T + P_T N_T - I_{T_{train}+1}$.

The first trading strategy is the so-called 'long-term strategy', where trading is performed with respect to the constant value of $\widehat{PE^*}$, so $A_t = \widehat{PE^*}$. The alternative strategy is the 'medium-term strategy', where trading is performed with respect to the dynamic values of $\widehat{PE^*}(1 + \widehat{Z}_t)$, where each $\widehat{Z}_t$ is dynamically estimated through filtering. Therefore, for the medium-term strategy, $A_t = \widehat{PE^*}(1 + \widehat{Z}_t)$.

The results on the BLK and ITUB stocks are presented in graphical detail in this paper. Firstly, the long-term strategy on the BLK stock data suggests that the investor buys the stock at time point $2265 = T_{train} + 1$ and holds the stock for the rest of the period. Clearly, this strategy coincides with the B&H strategy and, as a result, profit from the long-term-strategy would be equal to profit from the B&H strategy which is equal to USD1298.04, equaling a 12.98% return on the initial investment of USD 10,000. On the other hand, the

medium-term strategy suggests that the investor buys the stock at time points 2265, 2301 and 2437, sells it at time points 2289 and 2431, and holds it for the rest of the time points. This would yield a profit of USD 3391.85; equalling a 33.92% return on the initial investment of USD 10,000 in the nine-month period that the testing set covers. This means that the medium-term strategy beats the B&H strategy by 20.94%. Figure 2 illustrates the long-term strategy and medium-term strategy.

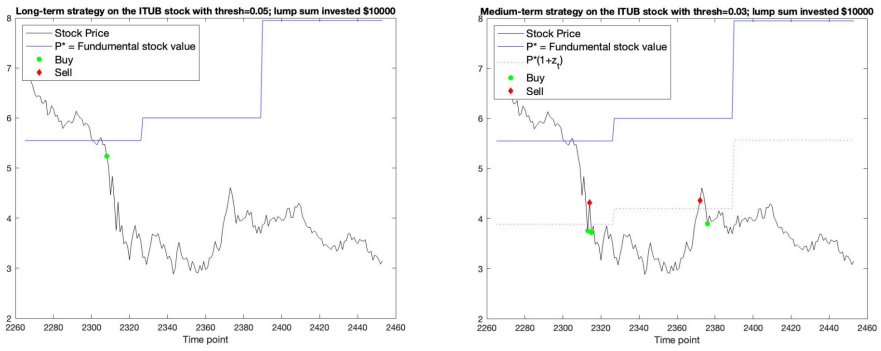
Next, we discuss the ITUB stock data. The long-term strategy suggests that the investor buys the stock at time point 2308 and holds the stock for the rest of the period. Implementing this strategy would yield a loss of USD 3986.16, which equates to a 39.86% loss on the initial investment of USD 10,000. The buy-and-hold strategy, however, would yield a greater loss of USD 5584.67. In contrast, the medium-term strategy suggests that the investor buys the stock at time points 2313, 2315 and 2376, sells it at time points 2314 and 2372, and holds it for the rest of the time points. This would actually yield a profit of USD 909.81. Although this is only a 9.1% return on the initial investment of USD 10,000, the medium-term strategy provides the investor with a strategy whereby he or she can make a profit in a period when the stock is actually crashing. Figure 3 illustrates the long-term strategy and medium-term strategy, respectively.



(**a**) Long-term strategy (**b**) Medium-term strategy

Figure 2. Simulated strategies for the BLK stock.

More generally, we see in Table 1 that the medium strategy has been consistently superior (for various thresholds) for BLK, COST, HD, ITUB, MA, MCD, NVDA and UNH. For ITUB, for all but one threshold, the strategy turns a slight profit even though a loss is registered for other strategies. The long-term strategy has been consistently superior for NTES and SAN. Neither strategy has offered any advantages on ADBE, while for AAPL, the success of the medium-term strategy depends on the choice of threshold. Some further analysis on these results will be given in the conclusion.



(**a**) Long-term strategy (**b**) Medium-term strategy

Figure 3. Simulated strategies for the IUTB stock.

4. Conclusions

Through the use of DBNs, the model for stock movement by Wang [22] is built for our existing equity dataset. This model includes two latent states—one modeling the medium-term noise effects, and the other modeling the true fundamental PE ratio of a firm, the latter assumed to be constant throughout the period under study. The forward–backward algorithm and the Baum–Welch algorithm (variant of the EM algorithm) are used to perform parameter learning and inference. Based on this fitted model, the long- and medium-term strategies are applied to the twelve stocks studied here, with nine of these stocks trading on a developed market and the rest on an emerging one. Overall, both the long-term and medium-term strategies outperform the benchmark B&H trading strategy 17 and 31 times, respectively, out of a total of 48 experiment runs for each strategy (four for each of the twelve stocks). The strategies proposed by Wang only lose out to the B&H four and three times, respectively. Furthermore, the outperformance of these trading strategies is substantial. Whereas the average profit over all stocks when using the B&H is 20.83%, the average profit over all stocks and over all thresholds for the long-term strategy is 27.78% and that for the medium-term strategy is 36.23%. Lastly, it results that these trading strategies provide the investor with trading suggestions that, in certain cases, can even turn a loss under the B&H strategy into a profit as was shown to be the case for the ITUB stock. This stock was on a downfall, but the medium-term strategy still yielded a profit on the sum invested. All these results are as displayed in Table 1.

As with any statistical model, the model implemented in this work has room for improvement. The first and most significant limitation is the fact that only stocks that had a positive EPS during the period under study can be modeled by the model. This is due to the left-hand side of the model equation, that is, Equation (1). In certain times such as during pandemics or during recessions, it may be considerably difficult to find a firm who has not registered a loss in at least one quarter for the period under study. A second limitation is the lack of use of expert knowledge. In real-world investing, expertise in the field is considered highly valuable, and only seasoned investors are generally advised to trade actively. Since the model in this paper allows for the incorporation of expert knowledge, it is indeed a limitation that such knowledge is not made use of. Another limitation related to expert knowledge is the limitation that no trading fees or commissions are taken into account in this analysis. Although the proposed model allows for commissions to be considered in the trading strategies, no knowledge on the actual values of fees or commissions charged was available at the time of writing, and hence, such expenses could not be taken into account. It is well known that certain trading fees can sometimes tally up when executing numerous trades in such a way that a profit can turn into a loss when these fees are brought to the fore, which would make the buy-and-hold strategy more profitable.

Improvements on the above-mentioned limitations can add value to the model and should be considered in future works. In addition, another improvement that makes the model here more applicable in real-life scenarios is the incorporation of some variable that measures the liquidity of a stock. This is because the suggestions provided by the trading strategies on when to buy and sell a stock are rendered useless if the stock itself is not liquid. Prices may change by the time the stock becomes liquid enough for an actual trading opportunity to arise, and if the price changes significantly, that obviously renders the suggestion itself useless. In this study, liquidity was not envisaged to be of concern since all 12 stocks considered are large-cap stocks.

The choice of historical data used is always an important decision due to the fact that the length of the time series plays a major role. In principle, longer time series are preferred to shorter ones, but if the historical data contain changes in regime, this may inhibit the model in its forecasting performance. Furthermore, future regime changes that may occur in the test period may also impact the model's forecasting performance.

Finally, other improvements that can be implemented in future studies are further and deeper experimentation and testing, and the possibility of short-selling stock. To begin with, experimentation at a portfolio level can be implemented with the stocks studied here

to understand how the profits change when a group of stocks are considered together—for instance, samples of nine stocks from the twelve studied here can be taken to form portfolios, and the 'optimal' portfolio can be identified. Apart from this, more time can be spent on cross-validation in future studies to further improve on the suggestions provided by the trading strategies in this work. Also, the properties of all the estimators used could be derived in future works, to better understand their behavior. Finally, this strategy does not allow the short-selling of stock; extensions to the model which allow for this could be proposed.

Author Contributions: All the three authors worked equally hard to write this paper. K.L. took care of the coding in MATLAB and the results section, M.A.C. and D.P.S. took care of the methodology and theoretical content of the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: All the data is openly available from Yahoo Finance at https://finance.yahoo.com/.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

BN	Bayesian Network
DAG	Directed Acyclic Graph
DBN	Dynamic Bayesian Network
DGM	Directed Graphical Model
EM	Expectation–Maximization
HNN	Hidden Markov Model
PDF	Probability Density Function
MAP	Maximum A Posteriori
ML	Maximum Likelihood

References

1. Russell, S.; Norvig, P. *Artificial Intelligence: A Modern Approach*, 2nd ed.; Prentice Hall: Upper Saddle River, NJ, USA, 1995.
2. Murphy, K. Dynamic Bayesian Networks: Representation, Inference and Learning. Ph.D. Dissertation, University of California, Berkeley, CA, USA, 2002.
3. Basu, S. Investment Performance of common stocks in relation to their price-earnings ratios: A Test of the efficient market hypothesis. *J. Financ.* **1977**, *23*, 3. 663–673.
4. Lleo, S.; Ziemba, W.T. A Tale of Two Indexes: Predicting Equity Market Downturns in China. 2020. [CrossRef]
5. Angelini, N.; Bormetti, G.; Marmi, S.; Nardini, F. Value Matters: Predictability of Stock Index Returns. *SSRN* **2012**. [CrossRef]
6. Chang, K.C.; Tian, Z. Market analysis and trading strategies with Bayesian networks. In Proceedings of the 18th International Conference on Information Fusion, FUSION 2015, Washington, DC, USA, 6–9 July 2015; pp. 1922–1929.
7. Damiano, L.; Peterson, B.; Weylandt, M. Regime Switching and Technical Trading with Dynamic Bayesian Networks in High-Frequency Stock Markets, 2017.
8. Tayal, A. *Regime Switching and Technical Trading with Dynamic Bayesian Networks in High-Frequency Stock Markets;* University of Waterloo: Waterloo, ON, Canada, 2009.
9. Dagum, P.; Galper, A.; Horvitz, E. Dynamic Network Models for Forecasting. In Proceedings of the Eighth Conference on Uncertainty in Artificial Intelligence, Stanford, CA, USA, 17–19 July 1992; pp. 41–48.
10. Dagum, P.; Galper, A.; Horvitz E. Uncertain Reasoning and Forecasting. *Int. J. Forecast.* **1995**, *11*, 73–87. [CrossRef]
11. Zweig, Z.; Russell S. Speech Recognition with Dynamic Bayesian Networks. 1998. Available online: http://www.cse.msu.edu/~cse841/papers/zweig.pdf (accessed on 28 April 2024).
12. Nefian, A.; Liang, L.; Pi, X.; Liu, X.; Murphy, K. Dynamic Bayesian Networks for Audio-Visual Speech Recognition. *EURASIP J. Adv. Signal Process.* **2002**, *11*, 783042. [CrossRef]
13. Yao, X.Q.; Zhu, H.; She, Z. A dynamic Bayesian network approach to protein secondary structure prediction. *BMC Bioinform.* **2008**, *9*, 49. [CrossRef] [PubMed]
14. Raval, A.; Ghahramani, Z.; Wild, D.L. A Bayesian network model for protein fold and remote homologue recognition. *Bioinformatics* **2002**, *18*, 788–801. [CrossRef] [PubMed]

15. Murphy, K.; Mian, S. Modelling Gene Expression Data Using Dynamic Bayesian Networks. 1996. Available online: https://www.cs.ubc.ca/~murphyk/Papers/ismb99.pdf (accessed on 28 April 2024).
16. Prembida, C.; Souza, F.A.A.; Faria, D.R. Dynamic Bayesian network for semantic place classification in mobile robotics. *Auton. Robot.* **2016**, *5*, 1161–1172. [CrossRef]
17. Patel, M.; Miro, J.V. ; Dissanayake, G. Dynamic Bayesian Networks for Learning Interactions Between Assistive Robotic Walker and Human Users. 2010. Available online: https://opus.lib.uts.edu.au/bitstream/10453/16294/1/2010000382.pdf (accessed on 28 April 2024).
18. Delage, E.; Lee, H.A.Y. A Dynamic Bayesian Network Model for Autonomous 3D Reconstruction from a Single Indoor Image. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), New York, NY, USA, 17–22 June 2006; Volume 2, pp. 2418–2428. [CrossRef]
19. Cozar, J. ; Puerta J. ;Gamez, J.A. An Application of Dynamic Bayesian Networks to Condition Monitoring and Fault Prediction in a Sensored System: A Case Study. *Int. J. Comput. Intell. Syst.* **2017**, *10* , 176. [CrossRef]
20. Benhamou, E.; Atif, J.; Laraki, R. A New Approach to Learning in Dynamic Bayesian Networks (DBNs). *SSRN Electron. J.* **2018**. [CrossRef]
21. Murphy, K. An introduction to Graphical Models. 2001. Available online: https://www.cs.ubc.ca/~murphyk/Bayes/bnintro.html (accessed on 28 April 2024).
22. Wang, H. Bayesian Probabilistic Inference on Firm-Level Stock Price Dynamics. Ph.D. Dissertation, Suranaree University of Technology, Nakhon Ratchasima, Thailand, 2016.
23. Murphy, K. *Machine Learning: A Probabilistic Perspective*; MIT Press: Cambridge, MA, USA, 2012.
24. Baum, L.E.; Petries, T.; Soules, G.; Weiss, N. A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *Ann. Math. Stat.* **1970**, *41*, 1, 167–171. [CrossRef]
25. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum Likelihood from Incomplete Data Via the EM Algorithm. *J. R. Stat. Soc. Ser. B* **1977**, *39*, 1–22 . [CrossRef]
26. PortfoliosLab. Available online: https://portfolioslab.com/ (accessed on 28 April 2024).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

AppliedMath

MDPI

Article

From Algebro Geometric Solutions of the Toda Equation to Sato Formulas

Pierre Gaillard

Faculté des Sciences Mirande, Université de Bourgogne Franche-Comté, 2100 Dijon, France; pierre.gaillard@u-bourgogne.fr

Abstract: We know that the degeneracy of solutions to PDEs, given in terms of theta functions on Riemann surfaces, provides important results about particular solutions, as in the case of the NLS equation. Here, we degenerate the so called finite gap solutions of the Toda lattice equation from the general formulation in terms of abelian functions when the gaps tend to points. This degeneracy allows us to recover the Sato formulas without using inverse scattering theory or geometric or representation theoretic methods.

Keywords: Riemann surfaces; theta functions; abelian integrals; Baker Akhiezer functions; Sato formulas

MSC: 35C99, 35Q35, 35Q53

1. Introduction

Different formulations can be chosen to represent the Toda equation. In the original work of 1967 [1], the vibration of a uniform chain of particles with nonlinear interaction was studied, and Toda considered the following equation for the n-th particle in the chain:

$$m\partial_x^2(u_n) = -\partial_r(\phi)(u_n - u_{n-1}) + \partial_r(\phi)(u_{n+1} - u_n), \tag{1}$$

where m stands for the mass of the particles, and $\phi(r)$ is the interaction energy between adjacent particles. Some solutions were constructed for particular interactions ϕ.

The same equation was considered with $m = 1$ by Date and Tanaka in 1976 [2], and they constructed some general solutions but in terms of integrals that are difficult to use.

In the appendix of the paper of Dubrovin [3], published in 1981, Krichever considered the non-abelian version of this equation and provided solutions in terms of the theta function on Riemann surfaces.

In 1995, Matveev and Stahlhoffen [4] considered the following version of the Toda equation:

$$\partial_x^2(u_n) = \exp(u_{n-1} - u_n) - \exp(u_n - u_{n+1}). \tag{2}$$

They used the Darboux transformation to contruct solutions to this equation in terms of Casoratis.

In 2014, Zhang and Zhou [5] used the following representation of the Toda equation

$$\partial_x\partial_t(u_n) = (\partial_t(u_n) + \alpha(t))(u_{n-1} - 2u_n + u_{n+1}). \tag{3}$$

They used the generalization of the exp-function to contruct multiwave solutions to this equation.

More recently, Sun, Ma, and Yu [5] examined the following representation of the Toda equation:

$$\Delta(u_n) = 4(\exp(u_{n-1} - u_n) - \exp(u_n - u_{n+1})). \tag{4}$$

check for updates

Citation: Gaillard, P. From Algebro Geometric Solutions of the Toda Equation to Sato Formulas. *AppliedMath* **2024**, *4*, 856–867. https://doi.org/10.3390/appliedmath4030046

Academic Editor: Carlo Bianca

Received: 26 April 2024
Revised: 2 July 2024
Accepted: 2 July 2024
Published: 9 July 2024

Copyright: © 2024 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

They used a logarithmic transformation to obtain some particular solutions in terms of logarithms.

In 2020, Duarte [6] considered the following representation of the Toda equation:

$$\alpha\partial_x^2(u(x,y,n)) + \beta\partial_y^2(u(x,y,n)) \\ = \exp(u(x,y,n-1) - u(x,y,n)) - \exp(u(x,y,n) - u(x,y,n+1)).$$

He used a particular ansatz combined with the properties of Laplace's equation to contruct some solutions in terms of trigonometric functions.

Also in 2020, Schiebold and Nilson [7] studied another version of the Toda equation in the form

$$\partial_{xy}^2(\log(1+u_n)) = u_{n+1} - 2u_n + u_{n+1}. \tag{5}$$

They constructed solutions in the frame of linear algebra by means of determinants.

Here, we consider the Toda equation in the following representation [8–11]:

$$\begin{cases} \dfrac{r_n}{r_{n+1}}\sqrt{c_n}\psi_{n+1} + v_n\psi_n + \dfrac{r_n}{r_{n-1}}\sqrt{c_{n-1}}\psi_{n-1} = \lambda\psi_n \\ \partial_t\psi_n = \dfrac{r_n}{2r_{n+1}}\sqrt{c_n}\psi_{n+1} + \partial_t(\ln r_n)\psi_n - \dfrac{r_n}{2r_{n-1}}\sqrt{c_{n-1}}\psi_{n-1}, \end{cases} \tag{6}$$

with

$$c_n = \frac{e^{2I_0}\theta^2(tV'' + nD + Z)}{\theta(tV'' + (n+1)D + Z)\theta(tV'' + (n-1)D + Z)} \tag{7}$$

$$v_n = -R_0 + \partial_t \ln\frac{\theta(tV'' + (n-1)D + Z)}{\theta(tV'' + nD + Z)}, \tag{8}$$

where θ is the classical Riemann theta function.

We know that the degeneracy of the solutions to PDEs given in terms of Riemann theta functions provides some important particular solutions. In the case of the NLS equation, we have managed to construct a quasi-rational solution involving the determinant of order N for each positive integer N depending on $2N-2$ real parameters [12]. In the case of the KdV equation, we constructed solutions in terms of Fredholm determinants and Wronskians [13], from Baker–Akhiezer functions.

In this study, we recovered Sato formulas for the Toda equation using this method.

From finite gap solutions given in terms of Riemann theta functions, we constructed some quasi-rational solutions and recovered the Sato formulas, using degeneracy, as given in the frame of the NLS equation [14].

Precisely, we derived multisoliton solutions from finite gap solutions by a limit transition, i.e., by making gaps tend toward points in a certain Riemann surface. This was accomplished in the spirit of [14] or more recently [13]. One strength of this approach is that it does not rely on inverse scattering theory or geometric and representation theoretic methods, which offers a fresh perspective on the problem.

We consider the Riemann surface Γ represented by $\cup_{k=1}^{g} a_k b_k a_k^{-1} b_k^{-1}$ of the algebraic curve defined by [8,9]

$$\omega^2 = \prod_{j=1}^{2g+2}(z - E_j),$$

with $E_j \neq E_k$, $j \neq k$.

Let us consider Ω_1'' and Ω_0' [11] abelian integrals, verifying

$$\int_{a_k} d\Omega_1''(P) = \int_{a_k} d\Omega_0'(P) = 0 \qquad k = 1, \ldots, g, \tag{9}$$

with the following asymptotic behavior for $P = (\omega, z) \in \Gamma$ [11],

$$\Omega_1''(P) = \pm\frac{1}{2}(z + R_0 + O(z^{-1})), \qquad P \to P_\infty^\pm \tag{10}$$

$$\Omega_0'(P) = \pm(\ln z - I_0 + O(z^{-1})), \qquad P \to P_\infty^\pm, \tag{11}$$

$$\omega = \pm(z^{g+1} + O(z^g)), \qquad P \to P_\infty^\pm. \tag{12}$$

The vectors V'', D, X, and Z are defined by

$$V_j'' = \frac{1}{2\pi i}\int_{b_j} d\Omega_1'', \tag{13}$$

$$D = U(P_\infty^+) - U(P_\infty^-), \tag{14}$$

$$X = K + \sum_{j=1}^{g} U(P_j), \qquad Z = U(P_\infty^+) - X, \tag{15}$$

where K is the vector of Riemann constants.

$U(P)$ is the classical abelian integral $\int_{P_0}^{P} dU$.

The solutions to the system (6) can be written [10,11] as

$$\psi_n(P,t) = \tilde{r}_n(t)e^{n\Omega_0'(P)+t\Omega_1''(P)}\left(\frac{\theta(U(P)+tV''+nD-X)}{\theta(U(P)-X)}\right), \tag{16}$$

with

$$\tilde{r}_n(t) = r_n(t)\left(\frac{\theta(Z)\theta(Z-D)}{\theta(tV''+nD+Z)\theta(tV''+(n-1)D+Z)}\right)^{\frac{1}{2}}. \tag{17}$$

2. Degeneracy of Solutions

Let us suppose that E_j are real, $E_m < E_j$ if $m < j$, and try to evaluate the limits of all objects in Formula (16) when E_{2m}, E_{2m-1} tends to $-\alpha_m$, $\alpha_m = \kappa_m^2$, $\kappa_m > 0$, for $1 \le m \le g$.

As in the previous section, θ is constructed from the matrix of the B-periods of the surface Γ, the coefficients c_{jk} are related with abelian differentials dU_j by

$$dU_j = \frac{\sum_{k=1}^{g} c_{jk} z^{g-k}}{\sqrt{\prod_{k=1}^{2g+2}(z-E_k)}}dz, \tag{18}$$

and the coefficients c_{jk} can be obtained by solving the system of linear equations

$$\int_{a_k} dU_j = \delta_{jk}, \quad 1 \le j \le g, \quad 1 \le k \le g.$$

In the remainder of this article, we use the following notations:

$$\begin{array}{ll} \kappa_{mj} = \sqrt{(\alpha_m + E_{2g+1})(\alpha_j + E_{2g+2})}, & \kappa_{jm} = \sqrt{(\alpha_j + E_{2g+1})(\alpha_m + E_{2g+2})}, \\ C_k(z) = \sqrt{(z - E_{2g+1})(\alpha_k - E_{2g+2})}, & D_k(z) = \sqrt{(z - E_{2g+2})(\alpha_k - E_{2g+1})}, \\ L_k = \sqrt{\alpha_k + E_{2g+1}}, & M_k = \sqrt{\alpha_k + E_{2g+2}}, \\ F(z) = \sqrt{z - E_{2g+1}}, & G(z) = \sqrt{z - E_{2g+2}}. \end{array} \tag{19}$$

2.1. Limit of $P(z) = \prod_{j=1}^{2g+2}(z - E_j)$

Now, we study the limit of $P(z) = \prod_{j=1}^{2g+2}(z - E_j)$.

The limit of $P(z) = \prod_{j=1}^{2g+2}(z - E_j)$ is obviously equal to $\tilde{P}(z) = \prod_{j=1}^{g}(z + \alpha_j)^2(z - E_{2g+1})(z - E_{2g+2})$ or, with (19),

$$\tilde{P}(z) = \prod_{j=1}^{g}((z + \alpha_j)F(z)G(z))^2. \tag{20}$$

2.2. Limit of $dU_m = \frac{\sum_{k=1}^{g} c_{mk} z^{g-k}}{\sqrt{\prod_{k=1}^{2g+2}(z-E_k)}}dz$

Now, we study the limit of $dU_m = \dfrac{\sum_{k=1}^{g} c_{mk} z^{g-k}}{\sqrt{\prod_{k=1}^{2g+2}(z - E_k)}}dz$.

The limit of dU_m is equal to $d\tilde{U}_m = \dfrac{\varphi_m(z)}{\prod_{j=1}^{g}(z + \alpha_j)\sqrt{(z - E_{2g+1})(z - E_{2g+2})}}dz$, where $\varphi_m(z) = \sum_{k=1}^{g} \tilde{c}_{mk} z^{g-k}$. The normalization condition takes the form in the limit

$$\int_{a_k} dU_j \to \frac{2\pi i \varphi_j(-\alpha_k)}{\prod_{m\neq k}(\alpha_m - \alpha_k)\sqrt{(\alpha_k + E_{2g+1})(\alpha_k + E_{2g+2})}} = \delta_{kj}, \tag{21}$$

which proves that the numbers $-\alpha_m$, $m \neq k$ are the zeros of the polynomials $\varphi_k(z)$; hence, $\varphi_k(z)$ can be written as $\varphi_k(z) = \tilde{c}_{k1}\prod_{m\neq k}(z + \alpha_m)$.

By (21), we obtain in the limit

$$\tilde{c_{k1}} = \frac{\sqrt{(\alpha_k + E_{2g+1})(\alpha_k + E_{2g+2})}}{2\pi i}.$$

So,

$$\varphi_k(z) = \frac{\sqrt{(\alpha_k + E_{2g+1})(\alpha_k + E_{2g+2})}}{2\pi i}\prod_{m\neq k}(z + \alpha_m).$$

Moreover,

$$d\tilde{U}_k(z) = \frac{\varphi_k(z)dz}{\sqrt{(z - E_{2g+1})(z - E_{2g+2})}\prod_{m=1}^{g}(z + \alpha_m)};$$

in other words,

$$d\tilde{U}_k(z) = \frac{\sqrt{(\alpha_k + E_{2g+1})(\alpha_k + E_{2g+2})}}{2\pi i\sqrt{(z - E_{2g+1})(z - E_{2g+2})}(z + \alpha_k)}dz,$$

and with (19),

$$d\tilde{U}_k(z) = \frac{L_k M_k}{2\pi i F(z)G(z)(z + \alpha_k)}dz. \tag{22}$$

2.3. Limit of B_{mk}

The subject of this subsection is the study of the limit of B_{mk}.

We have

$$I = \int_{\alpha_m}^{E_{2g+2}} dU_k \to \frac{1}{2}\tilde{B}_{mk}.$$

The integral I can be easily evaluated along the real axis on the upper sheet of surface Γ, and we obtain

$$\tilde{B}_{mk} = \frac{i}{\pi} \ln \left| \frac{\sqrt{(\alpha_m + E_{2g+1})(\alpha_k + E_{2g+2})} + \sqrt{(\alpha_k + E_{2g+1})(\alpha_m + E_{2g+2})}}{\sqrt{(\alpha_m + E_{2g+1})(\alpha_k + E_{2g+2})} - \sqrt{(\alpha_k + E_{2g+1})(\alpha_m + E_{2g+2})}} \right|,$$

or with the previous notations (19),

$$\tilde{B}_{mk} = \frac{i}{\pi} \ln \left| \frac{\kappa_{mk} + \kappa_{km}}{\kappa_{mk} - \kappa_{km}} \right|. \tag{23}$$

So, iB_{kk} tends to $-\infty$.

Moreover, we have

$$\int_{E_{2g+2}}^{P} d\tilde{U}_k = \frac{-i}{2\pi} \ln \left| \frac{\sqrt{(z - E_{2g+1})(\alpha_k + E_{2g+2})} + \sqrt{(z - E_{2g+2})(\alpha_k + E_{2g+1})}}{\sqrt{(z - E_{2g+1})(\alpha_k + E_{2g+2})} - \sqrt{(z - E_{2g+2})(\alpha_k + E_{2g+1})}} \right|,$$

or with the previous notations (19),

$$\int_{E_{2g+2}}^{P} d\tilde{U}_k = \frac{-i}{2\pi} \ln \left| \frac{C_k(z) + D_k(z)}{C_k(z) - D_k(z)} \right|. \tag{24}$$

2.4. Limit of $\Omega_1^{''}$

In this subsection, we study the limit of $\Omega_1^{''}$.

$\Omega_1^{''}$ is an abelian integral of the second kind satisfying the conditions

$$\int_{a_k} d\Omega_1^{''}(P) = 0, \quad 1 \leq k \leq g,$$

such that

$$\Omega_1''(P) = \pm \frac{1}{2}(z + R_0 + O(z^{-1})), \qquad P \to P_\infty^{\pm}$$

The limit of $d\Omega_1''(P)$ is equal to $d\tilde{\Omega}_1''(P) = \dfrac{\phi(z)}{\prod_{j=1}^{g}(z + \alpha_j)\sqrt{(z - E_{2g+1})(z - E_{2g+2})}} dz$,

where $\phi(z) = \sum_{k=0}^{g+1} \tilde{c}_k z^{g+1-k}$.

For $\Omega_1''(P)$, satisfying the condition $\Omega_1''(P) = \pm\frac{1}{2}z + O(1)$ when $z \to \pm\infty$, we have $\tilde{c_0} = \frac{1}{2}$. Moreover, the conditions

$$\int_{a_k} d\Omega_1''(P) = 0, \quad 1 \leq k \leq g$$

prove that $-\alpha_1, \ldots, -\alpha_g$ are the zeros of ϕ; thus, $\phi(z) = \frac{1}{2}\prod_{m=1}^{g}(z + \alpha_m)(z - \beta)$.

We have that $P(z) = \sqrt{\prod_{k=1}^{2g+2}(z - E_k)}$ tends to $\tilde{P}(z) = \sqrt{(z - E_{2g+1})(z - E_{2g+2})}\prod_{m=1}^{g}(z + \alpha_m)$, and we obtain

$$d\tilde{\Omega}_1''(z) = \frac{z}{2\sqrt{(z - E_{2g+1})(z - E_{2g+2})}} dz;$$

$\Omega_1''(P)$ tends to

$$\tilde{\Omega}_1''(P) = \int_{E_{2g+2}}^{z} \frac{u}{2\sqrt{(u - E_{2g+1})(u - E_{2g+2})}} du.$$

This can be evaluated, and it gives

$$\tilde{\Omega}_1''(P) = \frac{E_{2g+1}+E_{2g+2}}{4}\ln\left|\frac{\sqrt{z-E_{2g+1}}+\sqrt{z-E_{2g+2}}}{\sqrt{z-E_{2g+1}}-\sqrt{z-E_{2g+2}}}\right| + \frac{1}{2}\sqrt{(z-E_{2g+1})(z-E_{2g+2})}.$$

With the notations defined in (19), it can be rewritten as

$$\tilde{\Omega}_1''(P) = \frac{E_{2g+1}+E_{2g+2}}{4}\ln\left|\frac{F(z)+G(z)}{F(z)-G(z)}\right| + \frac{1}{2}F(z)G(z). \tag{25}$$

2.5. Limit of $\Omega_0^{'}$

We consider $\Omega_0^{'}$, and we study its limit.

$\Omega_0^{'}$ is an abelian integral of the third kind satisfying the conditions

$$\int_{a_k} d\Omega_0^{'}(P) = 0, \quad 1 \le k \le g,$$

such that

$$\Omega_0'(P) = \pm(\ln z - I_0 + O(z^{-1})), \qquad P \to P_\infty^{\pm}.$$

The limit of $d\Omega_0'(P)$ is equal to

$$d\tilde{\Omega'}_0(P) = \frac{\sum_{k=0}^{g}\tilde{c}_k z^{g-k}}{\prod_{j=1}^{g}(z+\alpha_j)\sqrt{(z-E_{2g+1})(z-E_{2g+2})}}dz.$$

$\Omega_0'(P)$ satisfies the condition $\Omega_0'(P) = \pm\ln(z) + O(1)$, and when $z \to \pm\infty$, we have $\tilde{c_0} = 1$.

Moreover, the conditions

$$\int_{a_k} d\Omega_0'(P) = 0, \quad 1 \le k \le g$$

prove that $-\alpha_1, \ldots, -\alpha_g$ are the zeros of ϕ; thus, $\phi(z)$, defined by $\phi(z) = \sum_{k=0}^{g}\tilde{c}_k z^{g-k}$, can be written as $\phi(z) = \prod_{m=1}^{g}(z+\alpha_m)$.

As $P(z) = \sqrt{\prod_{k=1}^{2g+2}(z-E_k)}$ tends to $\tilde{P}(z) = \sqrt{(z-E_{2g+1})(z-E_{2g+2})}\prod_{m=1}^{g}(z+\alpha_m)$, we obtain

$$d\tilde{\Omega'}_0(z) = \frac{1}{\sqrt{(z-E_{2g+1})(z-E_{2g+2})}}dz,$$

and $\Omega_0'(P)$ tends to $\tilde{\Omega'}_1(P) = \int_{E_{2g+2}}^{z}\frac{1}{\sqrt{(u-E_{2g+1})(u-E_{2g+2})}}du$.

This can be calculated, and it gives

$$\tilde{\Omega}_0'(P) = \ln\left|\frac{\sqrt{z-E_{2g+1}}+\sqrt{z-E_{2g+2}}}{\sqrt{z-E_{2g+1}}-\sqrt{z-E_{2g+2}}}\right|.$$

It can be written with (19) as

$$\tilde{\Omega}_0'(P) = \ln\left|\frac{F(z)+G(z)}{F(z)-G(z)}\right|. \tag{26}$$

2.6. Limit of X

We deduce the limit of X.

From the previous sections,

$$\tilde{B}_{mk} = \frac{i}{\pi}\ln\left|\frac{\kappa_{mk}+\kappa_{km}}{\kappa_{mk}-\kappa_{km}}\right|,$$

$$\tilde{U}_k(P) = \frac{i}{2\pi} \ln \left| \frac{C_k(z) + D_k(z))}{C_k(z) - D_k(z)} \right|.$$

As K is defined by $K_m = \sum_{j=1}^{g} B_{jm} - \frac{m}{2}$, and $X = K + \sum_{j=1}^{g} U(P_j)$, we can write

$$X_m \to \tilde{X_m} = \frac{i}{\pi} \sum_{j=1, j\neq m}^{g} \ln \left| \frac{\kappa_{mj} + \kappa_{jm}}{\kappa_{mj} - \kappa_{jm}} \right| - \frac{m}{2} + \frac{i}{2\pi} \sum_{j=1}^{g} \ln \left| \frac{C_m(z_j) + D_m(z_j)}{C_m(z_j) - D_m(z_j)} \right| + i\infty. \tag{27}$$

2.7. Limit of $U(P_\infty^+)$, $U(P_\infty^-)$, D and Z

In this subsection, we study the limit of $U(P_\infty^+)$, $U(P_\infty^-)$, D, and Z.
From the previous sections, it is easy to obtain the limits of $\tilde{U}_k$ as

$$\lim_{P \to P_\infty^+} \tilde{U}_k(P) = \frac{-i}{2\pi} \ln \left| \frac{\sqrt{\alpha_k + E_{2g+2}} + \sqrt{\alpha_k + E_{2g+1}}}{\sqrt{\alpha_k + E_{2g+2}} - \sqrt{\alpha_k + E_{2g+1}}} \right|,$$

or

$$\lim_{P \to P_\infty^+} \tilde{U}_k(P) = \frac{-i}{2\pi} \ln \left| \frac{M_k + L_k}{M_k + L_k} \right|. \tag{28}$$

$$D_k = U_k(P_\infty^+) - U_k(P_\infty^-) \to \tilde{D_k} = -\frac{i}{2\pi} \ln \frac{M_k + iL_k}{M_k - iL_k}. \tag{29}$$

Thus,

$$\lim_{P \to P_\infty^-} \tilde{U}_k(P) = \frac{-i}{2\pi} \ln \left| \frac{\sqrt{\alpha_k + E_{2g+2}} + \sqrt{\alpha_k + E_{2g+1}}}{\sqrt{\alpha_k + E_{2g+2}} - \sqrt{\alpha_k + E_{2g+1}}} \right| - D_k,$$

or

$$\lim_{P \to P_\infty^-} \tilde{U}_k(P) = \frac{-i}{2\pi} \ln \left| \frac{M_k + L_k}{M_k + L_k} \right| + \frac{i}{2\pi} \ln \left| \frac{M_k + iL_k}{M_k + iL_k} \right|. \tag{30}$$

Therefore, the limit of $Z = U(P_\infty^+) - X$ is given by

$$Z_k \to \tilde{Z_k} = -\frac{i}{\pi} \sum_{j=1, j\neq k}^{g} \ln \left| \frac{\kappa_{kj} + \kappa_{jk}}{\kappa_{kj} - \kappa_{jk}} \right| + \frac{k}{2} - \frac{i}{2\pi} \sum_{j=1,}^{g} \ln \left| \frac{C_k(z_j) + D_k(z_j)}{C_k(z_j) - D_k(z_j)} \right| - i\infty - \frac{i}{2\pi} \ln \left| \frac{M_k + L_k}{M_k - L_k} \right| \tag{31}$$

2.8. Limit of V''

In this paragraph, the limit of V'' is studied.
$V_k'' = \frac{1}{2\pi i} \int_{b_k} d\Omega_1''$ is defined in (13).
We have $I = \int_{-\alpha_k}^{E_{2g+2}} d\Omega_1'' \to \frac{1}{2} \int_{b_k} d\tilde{\Omega}_1''$.
As in the determination of the limit of $\Omega_1^{''}$, we have

$$\int_{-\alpha_k}^{E_{2g+2}} d\tilde{\Omega}_1''(P) = -\frac{E_{2g+1} + E_{2g+2}}{4} \ln \left| \frac{\sqrt{\alpha_k + E_{2g+1}} + \sqrt{\alpha_k + E_{2g+2}}}{\sqrt{\alpha_k + E_{2g+1}} - \sqrt{\alpha_k + E_{2g+2}}} \right|$$

$$-\frac{1}{2} \sqrt{(\alpha_k + E_{2g+1})(\alpha_k + E_{2g+2})}.$$

So,

$$V_k'' = \frac{1}{2\pi i} \int_{b_k} d\Omega_1'' \to \tilde{V}'' = -\frac{E_{2g+1} + E_{2g+2}}{4\pi i} \ln \left| \frac{\sqrt{\alpha_k + E_{2g+1}} + \sqrt{\alpha_k + E_{2g+2}}}{\sqrt{\alpha_k + E_{2g+1}} - \sqrt{\alpha_k + E_{2g+2}}} \right|$$

$$-\frac{1}{2\pi i} \sqrt{(\alpha_k + E_{2g+1})(\alpha_k + E_{2g+2})},$$

$$\tilde{V}_k'' = \frac{E_{2g+1} + E_{2g+2}}{4\pi i} \ln \left| \frac{L_k - M_k}{L_k + M_k} \right| - \frac{1}{2\pi i} L_k M_k. \tag{32}$$

2.9. Limit of $\theta(p)$

We determine the limit of $\theta(p)$.

Let us denote A as the argument of $\theta(p) = \sum_{k\in\mathbf{Z}^g} \exp\{\pi i(Bk|k) + 2\pi i(k|p)\}$.

A can be rewritten in the form

$$A = \pi i \sum_{j=1}^{g} B_{jj}k_j(k_j-1) + 2\pi i \sum_{j>m} B_{mj}k_m k_j + \sum_{j=1}^{g} \pi i(2p_j + B_{jj})k_j.$$

Using the inequality $k_j(k_j-1) \geq 0$ for all $k \in \mathbf{Z}^g$ and the fact that iB_{kk} tends to $-\infty$, we can reduce the limit $\tilde{\theta}$ of $\theta(p)$ to a finite sum taken over vectors $k \in \mathbf{Z}^g$, such that each k_j must be equal to 0 or 1.

In this section, we compute the limit of all the terms in the expression of the solution $\psi(P,t)$ in θ. We denote p_i as the arguments of these different expressions.

We first study the terms of $\tilde{r}_n$ corresponding to the arguments $p_1 = Z = U(P_\infty^+) - X$, $p_2 = Z - D$, $p_3 = tV'' + nD - Z$, $p_4 = tV'' + (n-1)D - Z$.

Then, we study the remaining terms of ψ_n corresponding to the arguments $p_5 = U(P) - X$, $p_6 = U(P) + tV'' + nD - X$.

With these notations, the solution $\psi(n,,t,z)$ can be written as

$$\psi_n(P,z) = r_n(t)\left(\frac{\theta(p_1)\theta(p_2)}{\theta(p_3)\theta(p_4)}\right)^{\frac{1}{2}} \frac{\theta(p_6)}{\theta(p_5)} e^{n\Omega_0'(P)+t\Omega_1''(P)}.$$

2.10. Limit of $\theta(p_1)$

We denote p_1 as the term $Z = U(P_\infty^+) - X$.

We study its limit.

Then, $\theta(p_1) \to \tilde{\theta}_1$:

$$\tilde{\theta}_1 = \sum_{k\in\mathbf{Z}^g, k_j=0\,or\,1} \exp\left\{\sum_{j>l} 2\ln\left|\frac{\kappa_{lj}-\kappa_{jl}}{\kappa_{lj}+\kappa_{jl}}\right| k_j k_l + \sum_{j=1}^{g} k_j\left(\ln\left|\frac{M_j+L_j}{M_j-L_j}\right|\right.\right.$$

$$\left.\left. + \sum_{l\neq j} 2\ln\left|\frac{\kappa_{jl}+\kappa_{lj}}{\kappa_{jl}-\kappa_{lj}}\right| + j\pi i + \sum_{l=1}^{g}\ln\left|\frac{C_j(z_l)+D_j(z_l)}{C_j(z_l)-D_j(z_l)}\right|\right)\right\}.$$

2.11. Limit of $\theta(p_2)$

We consider $p_2 = Z - D$. We study its limit.

Then, $\theta(p_2) \to \tilde{\theta}_2$. As $D_j \to \tilde{D}_j$, it is easy to see that

$$\tilde{\theta}_2 = \sum_{k\in\mathbf{Z}^g, k_j=0\,or\,1} \exp\left\{\sum_{j>l} 2\ln\left|\frac{\kappa_{lj}-\kappa_{jl}}{\kappa_{lj}+\kappa_{jl}}\right| k_j k_l + \sum_{j=1}^{g} k_j\left(\ln\left|\frac{M_j+L_j}{M_j-L_j}\right|\right.\right.$$

$$\left.\left. + \sum_{l\neq j} 2\ln\left|\frac{\kappa_{jl}+\kappa_{lj}}{\kappa_{jl}-\kappa_{lj}}\right| + j\pi i + \sum_{l=1}^{g}\ln\left|\frac{C_j(z_l)+D_j(z_l)}{C_j(z_l)-D_j(z_l)}\right| + \sum_{j=1}^{g}\ln\left|\frac{M_j+iL_j}{M_j-iL_j}\right|\right)\right\}.$$

2.12. Limit of $\theta(p_3)$

Let p_3 be $tV'' + nD + Z$. We study its limit.

Then, $\theta(p_3) \to \tilde{\theta}_3$:

$$\tilde{\theta}_3 = \sum_{k\in\mathbf{Z}^g, k_j=0\,or\,1} \exp\left\{\sum_{j>l} 2\ln\left|\frac{\kappa_{lj}-\kappa_{jl}}{\kappa_{lj}+\kappa_{jl}}\right| k_j k_l + \sum_{j=1}^{g} k_j\left(\ln\left|\frac{L_j+M_j}{L_j-M_j}\right|\right.\right.$$

$$\left.\left. + t\left(\left(\frac{E_{g+1}+E_{2g+2}}{2}\right)\ln\left|\frac{L_j-M_j}{L_j+M_j}\right| - L_jM_j\right) + n\ln\left|\frac{M_j+iL_j}{M_j-iL_j}\right|\right.\right.$$

$$+\sum_{l\neq j} 2\ln\left|\frac{\kappa_{jl}+\kappa_{lj}}{\kappa_{jl}-\kappa_{lj}}\right| + j\pi i + \sum_{l=1}^{g}\ln\left|\frac{C_j(z_l)+D_j(z_l)}{C_j(z_l)-D_j(z_l)}\right|\Bigg)\Bigg\}.$$

2.13. Limit of $\theta(p_4)$

We consider $p_4 = tV'' + (n-1)D + Z$; $\theta(p_4) \to \tilde{\theta}_4$. We determine its limit.
As $D_j \to \tilde{D}_j$, it is easy to verify that

$$\tilde{\theta}_4 = \sum_{k\in\mathbf{Z}^{\mathbf{g}}, k_j=0\,or\;1} \exp\Bigg\{\sum_{j>l} 2\ln\left|\frac{\kappa_{lj}-\kappa_{jl}}{\kappa_{lj}+\kappa_{jl}}\right| k_j k_l + \sum_{j=1}^{g} k_j\Bigg(\ln\left|\frac{L_j+M_j}{L_j-M_j}\right|$$

$$+t\Bigg(\left(\frac{E_{g+1}+E_{2g+2}}{2}\right)\ln\left|\frac{L_j-M_j}{L_j+M_j}\right| + L_jM_j\Bigg) + (n-1)\ln\left|\frac{M_j+iL_j}{M_j-iL_j}\right|$$

$$+\sum_{l\neq j} 2\ln\left|\frac{\kappa_{jl}+\kappa_{lj}}{\kappa_{jl}-\kappa_{lj}}\right| + j\pi i + \sum_{l=1}^{g}\ln\left|\frac{C_j(z_l)+D_j(z_l)}{C_j(z_l)-D_j(z_l)}\right|\Bigg)\Bigg\}.$$

2.14. Limit of $\theta(p_5)$

The term $p_5 = U(P) - X$ tends to $\tilde{p}_5 = \tilde{U}(P) - \tilde{X}$.
We determine its limit.
Then, $\theta(p_5) \to \tilde{\theta}_5$:

$$\tilde{\theta}_5 = \sum_{k\in\mathbf{Z}^{\mathbf{g}}, k_j=0\,or\;1} \exp\Bigg\{\sum_{j>l} 2\ln\left|\frac{\kappa_{lj}-\kappa_{jl}}{\kappa_{lj}+\kappa_{jl}}\right| k_j k_l + \sum_{j=1}^{g} k_j\Bigg(\ln\left|\frac{C_j(z)-D_j(z)}{C_j(z)+D_j(z)}\right|$$

$$+\sum_{l\neq j} 2\ln\left|\frac{\kappa_{jl}+\kappa_{lj}}{\kappa_{jl}-\kappa_{lj}}\right| + j\pi i + \sum_{l=1}^{g}\ln\left|\frac{C_j(z_l+D_j(z_l}{C_j(z_l-D_j(z_l}\right|\Bigg)\Bigg\}.$$

2.15. Limit of $\theta(p_6)$

The term $p_6 = U(P) + tV'' + nD - X$ tends to $\tilde{p}_6 = \tilde{U}(P) + t\tilde{V}'' - \tilde{X} + n\tilde{D}$.
Then, we determine its limit. $\theta(p_6) \to \tilde{\theta}_6$:

$$\tilde{\theta}_6 = \sum_{k\in\mathbf{Z}^{\mathbf{g}}, k_j=0\,or\;1} \exp\Bigg\{\sum_{j>l} 2\ln\left|\frac{\kappa_{lj}-\kappa_{jl}}{\kappa_{lj}+\kappa_{jl}}\right| k_j k_l + \sum_{j=1}^{g} k_j\Bigg(\ln\left|\frac{L_j+M_j}{L_j-M_j}\right|$$

$$+t\Bigg(\left(\frac{E_{g+1}+E_{2g+2}}{2}\right)\ln\left|\frac{L_j-M_j}{L_j+M_j}\right| - L_jM_j\Bigg) + n\ln\left|\frac{M_j+iL_j}{M_j-iL_j}\right|$$

$$+\sum_{l\neq j} 2\ln\left|\frac{\kappa_{jl}+\kappa_{lj}}{\kappa_{jl}-\kappa_{lj}}\right| + j\pi i - \sum_{l=1}^{g}\ln\left|\frac{C_j(z_l)+D_j(z_l)}{C_j(z_l)-D_j(z_l)}\right| + \ln\left|\frac{C_j(z)-D_j(z)}{C_j(z)+D_j(z)}\right|\Bigg)\Bigg\}.$$

2.16. Limit of Solutions Ψ

From the previous section, we can give the limit $\tilde{\psi}$ of ψ. It takes the following form:

$$\tilde{\psi}(z,n,t)) = \frac{r_n(t)\left(\tilde{\theta_1}\tilde{\theta_2}\right)^{\frac{1}{2}}}{\tilde{\theta_5}} \times \frac{e^{\left\{n\ln\left|\frac{F(z)+G(z)}{F(z)-G(z)}\right| + t\left(\frac{E_{2g+1}+E_{2g+2}}{4}\ln\left|\frac{F(z)+G(z)}{F(z)-G(z)}\right| + \frac{1}{2}F(z)G(z)\right)\right\}}\tilde{\theta_6}}{\left(\tilde{\theta_3}\tilde{\theta_4}\right)^{\frac{1}{2}}}.$$

In the previous expression, $\tilde{\theta}_1$, $\tilde{\theta}_2$, and $\tilde{\theta}_5$ are independent of n and t.
Only $\tilde{\theta}_3$, $\tilde{\theta}_4$, and $\tilde{\theta}_6$ depend on n and t.
We choose the particular case in which $t = 0$.
We replace n by x.
We denote c as the coefficient defined by

$$c = \frac{r_n(0)\left(\tilde{\theta_1}\tilde{\theta_2}\right)^{\frac{1}{2}}}{\tilde{\theta_5}}.$$

We denote H as the function defined by

$$H(z) = \left|\frac{F(z) - G(z)}{F(z) + G(z)}\right|.$$

In the different sums involving θ, only terms with $k_j = 1$ remain. So, the sums can be reduced only on the subsets J of $[1; g]$ of integer $\mathbf{N}$. We denote a_j as the term $a_j = -\frac{1}{2}\left|\frac{M_j + iL_j}{M_j - iL_j}\right|$. Then, the term $\tilde{\theta}_6$ can be written as

$$\tilde{\theta}_6 = \sum_{J\subset\{1,\ldots,g\}} \prod_{j,l\in J, j<l} \ln\left|\frac{\kappa_{lj} - \kappa_{jl}}{\kappa_{lj} + \kappa_{jl}}\right|^2 \prod_{j\in J}\left|\frac{L_j + M_j}{L_j - M_j}\right|$$

$$\prod_{l\in Jl\neq j} \ln\left|\frac{\kappa_{jl} + \kappa_{lj}}{\kappa_{jl} - \kappa_{lj}}\right|^2 \prod_{j\in J}\left|\frac{C_j(z_l) + D_j(z_l)}{C_j(z_l) + D_j(z_l)}\right|\left|\frac{C_j(z) - D_j(z)}{C_j(z) + D_j(z)}\right| e^{\sum_{j\in J} j\pi i} e^{-2x\sum_{j\in J} a_j}.$$

We denote τ as the function defined by

$$\tau(x) = \tilde{\theta}_3(x) = \sum_{J\subset\{1,\ldots,g\}} \prod_{j,l\in J, j<l} \ln\left|\frac{\kappa_{lj} - \kappa_{jl}}{\kappa_{lj} + \kappa_{jl}}\right|^2 \prod_{j\in J}\left|\frac{L_j + M_j}{L_j - M_j}\right|$$

$$\prod_{l\in Jl\neq j} \ln\left|\frac{\kappa_{jl} + \kappa_{lj}}{\kappa_{jl} - \kappa_{lj}}\right|^2 \prod_{j\in J}\left|\frac{C_j(z_l) + D_j(z_l)}{C_j(z_l) - D_j(z_l)}\right| e^{\sum_{j\in J} j\pi i} e^{-2x\sum_{j\in J} a_j}.$$

Thus, the solution $\tilde{\Psi}$ can be written as

$$\tilde{\Psi}(x, z) = c\frac{\exp(xH(z))\chi(x, z)}{\sqrt{\tau(x)\tau(x-1)}}. \tag{33}$$

Here, we return to the Sato formulation of the solution, as given in [15].

This expression is very similar to the solution expressed on page 5830, and the expression of the solution is given by

$$\tilde{\Psi}(x, z) = \frac{\exp(xz)\chi(x, z)}{\sqrt{\tau(x)\tau(x-1)}}. \tag{34}$$

The difference between these two statements comes from the fact that these two equations are treated differently:

Van Diejen considers the equation

$$a(x)f(x+1) + b(x)f'x) + a(x-1)f(x-1) = \lambda f(x);$$

here, we consider (1)

$$\frac{r_n}{r_{n+1}}\sqrt{c_n}\psi_{n+1} + v_n\psi_n + \frac{r_n}{r_{n-1}}\sqrt{c_{n-1}}\psi_{n-1} = \lambda\psi_n.$$

2.17. Limit of the Associated Potentials

In this subsection, we determine the limit of the associated potentials.

The potentials $u_n = \sqrt{c_n}$ and v_n are described in (6) and (7) as

$$c_n = \frac{e^{2I_0}\theta^2(tV'' + nD + Z)}{\theta(tV'' + (n+1)D + Z)\theta(tV'' + (n-1)D + Z)}$$

$$v_n = -R_0 + \partial_t \ln \frac{\theta(tV'' + (n-1)D + Z)}{\theta(tV'' + nD + Z)}.$$

So, the limit of u_n is equal to

$$\tilde{u_n} = k_1 \frac{\tau(x)}{\sqrt{\tau(x+1)\tau(x-1)}},$$

which is the same as in [15] (p. 5830), up to the constant k_1.

Denoting $\sigma(x) = \partial_t \theta(t\tilde{V}'' + x\tilde{D} + \tilde{Z})_{t=0}$, the limit of v_n is equal to

$$\tilde{v_n} = k_2 + \frac{\sigma(x-1)}{\tau(x-1)} - \frac{\sigma(x)}{\tau(x)},$$

which is similar to the one written in [15] (p. 5830), up to the constant k_2.

3. Conclusions

We have used the degeneracy of solutions of some PDEs given in terms of Riemann theta functions to obtain some important solutions. In particular, in the case of the NLS equation, we have managed to construct a quasi-rational solution involving the determinant of order N for each positive integer N depending on $2N-2$ real parameters [12]. In the case of the KdV equation, from abelian functions, we constructed solutions in terms of the Fredholm determinants and Wronskians [13].

In this study, we have managed to recover Sato formulas using this method for the Toda equation. From solutions given in terms of the Baker–Akiezer functions, we succeeded to construct by degeneracy, as given in the frame of the NLS equation [14], some quasi-rational solutions, and we recovered the Sato formulas for the Toda equation.

Precisely, we derived multisoliton solutions from finite gap solutions by a limit transition, i.e., by making gaps tend toward points in a certain Riemann surface. One strength of this approach is that it does not rely on inverse scattering theory or geometric and representation theoretic methods, which offers another perspective on the problem.

The degeneracy of the solutions to the NLS equation has allowed building new quasi-rational solutions of order N depending on $2N-2$ real parameters and their construction up to order 23, which were previously unknown. In the case of the KdV equation, the degeneracy of the solutions made it possible to find the solutions given by the Darboux method, which constituted a bridge between the geometric algebra approach and the Darboux transformations framework. In this article, this is somewhat the same situation as that of the KdV equation, where we linked the algebro–geometric approach to the framework of the inverse scattering method.

Funding: This research received no external funding.

Data Availability Statement: The data that support the findings of this study are available within the article.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Toda, M. Vibration of a chain with nonlinear interaction. *J. Phys. Soc. Jap.* **1967**, *6*, 22–431. [CrossRef]
2. Date, E.; Tanaka, S. Periodic Multi-Soliton Solutions of Korteweg-de Vries Equation and Toda Lattice. *Suppl. Prog. Theor. Phys.* **1976**, *59*, 107–125. [CrossRef]
3. Dubrovin, B.A. Theta functions and nonlinear equations. *Russ. Math. Surv.* **1981**, *36*, 11. [CrossRef]
4. Stahlhofen, A.A.; Matveev, V.B. Positons for Toda lattice and related spectral problems. *Phys. A Math. Gen.* **1995**, *28*, 1957–1965. [CrossRef]
5. Zhang, S.; Zhou, Y.Y. Multiwave solutions for the Toda lattice equation by generalizing exp function method. *Int. J. Appl. Math.* **2014**, *44*, 1–6.
6. Duarte, V.N. A class of solutions of the two dimensional Toda lattice equation. *Phys. Lett.* **2021**, *385*, 126979. [CrossRef]
7. Nilson, T.; Schiebold, C. Solution formulas for the two dimensional Toda lattice and particle like solitons with unexpected asymptotic behaviour. *J. Nonliner. Math. Phys.* **2020**, *27*, 57–94. [CrossRef]

8. Bellokolos, E.D.; Bobenko, A.I.; Enol'skii, V.Z.; Its, A.R.; Matveev, V.B. *Algebro-Geometric Approach to Nonlinear Integrable Equations*; Series in Nonlinear Dynamics; Springer-Verlag: Berlin/Heidelberg, Germany, 1996.
9. Dubrovin, B.A.; Matveev, V.B.; Novikov, S.P. *Integrable Systems: Non-Linear Equations of Korteweg-De Vries Type, Finite-Zone Linear Operators, and Abelian Varieties*; Lecture Notes Series; London Mathematical Society: London, UK, 1976; Volume 60, pp. 53–140.
10. Krichever, I. The periodic non abelian Toda lattice and two dimensional generalization. *Uspekhi Math Nauka* **1981**, *36*, 72–77.
11. Smirnov, A.O. *Elliptic Solitons of the Integrable Nonlinear Equations*; St.-Petersburg State University: Saint Petersburg, Russia, 2002; pp. 1–170.
12. Gaillard, P. Families of quasi-rational solutions of the NLS equation and multi-rogue waves. *J. Phys. A Meth. Theor.* **2011**, *44*, 435204. [CrossRef]
13. Gaillard, P. Degenerate Riemann theta functions, Fredholm and wronskian representations of the solutions to the KdV equation and the degenerate rational case. *J. Geom. Phys.* **2021**, *161*, 104059. [CrossRef]
14. Its, A.R.; Rybin, A.V.; Salle, M.A. Exact integration of NLS equation. *Theor. Math. Phys.* **1988**, *74*, 29–32. [CrossRef]
15. Diejen, J.F.V. A Sato formula for reflectionless finite difference operators. *J. Math. Phys.* **1999**, *40*, 5822–5834. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

AppliedMath

MDPI

Article

Basic Circuit Model of Voltage Source Converters: Methodology and Modeling

Christian Bipongo Ndeke *, Marco Adonis and Ali Almaktoof

Department of Electrical, Electronic and Computer Engineering, Cape Peninsula University of Technology, Cape Town 7535, South Africa; adonisma@cput.ac.za (M.A.); almaktoofa@cput.ac.za (A.A.)
* Correspondence: ndekechristian@gmail.com or 219385017@mycput.ac.za; Tel.: +27-62-836-2466

Abstract: Voltage source converters (VSCs) have emerged as the key components in modern power systems, facilitating efficient energy conversion and flexible power flow control. Understanding the fundamental circuit model of VSCs is essential for their accurate modeling and analysis in power system studies. A basic voltage source converter circuit model connected to an LC filter is essential because it lowers the harmonic distortions and enhances the overall power quality of the micro-grid. This guarantees a clean and steady power supply, which is necessary for the integration of multiple renewable energy sources and sensitive loads. A comprehensive methodology for developing a basic circuit model of VSCs, focusing on the key components and principals involved, is presented in this paper. The methodology includes the modeling of space vector pulse-width modulation (SVPWM) as well as the direct quadrature zero synchronous reference frame. Different design controls, including the design of current control loop in the S-domain, the design of the direct current (DC) bus voltage control loop in the S-domain, and the design of the alternating current (AC) voltage control loop in the S-domain, are explored to capture the dynamic behavior and control strategies of VSCs accurately. The proposed methodology provides a systematic framework for modeling VSCs, enabling engineers and researchers to analyze their performance and assess their impact on power system stability and operation. Future studies can be conducted by using case studies and simulation scenarios to show the efficiency and applicability of the developed models in analyzing VSC-based power electronics applications, including high-voltage direct current (HVDC) transmission systems and flexible alternating current transmission systems (FACTS). The significance of this work lies in its potential to advance the understanding and application of VSCs, contributing to more resilient and efficient power systems. By providing a solid foundation for future research and development, this study supports the ongoing integration of renewable energy sources and the advancement of modern electrical infrastructure.

Keywords: voltage source converter; LC filter; current control; DC voltage control; AC voltage control

check for updates

Citation: Ndeke, C.B.; Adonis, M.; Almaktoof, A. Basic Circuit Model of Voltage Source Converters: Methodology and Modeling. *AppliedMath* **2024**, *4*, 889–907. https://doi.org/10.3390/appliedmath4030048

Academic Editor: Jun-Juh Yan

Received: 21 May 2024
Revised: 16 July 2024
Accepted: 18 July 2024
Published: 29 July 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Voltage source converters (VSCs) have become integral components in modern power systems, playing a crucial role in facilitating efficient energy conversion and enabling flexible power flow control [1]. As the demands for renewable energy integration, grid stabilization, and improved grid reliability continue to grow, the accurate modeling and analysis of VSCs have become paramount [2,3]. The primary power interface connecting distributed generations (DGs) to micro-grids is the AC-DC converter. The voltage source converter functions in two distinct modes, operating as an active rectifier for direct current (DC) bus control and, as depicted in Figure 1, acting as an inverter during grid-feeding and grid-forming operations using distinct control loops [4]. Understanding the basic circuit model of VSCs, along with methodologies for their control and modeling, is essential for engineers and researchers to effectively design, analyze, and optimize VSC-based power electronics applications [5]. The aim of this paper is to present a comprehensive

methodology overview, control strategies, and modeling techniques associated with the basic circuit model of VSCs. This paper delves into the fundamental principles underlying the operation of VSCs, focusing on the key components and their interactions within the circuit. An emphasis is placed on elucidating different methodologies, such as space vector pulse-width modulation (SVPWM), the direct quadrature zero synchronous reference frame, and their integration into VSC circuits.

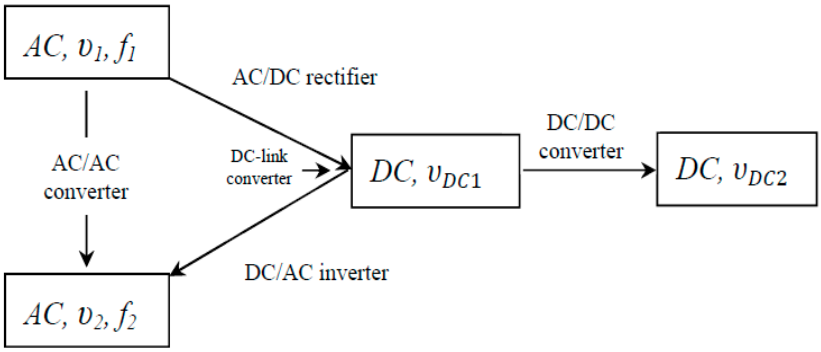


Figure 1. Solid-state power converter groups based on their operating transformation [6].

A fundamental voltage source converter circuit model connected to the LC filter is crucial because it helps in reducing harmonic distortions and improving the overall power quality in the micro-grid. This ensures a stable and clean power supply, which is essential for sensitive loads and the integration of various renewable energy sources. It also plays a key role in converting DC power from renewable sources like solar panels or batteries to AC power, which is required for most loads and for synchronization with the main grid. The LC filter aids in smoothing out the output, making the conversion process more efficient. Understanding the VSC circuit model and its interaction with the LC filter is essential for designing systems that can seamlessly integrate with the main grid or operate in island mode. This ensures that the micro-grid can handle fluctuations in supply and demand without compromising stability. In essence, the significance of introducing the fundamental VSC circuit model connected to the LC filter lies in its ability to address key challenges in energy conversion, stability, integration, and the management of hybrid renewable micro-grid systems.

Furthermore, this paper explores various control strategies employed in VSCs to regulate the voltage, current, and power flow in both grid-connected and standalone applications. Various control techniques using pulse width modulation (PWM) and proportional integral (PI) control are discussed. Moreover, the modeling aspects of the design and control of VSCs are also introduced to develop accurate and dynamic models that capture the transient and steady-state behavior of these devices. To ensure the relevance and applicability of the presented models in real-world scenarios, practical considerations, such as parameter estimation, validation, and implementation challenges, need to be addressed. Case studies and simulations are recommended to be conducted to shown the efficiency and accuracy of the presented methodologies and control strategies in analyzing VSC-based power electronics applications, such as flexible alternating current transmission systems (FACTS), grid-connected renewable energy systems, and high-voltage direct current (HVDC) transmission systems.

The following sections describe the remaining paper parts: In Section 2, two-level grid-connected VSC modeling with an LC filter is presented. The design of the current control loop in the S-domain is introduced in Section 3. Section 4 presents the design of a DC bus voltage control loop in the S-domain. In Section 5, the design of an AC voltage control loop in the S-domain is provided. Section 6 gives a brief discussion, and Section 7 concludes the paper.

2. Two-Level Grid-Connected VSC Modeling with LC Filter

A typical circuit model of a two-level inverter with an LC filter connected to a three-phase load is presented in Figure 2. The signal applied to terminals a, $\grave{a}$, b, $\grave{b}$, c, and $\grave{c}$ regulates the power of the six switches that form the output, denoted as S1 to S6. It is assumed that switches S1 through S6 are complementary [7]. Typically, a voltage source converter with an LCL filter is regulated by a conventional linear controller augmented by a damping loop to ensure the stability of the system [8]. Furthermore, both the three-phase AC system and the DC-side voltage source can directly receive power from the VSC. Consequently, there are two different modes of operation for the VSC, either the inverter (DC/AC) or rectifier (AC/DC) modes.

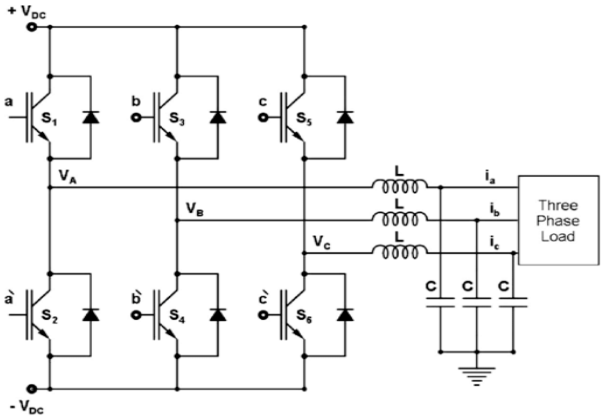


Figure 2. Voltage source converter with three phase-2 levels [7].

To connect the distributed generation systems to the grid, three-phase pulse-width modulation voltage source converters (PWM-VSCs) are commonly employed because of their superior controllability and higher efficiency compared to the other converters [9]. To enable DC-side voltage control and power factor at the point of common coupling (PCC), a two-level PWM-VSC topology is used. However, the implementation of the PWM results in baseband and sideband harmonics appearing in the AC voltage harmonic spectrum, necessitating the use of filters to reduce the harmonics of the grid current [10].

2.1. Space Vector Pulse-Width Modulation (SVPWM) Methodology

Using the triangle comparison strategy simplifies the execution of this procedure. Compared to triangle comparison PWM, it achieves a 15% increase in AC voltage and reduces voltage and current total harmonic distortion (THD). A standard space vector diagram of a two-level voltage source inverter (VSI) with switching states in the six sectors is presented in Figure 3 [11]. Table 1 illustrates that a typical three-phase, two-level voltage source inverter has eight switching states, with six active states producing voltage vectors of either $+V_{dc}$ or $-V_{dc}$, and two null states generating voltage vectors with zero amplitude. In Figure 3, the switch states 000 and 111 represent the null states, while the space vector plane is divided into six equal sectors by the active states. This methodology employs a revolving reference vector, as depicted in Figure 3, sampled once during each sub-cycle, T_s [12]. The main objective is to estimate the vector of the reference voltage V_{ref} utilizing eight switching models.

Figure 3 illustrates that two of the potential output voltage vectors ($\vec{V}_1$ and $\vec{V}_7$) are null or zero vectors, represented by equal values in all three phases (000 or 111). The remaining six vectors ($\vec{V}_1, \ldots, \vec{V}_6$) are non-zero and are all spatially separated by 60°.

The null voltage vector time was arbitrary assumed to be divided equally between t_0 and t_7 for the SVPWM approach, meaning that $T_0 = T_7$.

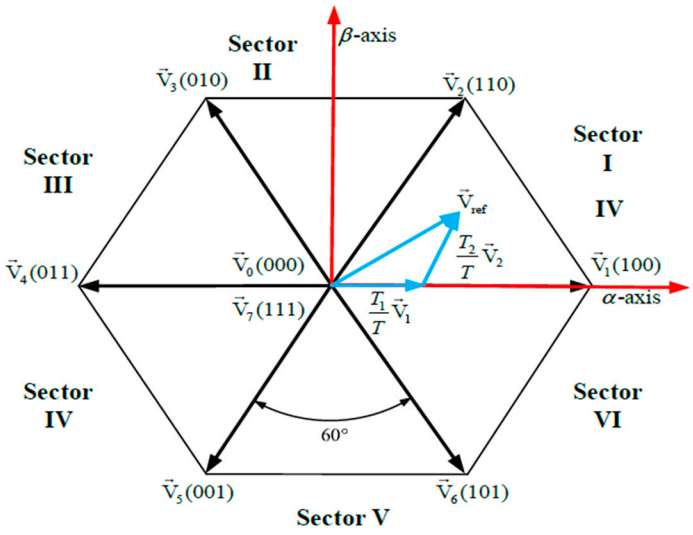


Figure 3. A typical SVPWM diagram for a voltage source inverter with two levels.

Table 1. Phase voltages, output line-to-line voltages, and switching vectors.

Voltage Vectors	Switching Vectors			Line-to-Neutral Voltage			Line-to-Line Voltage		
	S_1	S_2	S_3	V_{an}	V_{bn}	V_{cn}	V_{ab}	V_{bc}	V_{ca}
$\vec{V}_0$	0	0	0	0	0	0	0	0	0
$\vec{V}_1$	1	0	0	$\frac{2}{3}V_{DC}$	$-\frac{1}{3}V_{DC}$	$-\frac{1}{3}V_{DC}$	V_{DC}	0	$-V_{DC}$
$\vec{V}_2$	1	1	0	$\frac{1}{3}V_{DC}$	$\frac{1}{3}V_{DC}$	$-\frac{2}{3}V_{DC}$	0	V_{DC}	$-V_{DC}$
$\vec{V}_3$	0	1	0	$-\frac{1}{3}V_{DC}$	$\frac{2}{3}V_{DC}$	$-\frac{1}{3}V_{DC}$	$-V_{DC}$	V_{DC}	0
$\vec{V}_4$	0	1	1	$-\frac{2}{3}V_{DC}$	$\frac{1}{3}V_{DC}$	$\frac{1}{3}V_{DC}$	$-V_{DC}$	0	V_{DC}
$\vec{V}_5$	0	0	1	$-\frac{1}{3}V_{DC}$	$-\frac{1}{3}V_{DC}$	$\frac{2}{3}V_{DC}$	0	$-V_{DC}$	V_{DC}
$\vec{V}_6$	1	0	1	$\frac{1}{3}V_{DC}$	$-\frac{2}{3}V_{DC}$	$\frac{1}{3}V_{DC}$	V_{DC}	$-V_{DC}$	0
$\vec{V}_7$	1	1	1	0	0	0	0	0	0

The following is used to calculate the sampling interval:

$$T_s = T_0 + T_1 + T_2 \tag{1}$$

or

$$T_S = T_0 + T_{mod} \tag{2}$$

with

$$T_{mod} = T_1 + T_2 \tag{3}$$

where T_{mod} is the modulation time; T_0, T_1, and T_2 are the time intervals; and T_s is the sampling interval. When examining the first sector, the reference voltage vector $\vec{V}_{ref}$ may be presented as function vectors:

$$\vec{V}_{ref} = \frac{T_1}{T_S}\vec{V}_1 + \frac{T_2}{T_S}\vec{V}_2 \tag{4}$$

$$\left|\vec{V}_{ref}\right|\begin{bmatrix}\cos\alpha \\ \sin\alpha\end{bmatrix} = \frac{T_1}{T_S}\left|\vec{V}_1\right|\begin{bmatrix}\cos(0) \\ \sin(0)\end{bmatrix} + \frac{T_2}{T_S}\left|\vec{V}_2\right|\begin{bmatrix}\cos(\pi/3) \\ \sin(\pi/3)\end{bmatrix} \tag{5}$$

Equation (5) can be divided into two parts and expressed as follows:

$$T_S\left|\vec{V}_{ref}\right| \cos\alpha = T_1\left|\vec{V}_1\right|\cos(0) + T_2\left|\vec{V}_2\right|\cos(\pi/3) \tag{6}$$

$$T_S\left|\vec{V}_{ref}\right| \sin\alpha = T_1\left|\vec{V}_1\right|\sin(0) + T_2\left|\vec{V}_2\right|\sin(\pi/3) \tag{7}$$

The sample interval equations are obtained by solving for T_2 using Equation (7), and then substituting the results in Equation (6).

$$T_1 = mT_s\frac{\sin\left(\frac{\pi}{3}-\alpha\right)}{\sin\left(\frac{\pi}{3}\right)} \tag{8}$$

$$T_2 = mT_s\frac{\sin\left(\alpha\right)}{\sin\left(\frac{\pi}{3}\right)} \tag{9}$$

$$T_0 = T_S - T_1 - T_2 \tag{10}$$

The modulation index, $m = \frac{V_{ref}}{V_{dc}}$, is represented as $0 \leq \alpha \leq 60°$, and T_S is the half-period for switching carriers. The following formulas represent the dwell periods for each sector:

$$T_1 = mT_S\sin\left(\frac{n\pi}{3}-\alpha\right) \tag{11}$$

$$T_2 = mT_s\sin\left(\alpha - \frac{\pi(n-1)}{3}\right) \tag{12}$$

Figure 4 shows the SVPWM output switching sequence in the first sector.

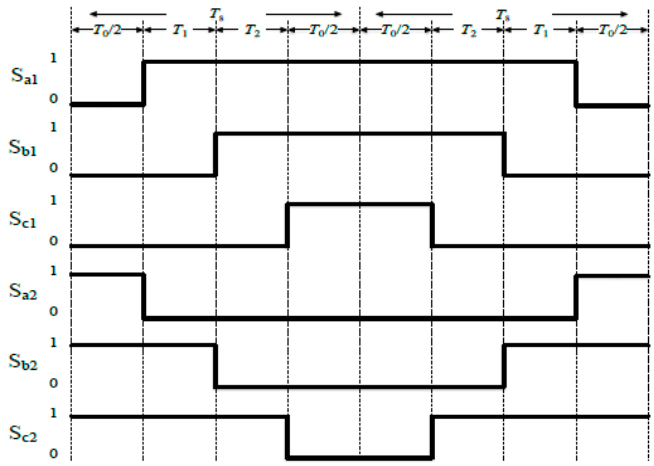


Figure 4. Output switching sequence at the first sector.

The switching periods for the upper and lower switches for each sector are summarized in Table 2 [13].

Table 2. Switching time calculation for every sector.

Sectors	**Upper Switches (S_1,S_3,S_5)**	**Lower Switches (S_4, S_6,S_2)**
Sector 1	$S_1 = T_1 + T_2 + T_0/2$ $S_3 = T_2 + T_0/2$ $S_5 = T_0/2$	$S_4 = T_0/2$ $S_6 = T_1 + T_0/2$ $S_2 = T_1 + T_2 + T_0/2$
Sector 2	$S_1 = T_1 + T_0/2$ $S_3 = T_1 + T_2 + T_0/2$ $S_5 = T_0/2$	$S_4 = T_2 + T_0/2$ $S_6 = T_0/2$ $S_2 = T_1 + T_2 + T_0/2$

Table 2. *Cont.*

Sectors	Upper Switches (S_1,S_3,S_5)	Lower Switches (S_4, S_6,S_2)
Sector 3	$S_1 = T_0/2$ $S_3 = T_1 + T_2 + T_0/2$ $S_5 = T_2 + T_0/2$	$S_4 = T_1 + T_2 + T_0/2$ $S_6 = T_0/2$ $S_2 = T_1 + T_0/2$
Sector 4	$S_1 = T_0/2$ $S_3 = T_1 + T_0/2$ $S_5 = T_1 + T_2 + T_0/2$	$S_4 = T_1 + T_2 + T_0/2$ $S_6 = T_2 + T_0/2$ $S_2 = T_0/2$
Sector 5	$S_1 = T_2 + T_0/2$ $S_3 = T_0/2$ $S_5 = T_1 + T_2 + T_0/2$	$S_4 = T_1 + T_0/2$ $S_6 = T_1 + T_2 + T_0/2$ $S_2 = T_0/2$
Sector 6	$S_1 = T_1 + T_2 + T_0/2$ $S_3 = T_0/2$ $S_5 = T_1 + T_0/2$	$S_4 = T_0/2$ $S_6 = T_1 + T_2 + T_0/2$ $S_2 = T_2 + T_0/2$

2.2. Direct Quadrature Zero Synchronous Reference Frame Methodology

The Direct Quadrature Zero (DQZ) synchronous reference frame methodology is a control technique widely used in power electronics, particularly in VSIs, to achieve the precise and efficient control of three-phase AC systems [14]. This methodology is widely applied in grid-connected renewable energy systems (such as wind and solar inverters), adjustable speed drives for motors, active power filters for harmonic mitigation, and other applications requiring the precise control of the AC voltage and the current waveform [15]. The DQZ synchronous reference frame methodology involves transforming the three-phase AC signals (typically voltages or currents) into a rotating reference frame aligned with the AC frequency of the system. This transformation simplifies the control of the VSIs by decoupling the components of interest (such as active power, reactive power, and harmonic components) from each other.

The three-phase AC signals are transformed from the time domain (a, b, c reference frame) to the rotating DQZ reference frame using Clarke and Park transformations [16]. Park transformation rotates the $\alpha\beta$ components to a stationary DQZ reference frame (d, q, 0), where d and q components represent the direct and quadrature components of the signals, respectively, and 0 represents the zero sequence component [17]. Clarke transformation converts three-phase quantities (a, b, c) into two-phase quantities (α, β), representing the symmetrical components of the AC signals. In the DQZ reference frame, control strategies such as proportional integral (PI) controllers or more advanced algorithms are applied to regulate the d and q components of the reference signals. These controllers adjust the switching patterns of the VSI's semiconductor devices typically insulated gate bipolar transistors (IGBTs) to generate the desired AC output voltage waveform, ensuring the precise control of the voltage magnitude, frequency, and waveform quality [18]. Figure 5 shows a space phasor with the $\alpha\beta0$ reference frame.

The revised three-phase current values (i_a, i_b, i_c) are shown in Figure 5 as projections onto the new reference axis (i_α, i_β). When a three-phase, three-wire system is balanced and the zero axis represents the common mode component, the total phase current is nullified. Thus, the variables α and β can fully characterize a system defined in the ABC reference frame.

The output variables ABC are described in Equation (13) below:

$$i_a = I_M\cos(\omega t);\ i_b = I_M\cos\left(\omega t - \frac{2\pi}{3}\right);\ i_c = I_M\cos\left(\omega t + \frac{2\pi}{3}\right) \tag{13}$$

where t is the amount of time needed to reach steady state and ω is the constant synchronous frequency. Equation (14) expresses the $\alpha\beta0$ transformation in complex form, while Equation (15) expresses it in matrix form.

$$\vec{I_s} = i_\alpha + ji_\beta = \frac{2}{3}\left(i_a + i_b e^{\frac{j2\pi}{3}} + i_c e^{\frac{-j2\pi}{3}}\right) \tag{14}$$

$$\begin{bmatrix} i_\alpha \\ i_\beta \\ i_0 \end{bmatrix} = \frac{2}{3}\begin{bmatrix} 1 & \frac{-1}{2} & \frac{-1}{2} \\ 0 & \frac{\sqrt{3}}{2} & \frac{-\sqrt{3}}{2} \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \end{bmatrix}\begin{bmatrix} i_a \\ i_b \\ i_c \end{bmatrix} \tag{15}$$

The reference axis currents are i_α, i_β, and i_0, while the three-phase currents are i_α, i_b, and i_c.

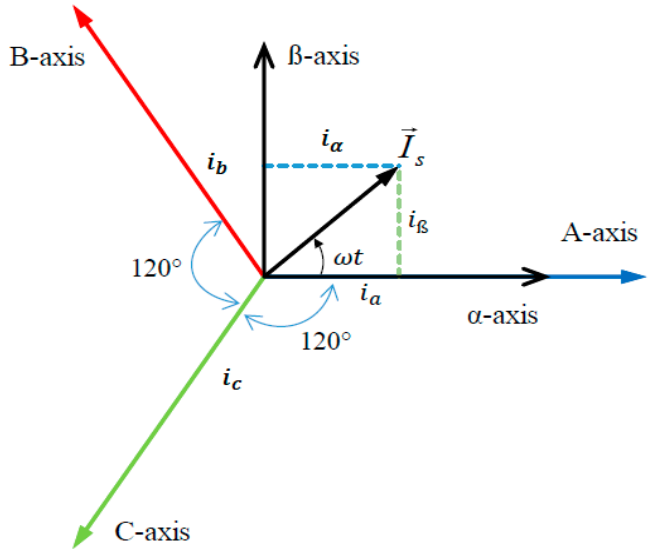


Figure 5. Space phasor and $\alpha\beta 0$ reference frame.

Using Park transformation with a $-90°$ shift, the frame rotates around the 0 axis at the same frequency as the sinusoids defining the phasors, while the system is in the $\alpha\beta 0$ frame. As it can be seen in Figure 6, the component is transformed to Q axis, which is positioned 90° at a quadrature angle to the direct component, and the $-\beta$ component is transformed to the d-axis, which is in line with the rotating vector.

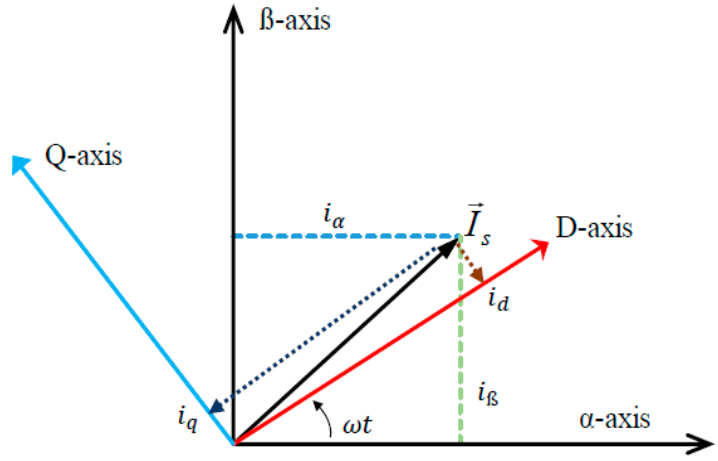


Figure 6. Space phasor and DQ0 reference frame.

Equations (16) and (17) express the DQ0 transformation in complex and matrix representations, respectively.

$$\vec{I}_s = i_d + ji_q = (i_\alpha + ji_\beta)e^{-j(\omega t - \frac{\pi}{2})} \tag{16}$$

$$\begin{bmatrix} i_d \\ i_q \\ i_0 \end{bmatrix} = \begin{bmatrix} \sin(\omega t) & -\cos(\omega t) & 0 \\ \cos(\omega t) & \sin(\omega t) & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} i_\alpha \\ i_\beta \\ i_0 \end{bmatrix} \quad (17)$$

Consequently, the rotation generates DC values from the periodic signals. Since the spinning frame trails the a-axis by 90 degrees, at $t = 0$, the Q- and A-axes coincide. When I_s is precisely aligned with the reference angle ωt, $i_d = 1$, $i_q = 0$, and $i_0 = 0$ are the DQ0 components. The full transformations from DQ0 to ABC and their corresponding inverses are presented as follow:

$$\begin{bmatrix} i_d \\ i_q \\ i_0 \end{bmatrix} = \begin{bmatrix} \sin(\omega t) & \sin(\omega t - \frac{2\pi}{3}) & \sin(\omega t + \frac{2\pi}{3}) \\ \cos(\omega t) & \cos(\omega t - \frac{2\pi}{3}) & \cos(\omega t + \frac{2\pi}{3}) \\ \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} i_a \\ i_b \\ i_c \end{bmatrix} \quad (18)$$

$$\begin{bmatrix} i_a \\ i_b \\ i_c \end{bmatrix} = \begin{bmatrix} \sin(\omega t) & \cos(\omega t) & 1 \\ \sin(\omega t - \frac{2\pi}{3}) & \cos(\omega t - \frac{2\pi}{3}) & 1 \\ \sin(\omega t + \frac{2\pi}{3}) & \cos(\omega t + \frac{2\pi}{3}) & 1 \end{bmatrix} \begin{bmatrix} i_d \\ i_q \\ i_0 \end{bmatrix} \quad (19)$$

3. Design of the Control Loop for the Current in the S-Domain

The current control loop design in the S-domain for a three-phase two-level VSI is a crucial aspect of ensuring the precise and stable operation of the inverter. This design methodology leverages the classical control theory in the S-domain to achieve the effective regulation of the output currents of the inverter. According to Wencong et al., this control loop modifies the output voltage of the inverter to allow the injection of the necessary current in the utility grid [19]. The three-phase VSI is typically modeled using a set of differential equations that describe the relationship between the input DC voltage, the switching states of the inverter, and the output AC currents. These equations are transformed into the synchronous reference frame (dq-frame) to decouple the three-phase AC currents into two orthogonal components (d-axis and q-axis), simplifying the control design. The system dynamics are represented by transfer functions in the S-domain, which relate the control inputs (reference voltages) to the outputs (actual currents). For a VSI, the transfer function typically includes the LC filter dynamics of the inverter and the influence of the load impedance. Figure 7 illustrates the schematic representation of a three-phase, two-level VSC with an L filter. R_f and L_f represent the resistance and the inductance, respectively, between the converter switches and the PCC.

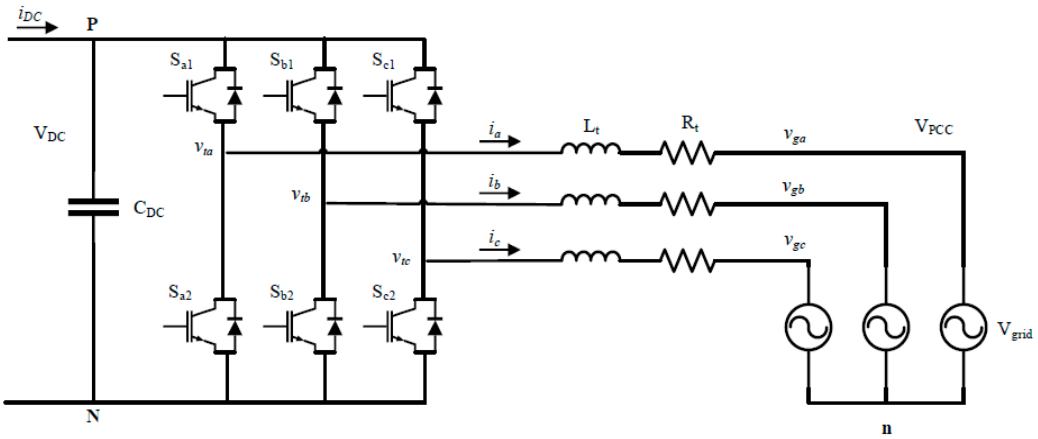


Figure 7. A two-level VSC circuit diagram with an L filter included.

A reduced equivalent per-phase circuit can be produced by applying the voltage law of Kirchhoff to the circuit shown above, as illustrated in Figure 8.

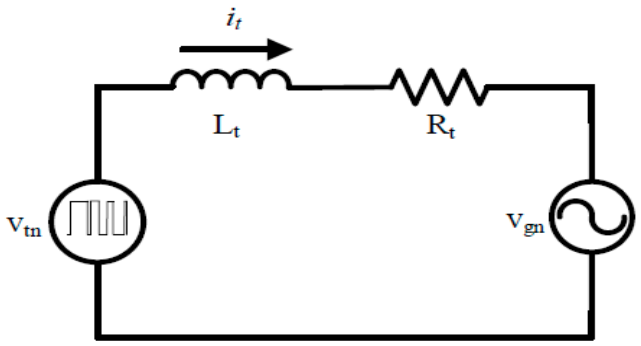


Figure 8. An L filter in a simplified VSC per-phase circuit.

The grid voltage, v_{gn}, is shown in Figure 8 as a relatively fixed component that may be considered disruptive. As a result, the current, i_{tn}, can be regulated by altering the output voltage of the inverter v_{tn}. Consequently, the VSC functions as an energy source that feeds electricity into the grid. The voltage law of Kirchhoff, when applied to the three circuits per phase, yields the following:

$$\begin{aligned} v_{tan} &= L_f \frac{di_a}{dt} + R_f i_a + v_{gan} \\ v_{tbn} &= L_f \frac{di_b}{dt} + R_f i_b + v_{gbn} \\ v_{tcn} &= L_f \frac{di_c}{dt} + R_f i_c + v_{gcn} \end{aligned} \tag{20}$$

The variables v_{tan}, v_{tbn}, and v_{tcn} represent the three per-phase inverter output voltages; L_f and R_f denote the inductance and resistance between the converter switches and the point of common coupling (PCC); v_{gan}, v_{gbn}, and v_{gcn} stand for the three per-phase grid voltages; and i_a, i_b, and i_c represent the three per-phase inverter currents. By consolidating all three equations from (20) into a single equation, we obtain the following:

$$v_{t,abc} = L_f \frac{di_{abc}}{dt} + R_f i_{abc} + v_{g,abc} \tag{21}$$

Equation (15) is used to translate Equation (21) into the reference frame $\alpha\beta 0$, which results in

$$v_{t,\alpha\beta} = L_f \frac{di_{\alpha\beta}}{dt} + R_f i_{\alpha\beta} + v_{g,\alpha\beta} \tag{22}$$

Using the reference frame $dq0$ in (16) and repeating the process for Equation (22), the result is

$$\begin{aligned} v_{t,dq} &= v_{t,\alpha\beta} e^{-j(\omega t - \frac{\pi}{2})} \\ v_{t,dq} L_f \frac{di_{dq}}{dt} &+ R_f i_{dq} + v_{g,dq} \end{aligned} \tag{23}$$

The a, b, c frame can be split into the two components, d and q, as was indicated in Section 2.1.

$$L_f \frac{di_d}{dt} = v_{td} + L_f \omega i_q - R_f i_d - v_{gd} \tag{24}$$

$$L_f \frac{di_q}{dt} = v_{tq} + L_f \omega i_q - R_f i_q - v_{gq} \tag{25}$$

with $v_{td} = \frac{V_{DC}}{\sqrt{3}} m_d$ and $v_{tq} = \frac{V_{DC}}{\sqrt{3}} m_q$, where m_{dq} is the index of modulation, and $\frac{V_{DC}}{\sqrt{3}}$ is the SVPWM gain converter.

The i_d and i_q dynamics are related because of the $L_f \omega$ term found in Equations (24) and (25). To decouple the d and q subsystems and account for the grid disruption inputs v_{gd} and v_{gq}, the index of modulation for each must be stated as follows:

$$m_d = \frac{\sqrt{3}}{V_{DC}} \left(u_d - L_f \omega i_q + v_{gd} \right) \tag{26}$$

$$m_q = \frac{\sqrt{3}}{V_{DC}}\left(u_q - L_f\omega i_d + v_{gq}\right) \tag{27}$$

For the d subsystem, the anticipatory term is $-L_f\omega i_q + v_{gd}$, while for the q subsystem, it is $-L_f\omega i_d + v_{gq}$. Equation (26) can be substituted into (24), yielding

$$L_f\frac{di_d}{dt} = u_d - R_f i_d \tag{28}$$

Equation (27) can be substituted into (25) in the same way to obtain

$$L_f\frac{di_q}{dt} = u_q - R_f i_q \tag{29}$$

with u_q as the output of PI compensator.

The control systems of the current control loops are the same in the d and q axes, as seen in (28) and (29). Therefore, the linked compensators can be the same. As a result, only the current control compensator gains in the d-axis will be determined. When Equation (29) is transformed using the Laplace method, the following results are obtained:

$$\begin{array}{c} L_f s I_d(s) = u_d(s) - R_f I_d(s) \\ I_d(s) = \frac{u_d(s)}{L_f + R_f} \end{array} \tag{30}$$

The control plant can therefore be expressed as follows:

$$G_P(s) = \frac{1}{L_f s + R_f} \tag{31}$$

The following Equation (32) indicates that the compensator is a PI controller:

$$G_c(s) = \frac{k_p s + k_i}{s} \tag{32}$$

Figure 9 illustrates the transfer function of the closed loop in the d-axis:

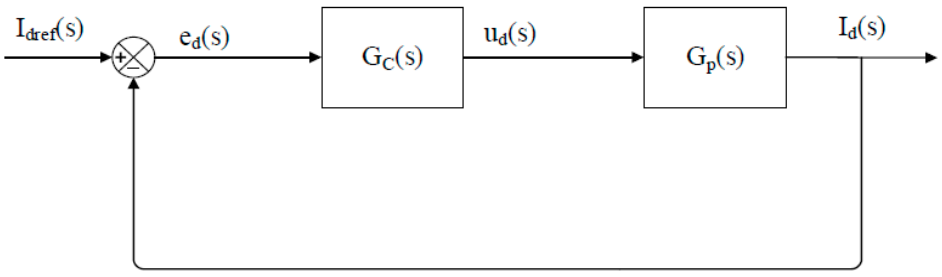


Figure 9. Current-controlled VSC system block diagram in simplified form.

Equation (33) gives the following open loop gain:

$$G_{ol}(s) = G_C(s)G_P(s) \tag{33}$$

By replacing the compensator and control plant Equations by their terms, Equation (33), will be:

$$G_{ol}(s) = \frac{k_p s + k_i}{s} \times \frac{1}{L_f s + R_f} = \frac{k_p s + k_i}{L_f s^2 + R_f s} = \left(\frac{k_p}{L_f s}\right)\frac{s + \frac{k_i}{k_p}}{s + \frac{R_f}{L_f}} \tag{34}$$

Once the terms $\frac{k_i}{k_p} = \frac{R_f}{L_f}$ are equalized by applying the pole cancelation approach, the open-loop gain Equation (34) is reduced from the second to the first order transfer function as follows:

$$G_{ol}(s) = \frac{k_p}{L_f s} \tag{35}$$

Equation (36) represents the current control gain in the closed loop in the *d*-axis:

$$G_i(s) = G_{cl}(s) = \frac{G_{ol}(s)}{1 + G_{ol}(s)} \tag{36}$$

Equation (35) can be substituted into (36) to obtain the following expression for the closed-look gain:

$$\frac{I_d(s)}{I_{dref}(s)} = G_i(s) = \frac{k_p}{k_p + L_f s} \tag{37}$$

By replacing $k_p = \frac{L_f}{\tau_i}$ and $k_i = \frac{R_f}{\tau_i}$, the closed-look gain will become

$$G_i(s) = \frac{1}{\tau_i s + 1} \tag{38}$$

The resultant closed-loop control constant time is written as τ_i. Therefore, the following formula provides the first-order settling time in the closed-loop system:

$$t_{i,\ settling} = 4.6\tau_i \tag{39}$$

4. Design of the Control Loop for the DC Bus Voltage in the S-Domain

For maintaining balanced power flow, the DC bus voltage control technique is essential. Implementing this control is imperative as it enables the voltage source converter to function as a rectifier, converting AC power to DC power and facilitating control of the DC link capacitor voltage to supply an output load. Moreover, for intermittent DC power sources such as PV systems, this control technique assists in adjusting the current control loop reference to regulate the input of DC power effectively [20]. The circuit from Figure 7 is shown in Figure 10, with a bleeding resistor R_B and extra feedback signals (i_{DC} and i_L) added to simplify DC control modeling.

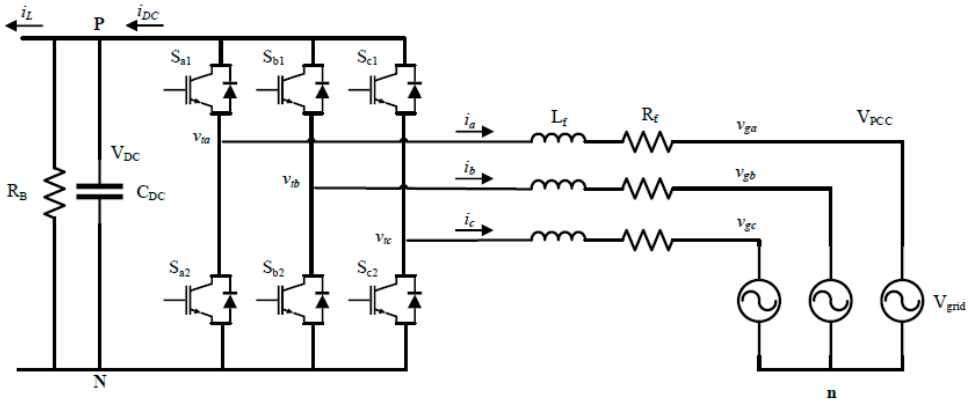


Figure 10. Two-level VSC with an L filter and a DC link circuit bleed resistor.

The resistance and inductance between the PCC and the inverter output are R_f and L_f. The fact that the active power transfer enables to charge and discharge the DC link capacitor, the circuit shown in the DQ synchronous reference frame configuration can only be modeled for the *d*-axis component. As a result, an instantaneous reactive current is not needed, neither the *q*-axis component for this procedure, as Figure 11 shows.

The following equations illustrate the voltage dynamic equation of the DC link capacitor:

$$\begin{aligned} i_{DC} &= i_{bus} + i_{RB} + i_L \\ i_{bus} &= i_{DC} - i_{RB} - i_L \\ C_{DC}\frac{dv_{DC}}{dt} &= i_{DC} - \frac{v_{DC}}{R_B} - i_L \end{aligned} \tag{40}$$

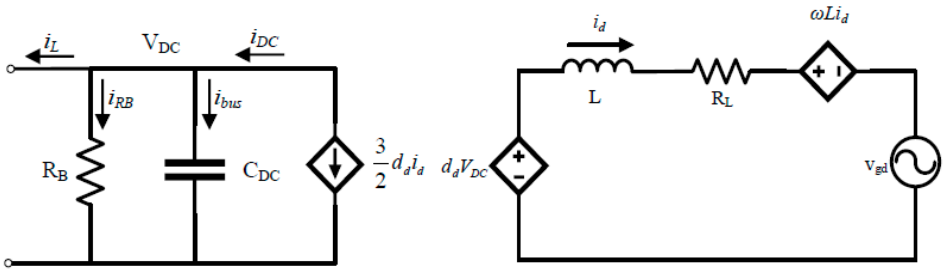


Figure 11. VSC *d*-axis component configuration with a bleed resistor and a L filter.

Equation (40) can be transformed using the Laplace transform to obtain

$$\begin{gathered}\left[sC_{DC}+\frac{1}{R_B}\right]V_{DC}(s)=I_{DC}(S)-I_L(s)\\ V_{DC}(s)=\frac{1}{\left[sC_{DC}+\frac{1}{R_B}\right]}u_d(s)=\frac{R_B}{sC_{DC}R_B+1}u_d(s)\end{gathered} \tag{41}$$

The formula for calculating the PI compensator output, $u_d(s)$, is $u_d(s) = I_{DC}(s) - I_L(s)$, where $I_L(s)$ is the feedback term. This allows for the controller to avoid making up a measured value. Considering this, the control system is written as follows:

$$G_P = \frac{R_B}{sC_{DC}R_B+1} \tag{42}$$

The relationship between these currents is represented as follows, assuming a converter without losses and the fact that the i_{DC} can only be directly controlled by the current pulled from the grid:

$$P_{3\varnothing} = P_{DC} \tag{43}$$

The active power is given as

$$P_{3\varnothing} = \frac{3}{2}\left(v_{gd}i_d + v_{gq}i_q\right) \tag{44}$$

The DC bus power is obtained by replacing Equation (44) with (43), which yields the following:

$$\frac{3}{2}v_{gd}i_d = -i_{DC}V_{DC} \tag{45}$$

The following Equation describes the relationship between the time constant τ_{DC} of the control loop for the DC bus voltage and the time constant τ_i of the current control loop:

$$\tau_i \ll \tau_{DC} \tag{46}$$

The relationship between these two control loops may be expressed by the following Equation (47) if the condition in Equation (46) is satisfied, $i_d = I_{dref}$, and the two control loops are decoupled.

$$I_{dref} = -\frac{2}{3}\frac{V_{DC}}{v_{gd}}I_{DC} \tag{47}$$

The inner gain coefficient is obtained from the relationship between I_{dref} and I_{DC} and is stated as follows:

$$K_{inner} = \frac{I_{DC}}{I_{dref}} = -\frac{3}{2}\frac{v_{gd}}{V_{DC}} \tag{48}$$

Based on the following, it is presumed that the compensator is a PI controller expressed as

$$G_C(s) = \frac{k_p s + k_i}{s} \tag{49}$$

The selected method for DC voltage control adopts a cascade approach, including an outer loop for the voltage control of the DC bus and an inner loop for the current control. This tactic guarantees that the time reaction of current controller is faster than that for

the controller of the DC voltage in accordance with Equation (46). The DC voltage loop considers the current loop as a constant gain, which permits a gradual shift in the current reference, and hence an error-free response from the current controller. Consequently, the dynamics of the two circuits depicted in Figure 11 are effectively disconnected. Figure 12 shows a DC link voltage loop block diagram of a VSC system in a simplified form.

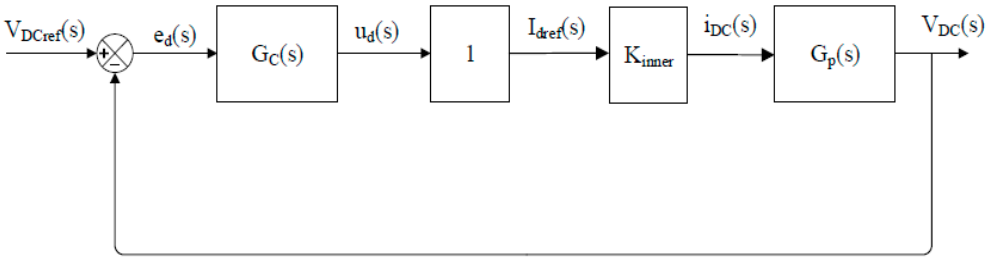


Figure 12. Simplified DC link voltage loop model of the VSC system.

Equation (50) provides the gain in the open loop based on Figure 12.

$$G_{ol}(s) = G_C(S)G_P(S)K_{inner} \tag{50}$$

Equations (42) and (49) can be substituted into (50) to obtain the following:

$$G_{ol}(s) = \frac{k_p s + k_i}{s} \times \frac{R_B}{sC_{DC}R_B + 1} \times K_{inner} = \frac{K_{inner}R_B(k_p s + k_i)}{s^2 C_{DC}R_B + s} \tag{51}$$

With respect to the current control in *d*-axis, the closed loop gain is expressed in Equation (52):

$$G_{DC}(S) = G_{cl}(s) = \frac{G_{ol}(s)}{1 + G_{ol}(s)} \tag{52}$$

Equation (51) substituted into (52) will result in

$$G_{DC}(s) = \frac{K_{inner}(k_p s + k_i)/C_{DC}}{S^2 + \left(\frac{K_{inner}R_B k_p + 1}{C_{DC}R_B}\right)s + \frac{K_{inner}k_i}{C_{DC}}} \tag{53}$$

Seeing that the Equation (53) has a first-order numerator, a pre-filter will be established to eliminate this term, enabling the alignment of (53) with the standard second-order transfer function, which is:

$$G_{PF}(s) = \frac{G_{DC,\,desired}(s)}{G_{DC,old}(s)} = \frac{k_i}{(k_p s + k_i)} \tag{54}$$

The gain in closed loop of the current control in *d*-axis can be obtained by multiplying Equations (53) and (54). The result is as follows:

$$G_{DC}(s) = \frac{\frac{K_{inner}k_i}{C_{DC}}}{s^2 + \left(\frac{K_{inner}R_B k_p + 1}{C_{DC}R_B}\right)S + \frac{K_{inner}k_i}{C_{DC}}} \tag{55}$$

The gain in closed loop of the current control in *d*-axis can be obtained by multiplying Equations (53) and (54). The result is as follows:

$$k_p = \frac{2\xi\omega_n C_{DC}R_B - 1}{K_{inner}R_B} \tag{56}$$

$$k_i = \frac{{\omega_n}^2 C_{DC}}{K_{inner}} \tag{57}$$

Thus, the following is the second closed-loop system settling time:

$$\begin{aligned} \tau_{DC} &= \frac{1}{\xi\omega_n} \\ t_{DC,setting} &= 4.6 \times \tau_{DC} \end{aligned} \tag{58}$$

5. Design of the Control Loop for the AC Voltage in the S-Domain

When a fault or power outage prevents the main grid from supplying power, a VSC system in grid-forming mode creates a grid reference and makes it easier to transfer active or reactive powers during micro-grid operation in isolated mode. This enables the micro-grid to continue operating [21]. The three-phase VSI is modeled using differential equations that describe the relationship between the input DC voltage, the switching states of the inverter, and the output AC voltages. The dynamics of the VSI system, including the LC filter (which is typically used to smooth the inverter output), as presented in Figure 13, are represented in the S-domain by transfer functions [22].

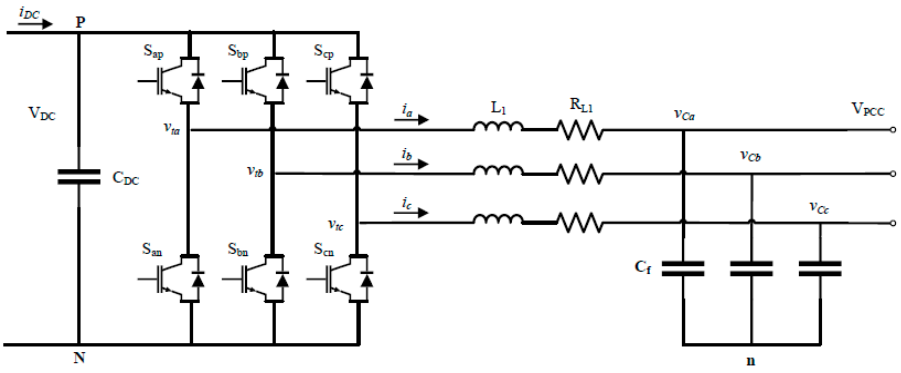


Figure 13. Two level VSC with an LC filter.

The resistances between the converter and the filter capacitor are presented as R_{L1} and L_1. A condensed per-phase equivalent circuit produced by using the voltage law of Kirchhoff is shown in Figure 14.

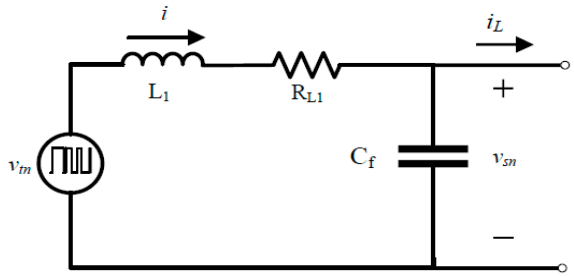


Figure 14. A VSC with an LC filter per-phase circuit in a simplified.

The phase currents for each of the three per-phase circuits can be obtained by using the current law of Kirchhoff as follows:

$$\begin{aligned} i_a &= i_{Cfa} + i_{La} \\ i_b &= i_{Cfb} + i_{Lb} \\ i_c &= i_{Cfc} + i_{Lc} \end{aligned} \tag{59}$$

Combining the three equations from (59) results in the following:

$$i_{abc} = i_{Cf,abc} + i_{L,abc} \tag{60}$$

Using the ABC-$\alpha\beta$0 transform on Equation (60), the following result is obtained:

$$i_{Cf,\alpha\beta} = i_{\alpha\beta} + i_{L,\alpha\beta} \tag{61}$$

Using Equation (16), the $\alpha\beta$0-dq0 transform can be converted to

$$i_{Cf,\,dq} = i_{Cf,\alpha\beta} e^{-j(\omega t - \frac{\pi}{2})} \tag{62}$$

$$i_{C,dq} = i_{dq} - i_{L,dq} \tag{63}$$

Consequently, as described in Section 2.1, the a, b, c frame reference may be divided into d and q components and cascaded via the current control loops in d and q axes, resulting in

$$C_f \frac{dv_{sd}}{dt} = C_f \omega v_{sq} + i_d - i_{Ld} \tag{64}$$

$$C_f \frac{dv_{sq}}{dt} = -C_f \omega v_{sd} + i_q - i_{Lq} \tag{65}$$

The existence of the $C_f\omega$ terms in Equations (64) and (65) causes the v_{sd} and v_{sq} dynamics to be coupled. The feedforward terms $\left(-C_f \omega v_{sq} + i_{Ld}\right)$ for the d-axis component and $\left(C_f \omega v_{sd} + i_{Lq}\right)$ for the q-axis component are introduced in order to decouple them and account for the load disturbance inputs i_{Ld} and i_{Lq}. Furthermore, the control units of the d-axis and q-axis components are the same, as shown in Equations (64) and (65); thus,

$$\frac{dV_{sd}}{dt} = \frac{1}{C_f}[u_d] \tag{66}$$

u_d is the PI compensator output. Equation (66) is solved by applying the Laplace transform, which yields the following:

$$V_{sd}(s) = \frac{u_d(s)}{sC_f} \tag{67}$$

As a result, the following is the control plant:

$$G_P(s) = \frac{1}{sC_f} \tag{68}$$

Furthermore, it is presumed that the control loop time constant of the AC voltage, τ_v, will be five times greater than that of the current control loop, τ_i. If this assumption is met, the two control loops will be decoupled. The circumstance is as follows:

$$\tau_i \ll \tau_v \tag{69}$$

Therefore, the inputs of the current control loops are presented in Equations (70) and (71), and $I_{dq}(s) = I_{dqref}(s)$ for the control loop of AC voltage.

$$I_{dref}(s) = u_d(s) - C_f \omega V_{sq}(s) + I_{Ld}(s) \tag{70}$$

$$I_{qref}(s) = u_q(s) + C_f \omega V_{sd}(s) + I_{Lq}(s) \tag{71}$$

Equation (72) gives the PI controller gain.

$$G_C(s) = \frac{k_p s + k_i}{s} \tag{72}$$

Since the control plants of Equation (68) for the d and q components are the same, the corresponding compensators may also be the same. With reference to Figure 15, the only information that needs to be collected next is the gains of the voltage control compensators in d-axis. The following is the open loop gain:

$$G_{ol}(s) = G_C(s) G_P(s) \tag{73}$$

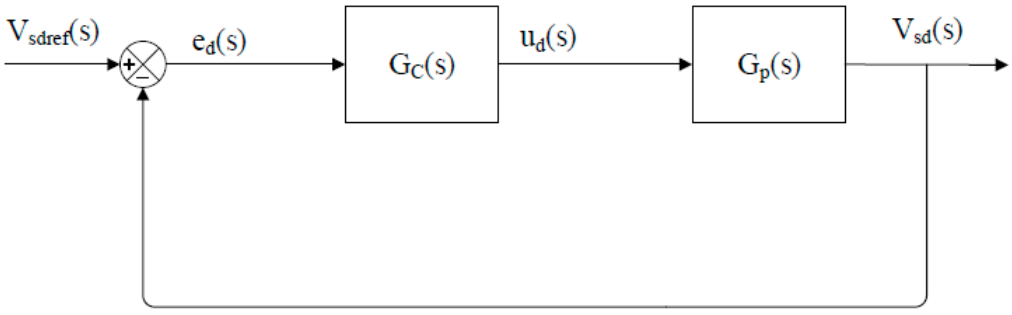


Figure 15. A block diagram for the VSC system AC voltage loop in a simplified form.

$G_C(s)$ and $G_P(s)$ can be substituted with their corresponding values to obtain the following result:

$$G_{ol}(s) = \frac{k_p s + k_i}{s} \times \frac{1}{sC_f} = \frac{k_p s + k_i}{s^2 C_f} \tag{74}$$

The gain of the current control in *d*-axis in closed loop can be represented as follows:

$$G_v(s) = G_{cl}(s) = \frac{G_{ol}(s)}{1 + G_{ol}(s)} \tag{75}$$

When Equation (74) is substituted into (75), this results in

$$G_v(s) = \frac{\frac{(k_p s + k_i)}{C_f}}{s^2 + \left(\frac{k_p}{C_f}\right)s + \frac{k_i}{C_f}} \tag{76}$$

The VSC system AC voltage loop block diagram in a simplified form is shown in Figure 15.

Given that the first-order numerator of Equation (76) must be cancelled, a pre-filter needs to be developed in order to allow for Equation (76) to correspond to the conventional transfer function of the second order. As stated below, a pre-filter is required.

$$G_{PF}(s) = \frac{G_{DC,\,desired}(s)}{G_{DC,\,old}(s)} = \frac{k_i}{k_p s + k_i} \tag{77}$$

The following is the result of multiplying Equations (76) and (77):

$$G_v(s) = \frac{k_i}{s^2 C_f + k_p s + k_i} \tag{78}$$

By resolving the results of the denominator coefficients of Equation (78), the gains of the proportional and integral controllers are solved for

$$k_p = 2\xi\omega_n C_f \tag{79}$$

$$k_i = {\omega_n}^2 C_f \tag{80}$$

As a result, the control loop time constant of the AC voltage is provided as

$$\tau_v = \frac{1}{\xi\omega_n} \tag{81}$$

To decouple the two control loops, the system settling time for this second order in the closed loop is considered to be five times longer than the existing control loop settling time. In other words,

$$t_{v,settling} = \frac{4.6}{\xi\omega_n} = 4.6 \times 5 \times \tau_i = 5 \times t_{i,settling} \quad (82)$$

6. Discussion

This Discussion section explores the key aspects of the basic circuit model of VSCs. The different methodologies and control strategies employed in VSCs, including the SVPWM methodology, the DQZ synchronous reference frame methodology, the design of the current, the DC bus voltage, and the AC voltage control loops in the S-domain, are presented in Table 3.

Table 3. Summary of the basic circuit model of voltage source converters.

SVPWM methodology	SVPWM is a widely used technique for controlling VSCs because it can produce excellent output waveforms with less harmonic distortion. By effectively synthesizing the desired output voltage vector, SVPWM enables the precise control of the VSC output. However, challenges, such as computational complexity and switching frequency constraints, need to be carefully addressed to ensure an optimal performance.
DQZ synchronous reference frame methodology	The DQ0 synchronous reference frame methodology provides a powerful tool for analyzing and controlling VSCs, particularly in grid-connected applications. By transforming the three-phase system into a two-dimensional space with direct (d) and quadrature (q) axes, this methodology simplifies the control and modeling processes. It allows for the accurate representation of the system dynamics and facilitates the design of control algorithms, enhancing the overall performance and stability of VSC-based systems.
Design of the control loop for the current in the S-domain	Effective current control is crucial for ensuring accurate and dynamic response in VSCs, especially in applications requiring the precise regulation of current flow. By designing current control loops in the S-domain, engineers can leverage advanced control techniques to achieve desired performance objectives, such as fast response, low overshoot, and robust stability. Implementing sophisticated control algorithms, such as proportional integral controllers or model predictive control (MPC), can further improve the current control loops efficiency in VSCs.
Design of the control loop for the DC Bus voltage in the S-domain	The control loop for DC bus voltage plays a critical role in maintaining stable operation and ensuring reliable performance of VSCs. By regulating the DC bus voltage, this control loop enables efficient energy conversion and smooth power transfer between the VSC and the grid or load. Designing the DC bus voltage control loop in the S-domain allows for the precise tuning of controller parameters and dynamic response characteristics, ensuring optimal performance under varying operating conditions and load profiles.
Design of the control loop for the AC voltage in the S-domain	In grid-connected VSC applications, the control loop for the AC voltage is essential for regulating the output voltage to meet grid requirements, such as voltage magnitude and frequency. By designing the control loop for the AC voltage in the S-domain, engineers can develop robust control strategies to maintain grid stability, mitigate voltage harmonics, and ensure the seamless integration of renewable energy sources. Advanced control techniques, such as predictive control and adaptive control, can be employed to enhance the AC voltage control loop efficiency and flexibility in VSCs.

7. Conclusions

The basic VSC circuit model with an output LC filter has been presented in this paper. The main power interface between distributed generators and the micro-grid is the AC-DC converter. Two forms of operation for a voltage source converter have been observed. It acts as an active rectifier when utilized as a DC bus control. It uses a variety of control loops to operate as an inverter when in the grid-feeding and grid-forming modes. A thorough manual that covers a variety of techniques has also been introduced, including the SVPWM, the DQZ synchronous frame, and the phase lock loop. In addition, this paper serves as a comprehensive guide to understanding the basic circuit model of VSCs, methodologies for their control, and techniques for their accurate modeling. By providing insights into the fundamental principles and practical considerations associated with VSCs, this paper aims to contribute to the advancement of VSC-based power electronics applications and their integration into modern power systems for enhanced efficiency,

reliability, and sustainability. Case studies and simulations need to be conducted to illustrate the efficacy and accuracy of the presented methodologies and control strategies in analyzing VSC-based power electronics applications.

Author Contributions: Conceptualization; methodology; formal analysis; writing—original draft preparation, C.B.N.; writing—review and editing; supervision.; project administration, M.A. and A.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Research Foundation (NRF) grant number PMDS22052313773 And The APC was funded by Cape Peninsula University of Technology.

Institutional Review Board Statement: This research does not involve human or animal subjects; hence, participant consent was not necessary in order to collect data for this project.

Data Availability Statement: Requests for access to the datasets can be made to the corresponding author.

Acknowledgments: As the Corresponding Author, I acknowledge that this paper has been completed as part of doctoral study research under the supervision and support of Marco Adonis, and Ali Almaktoof provided direction, editing, and reviewed the manuscript before publication.

Conflicts of Interest: There are no disclosed conflicts of interest pertaining to this publication.

References

1. Jing, T.; Maklakov, A.S. A Review of Voltage Source Converters for Energy Applications. In Proceedings of the 2018 International Ural Conference on Green Energy (UralCon), Chelyabinsk, Russian Federation, 4–6 October 2018; pp. 275–281.
2. Brodzicki, M.; Klucznik, J.; Czapp, S. Evaluation of VSC Impact on Power System Using Adequate P-Q Capability Curve. *Electronics* **2023**, *12*, 2462. [CrossRef]
3. Ndeke, C.B.; Adonis, M.; Almaktoof, A. Energy management strategy for a hybrid micro-grid system using renewable energy. *Discov. Energy* **2024**, *4*, 1. [CrossRef]
4. Monteiro, V.; Monteiro, L.F.; Franco, F.L.; Mandrioli, R.; Ricco, M.; Grandi, G.; Afonso, J.L. The role of front-end ac/dc converters in hybrid ac/dc smart homes: Analysis and experimental validation. *Electronics* **2021**, *10*, 2601. [CrossRef]
5. Xiong, L.; Liu, X.; Liu, Y.; Zhuo, F. Modeling and Stability Issues of Voltage-source Converter-dominated Power Systems: A Review. *CSEE J. Power Energy Syst.* **2022**, *8*, 1530–1549.
6. Almaktoof, A.M.A. Multilevel Inverters Using Finite Set-Model Predictive Current Control for Renewable Energy Systems Applications. Ph.D. Thesis, Cape Peninsula University, Cape, South Africa, March 2015.
7. Gaballah, M.; El-Bardini, M. Low cost digital signal generation for driving space vector PWM inverter. *Ain Shams Eng. J.* **2013**, *4*, 763–774. [CrossRef]
8. Jeong, M.; Gfrörer, T.; Biela, J. Extending Stable Parameter Range of LCL-filters for Grid-Connected Converters by Inherent Damping of Model Predictive Control (MPC). In Proceedings of the 2022 IEEE 23rd Workshop on Control and Modeling for Power Electronics (COMPEL), Tel Aviv, Israel, 20–23 June 2022.
9. Bernet, D.; Hiller, M. Grid-connected medium-voltage converters with parallel voltage-source active filters. *IET Electr. Power Appl.* **2019**, *13*, 1507–1513. [CrossRef]
10. Kouchaki, A.; Nymand, M. Analytical Design of Passive LCL Filter for Three-Phase Two-Level Power Factor Correction Rectifiers. *IEEE Trans. Power Electron.* **2018**, *33*, 3012–3022. [CrossRef]
11. Kumar Peddapelli, S. 2. Space vector pulse width modulation technique. In *Pulse Width Modulation: Analysis and Performance in Multilevel Inverters*; De Gruyter Oldenbourg: Berlin, Germany; Boston, MA, USA, 2017; pp. 13–22. [CrossRef]
12. Lewicki, A.; Odeh, C.I.; Kondratenko, D.; Morawiec, M. Hybridized Space-Vector Pulsewidth Modulation for Multiphase Two-Level Voltage Source Inverter. *IEEE Trans. Power Electron.* **2022**, *37*, 7663–7674. [CrossRef]
13. Doan, N.S.; Tsvetkov, A.N.; Nguyen, T.H. Study and implementation of space vector pulse width modulation inverter on an Arduino. *E3S Web Conf.* **2021**, *288*, 4–9. [CrossRef]
14. Waqas, M.; Ahmed, T.; Elavarasan, R.M.; Waqar, A.; Leong, K.; Pugazhendhi, R.; Das, N.; Jeelani, M.W. DQ Transformation Based Control of Single-Phase Grid-Tied Inverter. In Proceedings of the 2021 31st Australasian Universities Power Engineering Conference (AUPEC), Virtual, 26–30 September 2021.
15. Venkatramanan, D.; Singh, M.K.; Ajala, O.; Dominguez-Garcia, A.; Dhople, S. Integrated System Models for Networks with Generators & Inverters. *arXiv* **2022**, arXiv:2203.08253.
16. Altahir, A.A.R. Park and Clark Transformations: A Short Review. April 2020, pp. 2–5. Available online: https://www.researchgate.net/profile/Ali-Altahir/publication/340681625_Park_and_Clark_Transformations_Park_and_Clark_Transformations_A_Short_Review/links/5e98bb8a4585150839e3768b/Park-and-Clark-Transformations-Park-and-Clark-Transformations-A-Short-Review.pdf (accessed on 27 May 2024).

17. Casado-Machado, F.; Martinez-Ramos, J.L.; Barragan-Villarejo, M.; Maza-Ortega, J.M.; Rosendo-Macias, J.A. Reduced Reference Frame Transform: Deconstructing Three-Phase Four-Wire Systems. *IEEE Access* **2020**, *8*, 143021–143032. [CrossRef]
18. O'Rourke, C.J.; Qasim, M.M.; Overlin, M.R.; Kirtley, J.L. A Geometric Interpretation of Reference Frames and Transformations: Dq0, Clarke, and Park. *IEEE Trans. Energy Convers.* **2019**, *34*, 2070–2083. [CrossRef]
19. Hollweg, G.V.; Khan, S.A.; Chaturvedi, S.; Fan, Y.; Wang, M.; Su, W. Grid-Connected Converters: A Brief Survey of Topologies, Output Filters, Current Control, and Weak Grids Operation. *Energies* **2023**, *16*, 3611. [CrossRef]
20. Ali, M.; Yaqoob, M.; Cao, L.; Loo, K.H. Enhancement of DC-bus voltage regulation in cascaded converter system by a new sensorless load current feedforward control scheme. *IET Power Electron.* **2021**, *14*, 1457–1467. [CrossRef]
21. Wu, H.; Wang, X. Control of Grid-Forming VSCs: A Perspective of Adaptive Fast/Slow Internal Voltage Source. *IEEE Trans. Power Electron.* **2023**, *38*, 10151–10169. [CrossRef]
22. Sadeque, F.; Fateh, F. On Control Schemes for Grid-Forming Inverters. In Proceedings of the 2022 IEEE Kansas Power and Energy Conference (KPEC), Manhattan, KS, USA, 25–26 April 2022.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

AppliedMath

MDPI

Article

Solving Complex Optimisation Problems by Machine Learning †

Steven Prestwich

School of Computer Science and Information Technology, University College Cork, T12 XF62 Cork, Ireland; s.prestwich@cs.ucc.ie; Tel.: +353-36-21-420-5911

† This paper is an extended version of our paper published in the 9th International Conference on Machine Learning, Optimization, and Data Science, LOD 2023 (Grasmere, Lake District, UK, 22–26 September 2023).

Abstract: Most optimisation research focuses on relatively simple cases: one decision maker, one objective, and possibly a set of constraints. However, real-world optimisation problems often come with complications: they might be multi-objective, multi-agent, multi-stage or multi-level, and they might have uncertainty, partial knowledge or nonlinear objectives. Each has led to research areas with dedicated solution methods. However, when new hybrid problems are encountered, there is typically no solver available. We define a broad class of discrete optimisation problem called an influence program, and describe a lightweight algorithm based on multi-agent multi-objective reinforcement learning with sampling. We show that it can be used to solve problems from a wide range of literatures: constraint programming, Bayesian networks, stochastic programming, influence diagrams (standard, limited memory and multi-objective), and game theory (multi-level programming, Bayesian games and level-k reasoning). We expect it to be useful for the rapid prototyping of solution methods for new hybrid problems.

Keywords: multi-objective; multi-agent; reinforcement learning; optimisation

1. Introduction

Optimisation problems are ubiquitous in artificial intelligence, operations research, and a wide range of application areas. In their simplest form, they require us to make a set of decisions in order to optimise an objective function, possibly while satisfying some constraints. A decision is usually represented by the assignment of a domain value to a decision variable, where the domain might be a set of real numbers, integers, categories or data structures. Different domains, constraints and objectives are supported by different methods. Mixed integer linear programming (MILP) and more generally mathematical programming, constraint programming (CP), dynamic programming and metaheuristics (such as local search and genetic algorithms) can be applied to such problems.

However, many optimisation problems are more complex than this. In particular, they might involve uncertainty, which can be modelled by *chance variables* that are not controlled by a decision maker. Instead, they take values according to some probability distribution. Stochastic programming is an extension of MILP that models such problems. In multi-stage stochastic programs, decision and chance variables might be interleaved so that we must make decisions without knowing the values of some chance variables. Having made decisions, we can then observe some chance variables, but must then make next-stage decisions, and so on. A solution to a multi-stage problem is not a simple set of decisions, but a policy taking the form of a tree to allow reactions to random events. Probability distributions are usually assumed to be unaffected by decisions (*exogenous uncertainty*), but some applications have decision-dependent distributions (*endogenous uncertainty*). Stochastic dynamic programming is often used in the latter case.

Another possible complication is that a decision maker, or *agent*, might have multiple objectives: for example, to maximise profit while minimising environmental damage. Some form of compromise must be reached for such problems. Yet another complication is that a

check for updates

Citation: Prestwich, S. Solving Complex Optimisation Problems by Machine Learning. *AppliedMath* **2024**, *4*, 908–926. https://doi.org/10.3390/appliedmath4030049

Academic Editors: Libor Pekař and Jinyun Yuan

Received: 8 May 2024
Revised: 24 July 2024
Accepted: 29 July 2024
Published: 31 July 2024

Copyright: © 2024 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

problem might involve multiple agents. In multi-level programming, an agent must make a decision in the knowledge of how another agent will react, who might in turn have to take other agents into account. Knowledge might be partial: an agent might not know exactly how an adversary will act or has acted, and must make a decision based on assumptions.

Influence diagrams (IDs) are a particularly expressive formalism that can model multi-stage decision problems with endogenous uncertainty. Chance variables may be independent or related by conditional probability tables, forming a Bayesian network. *Observations* or the observed values of chance variables are also allowed, and Bayesian inference can be used to infer the distributions of unobserved chance variables based on these observations. Several exact and approximate methods are available for solving IDs.

A machine learning approach to complex optimisation problems is *reinforcement learning* (RL) [1] which can be used to solve multi-stage decision problems in stochastic environments. Multi-agent RL has famously been used to learn to play games such as Go and Chess to superhuman levels via self-play [2]. Multi-objective RL algorithms have also been devised, and these extensions have been combined in the field of multi-objective multi-agent RL. Similar developments have occurred in the evolutionary computation literature.

Thus, we have a wide range of available technologies, each able to solve certain types of problem efficiently. However, faced with a new optimisation problem with hybrid features, there might be no available solver. Some hybrids have been investigated, for example, multi-objective reinforcement learning with constraints [3], and bi-level multi-objective stochastic integer linear programming [4]. IDs have been extended to handle nonlinear utilities [5], multiple agents [6–10], multiple objectives [11,12], hard constraints [13], and partially observed variables [14]—though not all in one system. But no method exists for complex hybrids in general. To take a hypothetical example, there is no obvious way of modelling and solving a discrete bi-level optimisation problem whose leader is a non-linear chance-constrained stochastic program, and whose followers are a bi-objective constraint satisfaction problem and a weighted Max-SAT problem. A researcher faced with such a problem must invent a new approach or make simplifying assumptions to fit it into an existing framework. A general-purpose method able to handle all these complications would be very useful, at least for a preliminary investigation, though a more efficient specialised method might eventually be needed.

Another drawback of the fragmentation of technologies is that it is difficult to compare different approaches. Suppose we are faced with a new application with two decision makers but we are unsure how best to model it. Should we treat it as a bi-level program (see Section 3.8), a Bayesian game (see Section 3.9), using level-k thinking (see Section 3.10), model one decision maker as a random variable and implement a stochastic program (see Section 3.6) or take an expected value of one agent and reduce it to a constraint program (see Section 3.1)? Some simplifications might make little difference to the objective while others might make feasibility impossible. Each approach requires a different technology that takes considerable time to master, and possibly financial cost. A method that is general enough to solve all these models would be of great benefit for rapid prototyping, to explore the consequences of model design choices.

In this paper, we describe a very general class of discrete optimisation problem we call an *Influence Program* (IP) that generalises a wide range of problems from operations research, artificial intelligence and game theory (the name is inspired by the flexibility of IDs). We also present a simple solver for IPs based on a combination of multi-agent multi-objective reinforcement learning and Markov Chain Monte Carlo sampling. RL has been proposed as a unifying approach to sequential decision-making under uncertainty [15] and has solved complex adversarial games, making it a natural candidate for a general-purpose solver.

This paper extends our previous work. An early version using logic programming was described in [16], and a more recent C-based version appears in [17] where our problem class was called a "mixed influence diagram". This paper extends [17] by formalising the problem, providing more related work, replacing simple rejection sampling by Gibbs sampling, and solving a wider range of problems. Section 2 formalises the problem

and describes the algorithm. Section 3 applies it to a variety of problems from different optimisation literatures to demonstrate its flexibility and ease of use. Finally, Section 4 summarises the results and discusses future work.

2. The Influence Programming Framework

We now introduce our IP framework and an algorithm, after discussing related work.

2.1. Related Work

Considerable work has been performed on multi-objective RL (MORL). The MORL literature is too large to survey here, but a recent survey on MORL algorithms with a discussion on these issues is provided in [18]. Some methods use scalarisation to convert MORL into RL, making it a single-policy method like several others. Scalarisation has been criticised because it leads to only a single policy, and some MORL algorithms approximate Pareto fronts and learn policies for a range of weightings. Scalarisation has also been criticised on the grounds that it can be hard to choose appropriate weights, and that the choice puts decision power in the hands of engineers running the algorithm. Most such methods use linear scalarisation. Despite the shortcomings of scalarisation, we shall show that it can give good results on a range of problems. It is also easy to implement and has little runtime overhead, which is important for our lightweight approach. Results using weighted metrics are also less sensitive to the choice of weights [19].

Much work has also been performed on multi-agent RL (MARL), and a recent MARL survey can be found in [20]. Less work has been performed on the intersection of MARL and MORL: multi-objective multi-agent RL (MOMARL). A survey on MOMARL and related algorithms is provided in [21]. Because of the complexity of MOMARL problems, there is not even a single agreed definition of what constitutes a solution. MO-MIX [22] is a MOMARL algorithm using linear scalarisation, and deep RL in which an artificial neural network is used for state aggregation.

2.2. The Problem Class

We define a *(discrete) influence program* (IP) as a tuple $\langle \mathcal{V}, \mathcal{D}, \mathcal{A}, \mathcal{U}, \mathcal{L}, \mathcal{O}, \mathcal{P} \rangle$ where:

- $\mathcal{V} = \langle v_1, \ldots, v_n \rangle$ is an ordered list of variables v_i;
- $\mathcal{D} = \langle D_1, \ldots, D_n \rangle$ is a list of their corresponding finite domains D_i of possible values, typically ranges of integers or sets of symbolic names;
- $\mathcal{A} = \langle A_1, \ldots, A_n \rangle$ is a list of their corresponding *agents* (decision makers) A_i, which have symbolic names (in **bold**);
- $\mathcal{U}$ is a function mapping a total variable assignment to a utility vector for each agent;
- $\mathcal{L}$ is a set of directed links $v_i \to v_j$ $(i < j)$ between variables;
- $\mathcal{O}$ is a set of *observations*, where an observation is an assignment $v_i = d_j$ of a value $d_j \in D_i$ to a chance variable x_i;
- $\mathcal{P}$ is a function assigning a probability to each chance variable v_j assignment given an assignment for each variable v_i such that $v_i \to v_j \in \mathcal{L}$: it is typically represented by a (conditional) probability table.

Each variable is associated with exactly one agent, and we allow the possibility of chance variables whose agent has the name **chance**. The links $\mathcal{L}$ define which variables are visible to later decisions and chance variable distributions. A utility is a function mapping a total variable assignment (one value per variable) to a real value: we allow utilities to be *programmable*, requiring the user to write a small function to compute utilities from a total variable assignment. Each agent has at least one utility, apart from **chance**. The aim of an IP is to find a policy for each agent that Pareto-optimises its utilities in the context of observations and inter-variable visibility.

2.3. An Algorithm

We now describe the INFPROG algorithm shown in Algorithm 1. It is a lightweight research prototype using known techniques, used only to demonstrate the flexibility of our approach, and many other combinations of methods are potentially possible.

INFPROG is based on a simple RL algorithm: infinite-step tabular SARSA with ϵ-greedy action selection and learning rate α [1]. *Infinite-step* indicates that the reward is backed up equally to all values in the episode, which is more robust than Q-learning in the presence of unobserved variables [1]. *Tabular* indicates that state–action pairs have values in a table. The discount factor γ is set to 1 as our RL problem is episodic. If an agent has multiple utilities, these are scalarised so that there is one utility per agent (see Section 2.6). The scalarised objectives are used as rewards, computed at the end of each episode when all variables have been assigned updated values, and backed up to earlier states for those agents ($\mathbf{a}[v]$ denotes the agent $\mathbf{a}$ corresponding to variable v): the user must provide code for this step. However, the user is probably not interested in the value of the scalarised objective, so smoothed versions of the original objectives are printed out: these will be the values we report.

Algorithm 1 The INFPROG algorithm for solving IPs

Require: integers H, E and a utility hyperparameter Y
  initialise $\epsilon = 1$, $\alpha = 1$, $q_{H,v,i} = 0$
  for episode $e = 1 \dots E$ **do**
    for $v = v_1 \dots v_n$ **do**
      if v is chance **then**
        sample x_v from its distribution
      else if $U[0,1] < \epsilon$ **then**
        randomly sample $x_v \in D_v$
      else
        $x_v = \text{argmin}_{i \in D_v}(q_{H,v,i})$
      end if
    end for
    for each agent $\mathbf{a}$ **do**
      compute scalarised objective $f_{\mathbf{a}}(\vec{v})$ using Y
    end for
    for $v = v_1 \dots v_n$ **do**
      $q_{H,v,x_v} = \alpha f_{\mathbf{a}[v]}(\vec{v}) + (1-\alpha) q_{H,v,x_v}$
    end for
    update ϵ and α
  end for

The utility is computed at the end of an episode: in RL terms, this is a *sparse reward* which can make learning harder. In future work, we could allow for multiple value nodes as in IDs, or compensate for sparse rewards by adding RL techniques such as *hindsight experience replay* [23]. That technique was designed for off-policy RL algorithms but SARSA is on-policy, so some algorithm modification would be needed: for example, replacing SARSA by Q-learning or a more recent deep RL algorithm. There are many possibilities and this area is ripe for exploration.

The ϵ and α parameters start at 1 and decay to 0. Many decay schemes have been proposed in RL, and we arbitrarily choose $\alpha = \epsilon = ((E-e)/E)^3$. In our use of RL, a *state* is an assignment to variables $v_1 \dots v_i$ for some $i = 1 \dots n$, an *action* is the assignment of a value to a decision variable v_{i+1}, an *episode* assigns all the variables, and the *reward* is the utility (objective function value) computed at the end of an episode. The q_{H,v,x_v} are state–action values used to define the policy, which should optimise the expected reward.

Note that it is likely that for some problems it would be better to use SARSA with an eligibility trace and $0 < \lambda < 1$ (λ is a hyperparameter used with eligibility traces [1]) and we shall investigate this in future work. However, for our research, prototype we

effectively set $\lambda = 1$ by choosing infinite-step SARSA, thus simplifying the algorithm by removing a hyperparameter that requires tuning.

2.4. State Aggregation

Each decision might depend on all previous decisions and random events (but see Section 3.5), so a policy might involve a huge number of distinct states. To combat this problem, RL algorithms group together states via *state aggregation* methods. We choose a simple form called *random tile coding* [1], specifically *Zobrist hashing* [24] with H hash table entries for some large integer H. This works as follows. To each (decision or chance) variable–value pair $\langle v, x \rangle$, we assign a random integer r_{vx} which remains fixed. At any point during an episode, we have some set S of assignments $\langle v, x \rangle$, and we take the exclusive -or of the r_{vx} values (that is, their bit patterns) associated with assignments $X_S = \bigoplus_{\langle v,x \rangle \in S} r_{vx}$. Finally, we use X_S to index an array V with H entries: the value of q_{H,v,x_v} is stored in $V[X_S \bmod H]$ (in all our experiments, we fix $H = 2 \times 10^8$.)

The INFPROG algorithm takes two numerical hyperparameters: an integer H used for state aggregation and an integer E which is the number of iterations used by the solver. If H is sufficiently large then hash collisions are unlikely, and we will have a unique array element for each state encountered. It might be expected that Zobrist hashing will perform poorly when the number of states approaches or exceeds the size of the hash table, because hash collisions will confuse the values of different state–action pairs. Surprisingly, it can perform well even when hash collisions are frequent, and has been used in chess programming [25].

2.5. Sampling

INFPROG samples chance variables using a Markov chain Monte Carlo algorithm (MCMC) as in probabilistic programming. Thus, each SARSA episode is also a sweep through the chance variables. Our earlier work [16,17] used rejection sampling: at the end of a SARSA episode, if the chance variables did not match the observations, then the episode was rejected, in the sense that rewards were not backed up to earlier states. This worked on some problems but is impractical when probabilities are very low, and the use of an MCMC algorithm was proposed as future work. INFPROG uses Gibbs sampling and can handle such cases.

In the experiments, we found that some conditional probability table values that were set to 0 or 1 required adjustment: this is a known failure mode of Gibbs sampling. In fact, for optimisation problems without observations (which is true of most problems), we replace Gibbs with simple inverse transform sampling.

2.6. Multiple Objectives

Many real-world applications have multiple objectives and we must find a trade-off. Objectives can be combined in more than one way in RL, and a simple and popular approach is *linear scalarisation*: take a linear combination of the objectives, giving a single objective that can be used as an RL reward. This has the drawback that it cannot generate any solutions in a non-convex region of the Pareto front. It can also be hard to choose appropriate weights, especially when the objectives use different units.

Another approach is to rank the objectives by importance, then search for the lexicographically best result. This has the same problem as linear scalarisation (though recent work addresses this [26]) which can be fixed by thresholding all but the least important objective, a method called *thresholded lexicographic ordering* [27]. However, it has the drawback of using a specific RL algorithm. Moreover, in some of our applications, the most important objective is to minimise constraint violations, and this should not be thresholded.

The method we choose is to reduce the multiple objectives to a single objective via *weighted metric scalarisation* [19] in which the distance is minimised between the vector of values f_o for objectives o and a *utopian point* u^* in the multi-objective space:

$$\min_x \left(\sum_o w_o |f_o(x) - u_o^*|^p \right)^{1/p}$$

for some $p \geq 1$ and weights w_o chosen by the user. The need to choose weights is a disadvantage in terms of user-friendliness, but an advantage is that we can tune them to find different points on the Pareto front. The special case of *Chebyshev scalarisation* ($p = \infty$) is theoretically guaranteed to make the entire Pareto front reachable, but the L_p-norms ($p = 1, 2$) may perform better in practice [28]. In experiments, we found better results with L_1 than with L_2 or L_∞ despite Chebyshev's theoretical advantages, so we shall use L_1. However, in future work, it might be better to use Chebyshev and find ways of improving its results.

The utopian point is often adjusted during search to be just beyond the best point found so far, but INFPROG uses a fixed u^* provided by the user. Depending on whether each objective is to be maximised or minimised, we choose a value that is optimal or high/low enough to be unattainable.

Thus, to apply INFPROG to an optimisation problem, we must also provide a hyperparameter Y to guide scalarisation: a list of sublists of pairs. Each pair is a utopian value expressing a desired utility plus a weight, and each sublist corresponds to an agent and contains a pair for each of the agent's utilities. We shall always choose weights that sum to 1 for each agent, though this is not strictly necessary. The use of Y will be illustrated below via examples.

2.7. An Intuitive Explanation

To aid the reader, we now provide an intuitive explanation of how the algorithm works. The outer loop of the algorithm performs E *episodes* (a term taken from RL) where E is an integer hyperparameter chosen by the user. Each episode sweeps through the variables in the order specified by the list $\langle v_1, \ldots, v_n \rangle$, assigning each a value.

A decision variable is assigned a value as in the SARSA algorithm: during early episodes, assignment is largely random, but in later episodes, values are assigned more greedily in order to maximise the estimated expected *reward* (another RL term, corresponding to *utility* in influence diagrams and to *objective function value* in general optimisation). The degree of greediness is controlled by the ϵ parameter, which decays from 1 (completely random) in the first episode to 0 (completely greedy) in the last episode. In RL algorithms this strategy theoretically leads to an optimal policy.

A chance variable is assigned a value via sampling, thus an episode interleaves SARSA with sampling. In RL terms, the chance variable assignment is simply part of the environment in which the agent operates. In the special case where all variables are chance variables, the IP is a Bayesian network and an episode reduces to an MCMC sweep through the variables.

At the end of an episode, a reward is computed for each agent, and backed up to all earlier states as in SARSA, with the slight complication that in our multi-agent version each reward must be backed up to the appropriate agent.

The choice of infinite-step tabular SARSA was largely arbitrary, and has both advantages and disadvantages. An on-policy algorithm like SARSA is considered to be more consistent and stable than an off-policy algorithm like Q-learning, but also slower and less efficient. The choice of a tabular algorithm was made for its simplicity of implementation. The infinite-step choice corresponds to setting $\lambda = 1$ in an RL eligibility trace, making it a Monte Carlo method rather than a temporal difference method. The former are considered more robust in the presence of unobserved variables, as they are not based on the Markov property.

The restriction to rewards that occur only at the end of an episode was designed to simplify the implementation and user interface. In some cases, it would be preferable to allow intermediate rewards, if these occur naturally in the problem. However, this would require the user to write several pieces of code and to specify when each is to be executed. For our research prototype we took the simpler path.

Regarding the necessity for the user to provide programmable utility functions: although this might seem less user-friendly than some form of specification language, it is extremely flexible. We note that a similar approach is taken in the field of probabilistic programming [29,30].

3. Applications

In this section, we take small problems from a range of studies in the literature, model them as IPs, and solve them using InfProg. We shall not compare our method with others in terms of efficiency, as this is not the goal of the paper. For the same reason, we report few runtimes, though they are quite short—typically a few seconds and at most a few minutes. In fact, we do not expect it to be competitive on any particular class of problem, though as an RL-based method, it should perform reasonably well on applications from the RL literature. Our aim here is only to demonstrate that a single solver can solve a wide variety of optimisation problems, thus filling a gap in optimisation technology. We intend to implement faster, more scalable, and more user-friendly IP solvers in future work, and we hope that other researchers will also find better algorithms.

3.1. Constraint Programming

First we consider a "simple" optimisation problem: one agent and one objective. We take a well-known *constraint satisfaction problem* (CSP) known as eight queens: place eight chess queens on a standard 8×8 chessboard in such a way that no two queens attack each other (by being on the same row, column, or diagonal). This is a popular problem in constraint programming [31], and the smaller four queens problem is illustrated in Figure 1. The first example is a solution because no two queens are in the same row, column, or diagonal. The second example is a non-solution because it violates three constraints: the queens in the first two rows are in the same column, those in the last two rows lie on the same diagonal, and those in the first and third rows lie on another diagonal.

We treat eight queens as a *Max-CSP* problem in which the objective is to maximise the number of satisfied constraints (in this case, minimising the number of constraint violations to 0). We shall consider two possible IP models.

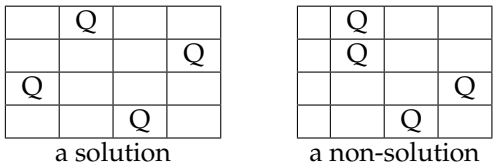


Figure 1. The 4-queens problem.

3.1.1. A Sparse Model

The problem can be modelled as follows: $\mathcal{V} = \langle v_1, \ldots, v_8 \rangle$, each $D_i = \{1, \ldots, 8\}$, $\mathcal{A} = \langle \mathbf{a}, \ldots, \mathbf{a} \rangle$ (all variables belong to the same agent **a**), $\mathcal{U} = \langle\langle viol \rangle\rangle$ where *viol* counts constraint violations:

$$\sum_{i=2}^{8} \sum_{j=1}^{i-1} [\![v_i = v_j \vee |i-j| = |v_i - v_j|]\!]$$

$[\![\cdot]\!]$ is the *Iverson bracket* that takes the value 1 if its argument is true and 0 if it is false. $\mathcal{P}$ is the empty function (there are no probability tables because there are no chance variables) and $\mathcal{L} = \mathcal{O} = \emptyset$. The utility hyperparameter is $\Upsilon = \langle\langle\langle 0, 1 \rangle\rangle\rangle$: one agent with one utility whose utopian point is 0 and associated weight is 1.

Though we shall not report runtimes in general, it might be of interest to the reader to see one example. Our solver is implemented in C and executed on a Intel(R) Core(TM) i5-6500 CPU @ 3.20 GHz with 8 GB of system RAM (Intel, Santa Clara, CA, USA). With 10^7 episodes, the solver takes 10.3 s per run, and in 100 runs, it found a correct solution 70 times. The other 30 times, the solution contained one constraint violation, illustrating that INFPROG can become trapped in local optima. It is also quite successful using only 10^6 episodes, finding 61 correct solutions.

This result is not competitive with methods such as constraint programming, integer programming or local search, which can all solve the problem more quickly and more reliably. But, as stated earlier, our aim is flexibility across a wide range of problems, not efficiency on any particular problem class.

3.1.2. A Dense Model

We referred to the previous model as *sparse* because it had few links (0). As an alternative *dense* model, we can add links to the IP from each decision variable to each earlier one: $\mathcal{L} = \{v_i \rightarrow v_j | i < j\}$. These are not logically necessary but they provide more information to the decision variables. With the dense IP, we found 100% success on eight queens using only 10^5 episodes, and even with only 30,000 episodes, it achieved 69% success. This suggests that making decisions visible to all subsequent decisions will improve performance. However, adding these links creates a number of states that are exponential in the number of variables, causing many hash collisions in INFPROG, which in turn might degrade performance.

This phenomenon is well known in RL. As an example, consider the elevator dispatching problem in which we must control a set of elevators in order to minimise the expected waiting times. Even if we restrict the problem to discrete time, it has a vast number of states, because of the many possible combinations of elevator positions, the direction of movement, and which buttons have been pressed. In principle, it would be best to treat every state separately, and the superior performance of our dense model supports this view, but combinatorial explosion makes it impractical for large problems.

One way of avoiding the explosion is to treat each elevator independently in a *distributed RL* approach—see, for example [32]. There is a decision-making agent for each elevator and they do not communicate, but they have a common objective and they indirectly learn to cooperate. A recent survey of distributed RL with applications is given in [33]. Our sparse IP works in the same way, though we only used one agent: we could instead use a separate agent for each variable, but it makes no difference because they have the same objective.

Controlling the links between variables makes it easy for INFPROG to emulate RL methods that are fully-, partially-, and non-distributed. For subsequent problems, we shall use a mixture of sparse and dense models.

3.2. Bayesian Networks

From a problem with all decision variables, we move to a problem with all chance variables: Pearl's alarm example [34]. This is an example of *probabilistic inference*, as performed in the field of probabilistic programming: inferring a conditional probability from a probabilistic program.

In this small problem, a house has an alarm that can be set off by an earthquake or a burglary, each with a prior probability. We also have conditional probabilities for the alarm under different circumstances: when there is/is not a burglary and there is/is not an earthquake. Given that the alarm has been activated, what is the probability that a burglary has occurred? The answer affects our actions.

We can model this as an IP. The variables are $\mathcal{V} = \{burglary, earthquake, alarm\}$, the domains $\mathcal{D}$ are all $\{yes, no\}$, the agents $\mathcal{A}$ are all **chance**, the links are $\mathcal{L} = \{burglary \rightarrow alarm, earthquake \rightarrow alarm\}$ as in the Bayesian network, the probabilities $\mathcal{P}$ are:

$$\begin{aligned} \Pr[burglary = yes] &= 0.7 \\ \Pr[earthquake = yes] &= 0.2 \\ \Pr[alarm = yes | burglary = yes, earthquake = yes] &= 0.9 \\ \Pr[alarm = yes | burglary = yes, earthquake = no] &= 0.8 \\ \Pr[alarm = yes | burglary = no, earthquake = yes] &= 0.1 \\ \Pr[alarm = yes | burglary = no, earthquake = no] &= 0.2 \end{aligned}$$

there is one observation $\mathcal{O} = \{alarm = yes\}$ and the utility vector is

$$\mathcal{U} = \langle\langle [\![earthquake = yes]\!] \rangle\rangle$$

Hence, the task is to compute the conditional probability $\Pr[earthquake = yes | alarm = yes]$. We can treat this as a utility, though the only agent is **chance** so we are not trying to optimise it by taking decisions. There are no decision-making agents, so the utility hyperparameter is $\mathrm{Y} = \langle\langle\langle\rangle\rangle\rangle$. InfProg applies Gibbs sampling and finds an expected "utility" of approximately 0.23, which is the correct probability.

We shall not use observations in subsequent examples as they are not a feature of most optimisation problems. However, they can be used to compute the *value of information* to determine which variables are most likely to reduce the uncertainty in a variable of interest.

3.3. Influence Diagrams

We now move to problems containing both decision and chance variables. IDs [35] are popular graphical models in decision analysis, and they can model important relationships between uncertainties, decisions, and values. They were initially conceived as tools for formulating problems but they have also emerged as efficient computational tools. They have many applications including medical diagnosis [36], cybersecurity [37], and risk management [38].

An ID is a directed acyclic graph with three types of node: *decision nodes* correspond to decision variables and are drawn as rectangles; *chance* (or *uncertain*) *nodes* correspond to chance variables and are drawn as ovals; *value nodes* correspond to preferences or objectives and are drawn as rounded rectangles, or polygons such as diamonds.

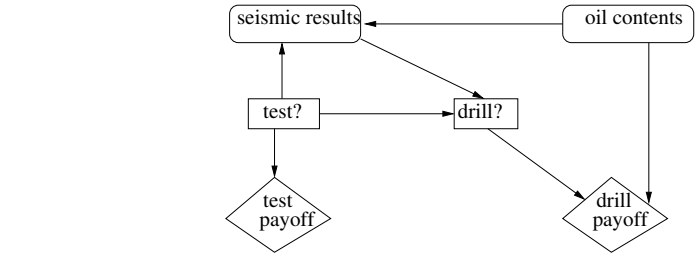
Each chance variable is associated with a conditional probability table that specifies its distribution for every combination of values for its parent variables in the graph. Each decision variable also has a set of parent variables in the graph, and its value depends only on their values. The decision variables are usually considered to be temporally ordered, and chance variables are observed at different points in the ordering. A standard ID assumption is *non-forgetting*, which means that the parents of any decision variable are all its ancestors in the graph: thus, a decision may depend on everything that has occurred before. All variables are discrete.

Each value node is associated with a table showing the *utility* (a real number) of each combination of parent variable values. A *decision policy* is a rule for each decision variable indicating how to choose its value from those of its parents. Any policy has a total *expected utility* (it is an expectation because of the chance variables) and *solving* an ID means computing its optimal policy, which has maximum expected utility.

Several methods exist for solving IDs. Some are exact and based on variable elimination [39–43] while others are approximate [44–48]. However, apart from our work ([16,17] and this paper), we know of no other applications of RL to solving IDs, which seems surprising as both IDs and RL can be used to model and solve sequential decision problems. The main connection usually made between the two is that IDs can model problems that can be tackled by RL, for example, causal IDs have been used to model artificial general intelligence safety frameworks which often use RL [49]. Our RL-based approach lies somewhere

between exact and approximate methods: given sufficient memory and training time, it has the potential to find an optimal solution but is not guaranteed to do so. We do not expect it to be as efficient as specialised algorithms, but it can tackle IDs with extensions such as limited memory (see Section 3.5), multiple agents, and multiple objectives. Moreover, the recent successes of deep RL make its application to large IDs a promising research direction. However, in this paper, we restrict ourselves to showing that RL can indeed solve IDs.

We use the Oil Wildcatter ID shown in Figure 2, a well-known problem published in [50]. An oil wildcatter must decide either to drill or not to drill for oil at a specific site. Before drilling, they may perform a seismic test that will help determine the geological structure of the site. The test result can be *closed* (indicating significant oil), *open* (indicating some oil), or *diffuse* (probably no oil). The special value *notest* means that test results are unavailable if the seismic test is not performed. The *test* decision does not depend on any other variable, but the *drill* decision depends on whether a test was made, and if so, on its result. The *oil* variable is unobservable so no decision depends on it (this is distinct from *forgetting* its value which is addressed in Section 3.5).



content	Pr[*O*]		
	dry	wet	soak
	0.5	0.3	0.2

test payoff	$U_1(T)$
test	
yes	−10
no	0

Seismic results		Pr[*S*|*O*, *T*]			
content	test	closed	open	diffuse	notest
dry	yes	0.1	0.3	0.6	0
dry	no	0	0	0	1
wet	yes	0.3	0.4	0.3	0
wet	no	0	0	0	1
soak	yes	0.5	0.4	0.1	0
soak	no	0	0	0	1

Drill payoff		$U_2(O, D)$
content	drill	
dry	yes	−70
dry	no	0
wet	yes	50
wet	no	0
soak	yes	200
soak	no	0

Figure 2. Oil wildcatter influence diagram.

An IP model is as follows. The variables are $\mathcal{V} = \langle oil, test, seismic, drill \rangle$, the domains are

$$\mathcal{D} = \langle \{dry, wet, soak\}, \{yes, no\}, \{closed, open, diffuse, notest\}, \{yes, no\} \rangle$$

the agents are $\mathcal{A} = \langle \textbf{chance}, \textbf{company}, \textbf{chance}, \textbf{company} \rangle$, the links are $\mathcal{L} = \{drill \rightarrow seismic, drill \rightarrow test\}$ (ID links to utilities are not part of an IP), there are no observations ($\mathcal{O} = \emptyset$), the utility vector is $\mathcal{U} = \langle \langle t + d \rangle \rangle$ where t is the test payoff and d is the drill payoff, and the probabilities are as shown in Figure 2. The utility hyperparameter is $\Upsilon = \langle \langle \langle 1000, 1 \rangle \rangle \rangle$: one agent with one objective to be maximised.

The known optimal policy given the above utilities and probabilities is as follows: apply the seismic test, and drill if the test result is open or closed. InfProg finds this solution and reports a close approximation to the correct expected utility: 22.5.

3.4. Multi-Objective Influence Diagrams

As a first example of multi-objective optimisation, we use a bi-objective oil wildcatter ID from [12]. In addition to maximising payoff, the aim is to minimise environmental damage. The IP is as in Section 3.3, except for additional payoffs: the utility of testing is now $(-10, 10)$ while the utility of drilling is $(-70, 18)$, $(50, 12)$, and $(200, 8)$ for dry, wet,

and soak, respectively: in each case, the first value is the original utility while the second value refers to the new utility. Instead of one optimal solution, there are four Pareto-optimal solutions, each with two utility values: (1) test, then drill if the result is closed or open, with utility $(22.5, 17.56)$; (2) do not test but drill, with the utility $(20, 14.2)$; (3) test, then drill if the result is closed, with utility $(11, 12.78)$; (4) do not test or drill, with utility $(0, 0)$.

The utilities are now $\mathcal{U} = \langle\langle\langle p, d\rangle\rangle\rangle$, where p is the payoff as before, and d is the environmental damage. With a utility hyperparameter $\mathrm{Y} = \langle\langle\langle 1000, 0.55\rangle, \langle -1000, 0.45\rangle\rangle\rangle$, in multiple runs, INFPROG found policies (1), (2), and (4), but not (3).

3.5. Limited Memory Influence Diagrams

Standard IDs are designed to handle situations involving a single, non-forgetful agent. Limited memory influence diagrams (LIMIDs) [14] are generalisations of IDs that allow decision making with limited information and simultaneous decisions, and can have much smaller policies. They relax the regularity (total variable ordering) and non-forgetting assumptions of IDs. LIMIDs are considered harder to solve optimally than IDs. We handle the limited memory feature in a simple way: during Zobrist hashing, the set S contains only assignments that are visible to the decision variable (as specified by the IP links $\mathcal{L}$).

For example, we use a pig breeding problem from [14]. A pig breeder grows pigs for four months and then sells them. During this period, the pig may or may not develop a disease. If it has the disease when it must be sold, then it must be sold for slaughter and its expected market price is 300. If it is disease-free, then its expected market price is 1000. Once a month, a veterinary surgeon tests the pig for the disease. If it is ill, then the test indicates this with a probability of 0.80, and if it is healthy, then the test indicates this with a probability of 0.90. At each monthly visit, the surgeon may or may not treat the pig, and the treatment costs 100. A pig has the disease in month 1 with a probability of 0.10. A healthy pig develops the disease in the next month with a probability of 0.20 without treatment and 0.10 with treatment. An unhealthy pig remains unhealthy in the next month with a probability of 0.90 without treatment, and 0.50 with treatment. The ID is shown in Figure 3.

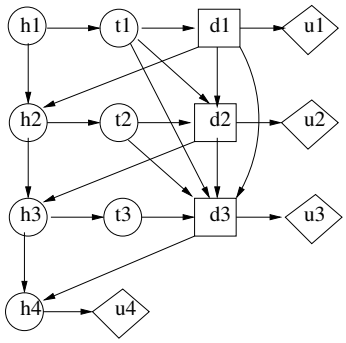


Figure 3. Pig breeding influence diagram.

An IP to model this problem is as follows. The variables are $\mathcal{V} = \langle h_i, t_i, d_i, \ldots, h_4\rangle$ $(i = 1, 2, 3)$, the domains $\mathcal{D}$ are $\{healthy, unhealthy\}$ for the h_i, $\{pos, neg\}$ for the t_i and $\{treat, leave\}$ for the d_i, there is one agent in $\mathcal{A}$ we call **breeder**, the links are $\mathcal{L} = \{h_i \to h_{i+1}, h_i \to t_j (i < j), t_i \to d_j (i \leq j), d_i \to h_{i+1}\}$, $\mathcal{P}$ contains the probabilities shown, $\mathcal{U} = \langle\langle u_1 + u_2 + u_3 + u_4\rangle\rangle$, and $\mathcal{O} = \emptyset$.

Using a utility hyperparameter $\mathrm{Y} = \langle\langle\langle 1000, 1\rangle\rangle\rangle$, INFPROG almost always finds the optimal policy with an expected utility of approximately 727: ignore the first test and do not treat, then follow the results of the other two tests (treat if positive). As noted in [16], a different policy is given in [14]: treat in month 3 if tests 1, and 2, or 3, are positive. We find that their policy has an expected utility of 725.884 while ours is optimal. They report the same expected utility as we do, so we believe this was simply a typographical error.

3.6. Multi-Stage Stochastic Programming

Stochastic programming [51] models and solves problems involving decision and chance variables, with known distributions for the latter. Problems may have one or more stages: in each stage, decisions are taken, then chance variables are observed. A solution is not a simple assignment of variables, but a policy that tells us how to make decisions given the assignments to variables from earlier stages. Stochastic programming dates back to the 1950s and is now a major area of research in mathematical programming and operations research.

We now show that discrete stochastic programs can be modelled and solved as IPs. We use a two-stage stochastic program from [52]:

$$\begin{array}{l}\max z = 1.5x_1 + 4x_2 + E[Q(x_1, x_2)] \\ \text{where } Q(x_1, x_2) \text{ is the value of} \\ \max 16y_1 + 19y_2 + 23y_3 + 28y_4 \text{ such that} \\ 2y_1 + 3y_2 + 4y_3 + 5y_4 \leq \xi_1 - x_1 \\ 6y_1 + y_2 + 3y_3 + 2y_4 \leq \xi_2 - x_2 \\ \text{where } 0 \leq x_1, x_2 \leq 5,\ y_i \in \{0, 1\}\ (i = 1, 2, 3, 4) \\ \xi_1, \xi_2 \text{ uniformly distributed on } \{5, 5.5, \ldots, 15\}\end{array}$$

A solution to this problem consists of fixed assignments to x_1, x_2, and assignments to the y_i that depend on the random ξ values. The optimum policy has an objective of 61.32.

We model the problem as an IP as follows. The variables are

$$\mathcal{V} = \langle x_1, x_2, \xi_1, \xi_2, y_1, y_2, y_3, y_4 \rangle$$

the agents are

$$\mathcal{A} = \langle \textbf{opt}, \textbf{opt}, \textbf{chance}, \textbf{chance}, \textbf{opt}, \textbf{opt}, \textbf{opt}, \textbf{opt} \rangle$$

(we choose the name **opt** for the non-random agent), the domains are

$$\mathcal{D} = \langle D_1, D_1, D_2, D_2, D_3, D_3, D_3, D_3 \rangle$$

where $D_1 = \{0, \ldots, 5\}$, $D_2 = \{5, 5.5, \ldots, 15\}$ and $D_3 = \{0, 1\}$, $\mathcal{P}$ assigns equal probabilities to the ξ_i domain values and $\mathcal{O} = \emptyset$.

As in Section 3.1, we model constraints by minimising their violations as an additional objective, choosing a large weight to prioritise feasibility. However, in this problem, we also have an objective so we shall use two utilities: the first to minimise the number of constraint violations to 0, and the second to maximise z. Hence, $\mathcal{U} = \langle \langle viol, z \rangle \rangle$. The utility hyperparameter is $\mathrm{Y} = \langle \langle \langle -10, 0.99 \rangle, \langle 100, 0.01 \rangle \rangle \rangle$.

3.6.1. A Sparse Model

In this IP, the links are $\mathcal{L} = \{\xi_i \rightarrow y_j | \forall i, j\}$. In multiple runs of up to 10^7 episodes, InfProg found policies with an expected utility of approximately 55. This is quite good but short of the known optimum of 61.32. Increasing to 10^9 episodes made no difference.

3.6.2. A Dense Model

In an attempt to improve the results, we added the following links to the sparse IP: $x_1 \rightarrow x_2$, $x_i \rightarrow y_j$ $(\forall i, j\}$ and $y_i \rightarrow y_j$ $(i < j)$. In multiple runs of 10^8 episodes, InfProg found policies with an expected reward approximately 60, which is close to the optimum.

3.7. Chance-Constrained Programming

A *chance* (or *probabilistic*) *constraint* is a constraint that should be satisfied with some probability threshold: $\Pr[C(x, r)] \geq \theta$ for a constraint C on decision variables x and chance variables r. Chance-constrained programming [53] is a method of optimising under uncertainty and has many applications, as it is a natural way of modelling uncertainty. Chance constraints are usually inequalities but we allow any form of constraint. Chance constraints

have also been used in the area of safe (or constrained) RL using various approaches [54,55], and our approach is related to that of [3]. They do not seem to have been added to IDs.

We model chance constraints by adding a new objective for each, with a reward of 1 for satisfaction, and 0 for violation, and setting the utopian value for that objective to the desired probability threshold. The weight attached to the objective should be high.

For example, we modify the stochastic program of Section 3.6 by attaching probabilities to the two hard constraints:

$$\Pr[2y_1 + 3y_2 + 4y_3 + 5y_4 \leq \xi_1 - x_1] = 0.7 \quad (a)$$
$$\Pr[6y_1 + y_2 + 3y_3 + 2y_4 \leq \xi_2 - x_2] = 0.9 \quad (b)$$

Notice that the chance constraints are of the form $\Pr[\cdot] = p$ instead of the usual $\Pr[\cdot] \geq p$. INFPROG is not guaranteed to satisfy the probability thresholds because of its multi-objective approach: instead of forcing the satisfaction of a chance constraint to exceed a threshold probability, it tries to match the threshold in a trade-off with other objectives. However, it can be used as an exploratory tool to find a policy with the desired characteristics, and the user can iteratively increase thresholds.

To model this problem, we modify the dense model from Section 3.6, adding a new objective for each chance constraint. The utility vector is now $\mathcal{U} = \langle\langle viol_a, viol_b, z\rangle\rangle$. The utility hyperparameter is

$$\mathrm{Y} = \langle\langle\langle 0.7, 0.496\rangle, \langle 0.9, 0.496\rangle, \langle 100, 0.008\rangle\rangle\rangle$$

To solve this tri-objective IP, we experimented with different weights and found a variety of policies, with different compromises between the chance constraints and original objective. Not all solutions were useful, but using weights $(0.496, 0.496, 0.008)$, we found a policy with objectives $(0.74, 0.95, 71)$ that satisfies the requirements. Relaxing the hard constraints to chance constraints allowed us to increase the objective z from 60 to 71.

3.8. Multi-Level Programming

Many problems in economics, diplomacy, war, politics, industry, gaming, and other areas involve multiple agents, which form part of each other's environment. Multi-agent RL can be applied to these problems, as can several forms of ID: bi-agent IDs [7], multi-agent IDs [8], game theory-based IDs [10], networks of IDs [6], and interactive dynamic IDs [9]. LIMIDs can model multiple agents, but only the cooperative case in which they all have the same objective.

The most common case is a *bi-level program* or *Stackelberg game*. These usually involve continuous variables and relatively little work has been performed on discrete bi-level programs [56], so new methods are of interest. However, as an even more challenging case, we use a discrete *tri-level program* studied in at least two publications [57,58] and shown in Figure 4.

$\max_{x_1} z_1 = (x_1 + x_2 + 2x_3 + 4)(-x_1 - x_2 + x_3 + 2x_4 + 1)$
where x_2 solves
$\max_{x_2} z_2 = 2x_2 + x_3 + 3x_4$
where $x_3, x_3, x_3, x_3, x_3, x_3$ solve, for given x_1, x_2
$\max_{x_3 \ldots x_8} z_3 = \frac{2x_1 + 3x_2 + 2x_3 - 3x_4}{5x_1 + 11x_2 + x_5 + 29}$
subject to
$-3x_1 + 7x_2 + x_3 + x_5 = 10 \quad 14x_1 + 4x_2 + x_6 = 6$
$x_1 + x_2 + x_3 - x_4 + x_7 = 5 \quad 2x_1 + x_2 + 2x_4 + x_8 = 8$
$0 \leq x_1 \leq 5 \quad 0 \leq x_2 \leq 2 \quad 5 \leq x_3 \leq 20 \quad 4 \leq x_4 \leq 30$
$1 \leq x_5 \leq 20 \quad 0 \leq x_6 \leq 30 \quad 0 \leq x_7 \leq 15 \quad 0 \leq x_8 \leq 20$

Figure 4. A tri-level non-linear program.

Tri-level programs have been used to model problems in supply chain management, network defence, planning, logistics, and economics. They involve three agents: the first is the *top-level leader*, the second is the *middle-level*) follower, and the third is the (*bottom-level*) follower. They are organised hierarchically: the leader makes decisions first, the

middle-level follower reacts, then the bottom-level follower reacts. A survey on multi-level programming is provided in [59], including a section on the tri-level case. Decentralised decision-making problems in a hierarchical management system often contain more than two levels. As a different application, we mention [60] which uses tri-level programming to model a defender–attacker–defender problem in defending an electrical power grid.

This problem is also of interest because its objectives are quadratic for the leader, linear for the middle-level follower, and fractional for the bottom-level follower. Non-linear objectives were added to IDs in [5] using non-linear optimal control approximations.

We model the problem using the following IP. The variables are $\mathcal{V} = \langle x_1, \ldots, x_8 \rangle$, and their domains are ranges of integer values $\mathcal{D} = \langle 0-5, 0-2, 5-20, 4-30, 1-20, 0-30, 0-15, 0-20 \rangle$. The agents are $\mathcal{A} = \langle \textbf{leader}, \textbf{middle}, \textbf{bottom}, \ldots, \textbf{bottom} \rangle$: v_1 is a top-level leader variable, v_2 is a middle-level leader variable, and v_3–v_8 are follower variables. The links are $\mathcal{L} = \{x_i \to x_j | i < j\}$ (a dense model in the sense of Section 3.1), the utilities are

$$\mathcal{U} = \langle \langle viols, z_1 \rangle, \langle viols, z_2 \rangle, \langle viols, z_3 \rangle \rangle$$

and $\mathcal{O} = \emptyset$. The first objective used by all agents is the number of constraint violations, which should be 0: the constraints ensure that any solution is *tri-level feasible* so all agents should be penalised for violations (in this problem, no agent is allowed to make a decision that leads to a constraint violation). The secondary objectives are z_1, z_2, z_3. Using a utility hyperparameter

$$\mathrm{Y} = \langle \langle \langle 0, 0.99999 \rangle, \langle 1000, 0.00001 \rangle \rangle, \langle \langle 0, 0.999 \rangle, \langle 100, 0.001 \rangle \rangle, \langle \langle 0, 0.99 \rangle, \langle 1, 0.01 \rangle \rangle \rangle$$

This is the most complex Y value we used, so we provide some explanation. The trilevel problem has been modelled using three agents, each with two objectives (one for the problem objectives and one to maximise constraint satisfaction). We use weighted metric scalarisation to reduce this to three agents with one objective each, so we must specify three utopian points (one per agent). Each utopian point has two ideal objective values and an associated weight for each value. So, for example, agent 1 has a utopian point $\langle 0, 0.99999 \rangle$ with associated weights $\langle 1000, 0.00001 \rangle$.

In repeated runs with 10^7 episodes, InfProg reliably finds the correct policy (0, 0, 9, 4, 1, 6, 0, 0) with objectives $((0, 396), (0, 21), (0, 0.2))$. As pointed out in [17], the first objective is given as 612 in [58], but it can be verified that the solution yields $z_1 = 396$. The suboptimal solutions that it sometimes finds can be recognised by their lower z_1 values.

3.9. Bayesian Games

The assumptions of classical game theory require complete information, which is often unrealistic. Bayesian games were a development that allowed incomplete (private or secret) information while avoiding infinite calculations [61]. An important feature is that a player may have a "type" that is not known by other players, representing their state of mind.

As a simple example, we take a well-known problem called the *Sheriff's dilemma*. There are two players: a sheriff and an armed suspect. The suspect has two possible types, a criminal or a civilian, and only he knows what type he is. The sheriff must know whether to shoot without knowing the suspect's type, while the suspect is allowed to use that information. Payoff matrices for the various possibilities are shown in Figure 5, where in each case, the first figure is a suspect payoff and the second a sheriff payoff.

We can model this as an IP as follows. The variables are $\mathcal{V} = \langle type, suspect, sheriff \rangle$ with domains

$$\mathcal{D} = \langle \{criminal, civilian\}, \{shoot, not\}, \{shoot, not\} \rangle$$

the agents are $\mathcal{A} = \langle \textbf{chance}, \textbf{suspect}, \textbf{sheriff} \rangle$. The links are $\mathcal{L} = \{type \to suspect\}$ allowing the suspect but not the sheriff to know the suspect's type. The type probabilities $\mathcal{P}$ are p for $\Pr[criminal]$ and $1 - p$ for $\Pr[civilian]$, the utilities are $\mathcal{U} = \langle \langle ususpect \rangle, \langle usheriff \rangle \rangle$, and $\mathcal{O} = \emptyset$.

Using a utility hyperparameter $\Upsilon = \langle\langle\langle 100, 1\rangle\rangle, \langle\langle 100, 1\rangle\rangle\rangle$ and 1000 episodes, INFPROG reliably finds the correct policy: the suspect shoots if and only if he is a criminal, and the sheriff shoots if and only if $p > \frac{1}{3}$. If p is close to $\frac{1}{3}$, then more episodes are needed to reduce sampling error.

type: criminal		sheriff action: shoot	sheriff action: not
suspect action	shoot	0,0	2,−2
	not	−2,−1	−1,1

type: civilian		sheriff action: shoot	sheriff action: not
suspect action	shoot	−3,−1	−1,−2
	not	−2,−1	0,0

Figure 5. Payoffs for the Sheriff's Dilemma.

3.10. Level-k Reasoning

Bayesian games have also been criticised for making unrealistic assumptions that do not always predict real-world behaviour. Another game theory approach to incomplete information is *level-k reasoning* [62,63], which assumes that players play strategically but with bounded rationality. It has recently been used in *adversarial risk analysis* for modelling problems in counter-terrorism [64] and other fields.

We consider a simple level-k problem: the *Keynesian beauty contest*. Contest participants are asked to choose a number that they hope will be as close as possible to some fraction p of the average of all participants' choices. Classical game theory predicts that all players will choose 0, which is the Nash equilibrium. The reasoning is that, if all players choose randomly, it is best to choose p times the mean: $50p$ or approximately 33 when $p = \frac{2}{3}$ (which is a popular value). But all players know this, so they should instead choose p times that value approximately 25. But all players know this, so the logical conclusion is that 0 is the only reasonable choice, yet in experiments, few people choose 0.

In level-k reasoning, we make assumptions about the other players. We might assume that all other players are level-0 thinkers who do not think strategically. They simply choose a number uniformly at random: $c_0 \in U[0, 100]$ (we shall restrict our numbers to be integers as our method is discrete). Under this assumption, we should use level-1 thinking and choose p times the expected level-0 choice: 33. If we instead assume that all other players are level-1, we should use level-2 thinking and choose 22. Similarly, if we assume that all others are level-2, we should use level-3 thinking and choose 15.

We can model and solve these problems as IPs, as we now show using the beauty contest. The variables represent the choices of the hypothetical level-k players (k $= 0, 1, 2$) and the real level-3 player: $\mathcal{V} = \langle c_0, c_1, c_2, c_3\rangle$. The domains in $\mathcal{D}$ are all $\{0, 1, 2, \ldots, 100\}$. The agents are $\mathcal{A} = \langle$**chance**, **P1**, **P2**, **P3**$\rangle$. The **chance** probabilities are all $\frac{1}{101}$. There are no links as all choices are made simultaneously, without knowing the other choices: $\mathcal{L} = \emptyset$. There are no observations so $\mathcal{O} = \emptyset$. The level-i player should guess as close as possible to p times the value they expect from the level-$(i-1)$ player, i.e., $|c_i - pc_{i-1}|$ should be minimised, so the utilities are $\langle\langle|c_1 - pc_0|\rangle, \langle|c_2 - pc_1|\rangle, \langle|c_3 - pc_2|\rangle\rangle$. Applying INFPROG with 10^6 episodes and a utility hyperparameter $\Upsilon = \langle\langle\langle 0, 1\rangle\rangle, \langle\langle 0, 1\rangle\rangle, \langle\langle 0, 1\rangle\rangle\rangle$ to minimise all utilities, we find c_1, taking values approximately 29–37, c_2 19–25 and c_3 13–15. Using more episodes reduces the variance in these numbers.

3.11. Practical Implications

We have demonstrated that a wide range of problems from diverse areas of literature can be modelled as IPs, and that these can be solved by the INFPROG solver. Of course, we do not claim that *any* problem from Operations Research, Game Theory and so on can be tackled by our approach, especially as it can currently handle only discrete variables. However, we believe that it has interesting practical implications.

Firstly, our approach is useful for rapid prototyping. Few researchers are masters of stochastic programming, dynamic programming, game theory, machine learning, and multiple other areas. When faced with a new complex optimisation problem, it is not always clear how to model it, and in practice, we often make simplifying assumptions based on our background. A stochastic programmer might simplify a multi-agent problem

by modelling an adversary using random variables, or pretend that decisions cannot affect chance variable distributions, or approximate a multi-stage problem by a two-stage one. A constraint programmer or integer programmer might approximate a chance variable by an expectation. A game theorist might assume a zero-sum game or complete knowledge, even when this is unrealistic. In general, it is hard to know whether our design choices are reasonable. We believe that a tool such as INFPROG is useful during the early stages of modelling, as it enables us to explore the consequences of various simplifications without the need to master multiple research areas: we can simply change the problem formulation and observe the result. Some simplifications might cause little change in solution quality, while others might greatly reduce quality or even make a problem infeasible. The eight-queens and stochastic program models can be seen as examples: in both cases, we implemented different models and compared the resulting objective values, leading to the conclusion that dense models give better results.

Secondly, we hope that our approach will be useful in its own right. It is based on multi-agent multi-objective RL, a very general paradigm with wide applications. For specific problems, we do not expect INFPROG to be competitive with specialised algorithms, but for problems that are inherently complex, it might be a useful tool, as there is little available software able to tackle problems involving multiple agents, multiple objectives, random events, and partial knowledge.

It should be noted that INFPROG is merely a research prototype and we certainly do not consider it to be a finished product. We hope that more competitive IP algorithms will emerge in the future, especially based on deep RL and implemented on highly parallel hardware. This should greatly enhance the scalability of the IP approach to complex optimisation problems, as RL has already shown its ability to tackle large multi-agent problems such as learning to play Go.

3.12. A Small Case Study

To illustrate our approach in more detail, we show the programmable utility for the Oil Wildcatter of Section 3.3 problem in Figure 6. The user must provide a C function called `utility` with two arguments: a one-dimensional array `v` of integer chance and decision variables and a two-dimensional array `u` of utilities. This function is called at the end of an episode, so we can assume that all variables in `v` have been assigned values. (The `*` and `**` are C notations indicating array dimensionality).

```
void utility(int* v, float** u)
{
   int oil=v[0], test=v[1], seismic=v[2], drill=v[3],
       no=0, yes=1, dry=0, wet=1, soak=2, company=1;
   float drill_payoff, test_payoff;

   if (test==yes) test_payoff= -10.0; else test_payoff=0.0;

   if (oil==dry  && drill==yes) drill_payoff= -70.0; else
   if (oil==wet  && drill==yes) drill_payoff=  50.0; else
   if (oil==soak && drill==yes) drill_payoff= 200.0; else
                                drill_payoff=   0.0;

   u[company][0]=drill_payoff+test_payoff;
}
```

Figure 6. Programmable utility for the oil wildcatter example.

For readability, we have created integer variables to name the IP variables (for example `oil=v[0]`), their possible values (for example `yes=0`) and the only agent (`company`—note that chance variables are assigned agent number 0). The total payoff is the sum of the drill and test payoffs, and this sum is assigned to the only utility `u[1][0]` in the problem. The second index is `0` because an agent's utilities are numbered from 0: for a bi-objective problem such as that in Section 3.4, we would also need to assign a value to `u[company][1]`.

The user must also provide a small text file describing the parameters of the problem (number of agents, variables, domain sizes, and so on). We do not show an example as INFPROG is currently a research prototype and lacks a user-friendly specification language.

4. Conclusions

This paper makes two contributions. Firstly, it defines a new class of optimisation problem called an influence program (IP) that is sufficiently general to encompass a wide variety of problems from operations research, artificial intelligence, game theory, and adversarial risk analysis. Secondly, it proposes an algorithm called INFPROG based on reinforcement learning plus sampling. INFPROG is a lightweight solver that does not need a graphics processing unit or other specialised hardware. We demonstrated the flexibility of IP using examples taken from diverse literatures: to the best of our knowledge, no other solver can tackle this range of optimisation problems. Our work also shows that many existing problems from other literatures can serve as benchmark problems for multi-objective and/or multi-agent reinforcement learning algorithms, of which there is a lack [18].

We expect IPs to be useful for hybrid optimisation problems that do not fit well into any particular class and have no available solvers. A fruitful source of applications might be safe reinforcement learning, in which constraints are used to ensure robust and safe policies. It might also be useful for the rapid prototyping of different models, to investigate how to model a new problem, to test the effects of hiding information, or making simplifying assumptions. Finally, it could be used as an educational tool for understanding the different models of optimisation problems.

In future work, we shall explore the application of our method to complex real-world problems. The current INFPROG implementation is a research prototype only, intended to demonstrate the potential of a unified approach to a wide range of complex optimisation problems. It can handle only discrete variables, uses a simple and quite weak form of state aggregation (random tile coding) and lacks a clear user-interface. We hope to design a more user-friendly version that also includes continuous variables, and to use more powerful state aggregation: the latter two improvements might be achieved by moving from SARSA to an actor–critic algorithm such as proximal policy optimisation [65].

Funding: This material is based upon works supported by the Science Foundation Ireland under Grant No. 12/RC/2289-P2 which is co-funded under the European Regional Development Fund. We would also like to acknowledge the support of the Science Foundation Ireland CONFIRM Centre for Smart Manufacturing, Research Code 16/RC/3918, and the EU H2020 ICT48 project TAILOR under contract #952215.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Sutton, R.S.; Barto, A.G. *Reinforcement Learning: An Introduction*; MIT Press: Cambridge, MA, USA, 1998.
2. Silver, D.; Hubert, T.; Schrittwieser, J.; Antonoglou, I.; Lai, M.; Guez, A.; Lanctot, M.; Sifre, L.; Kumaran, D.; Graepel, T.; Lillicrap, T.; Simonyan, K.; Hassabis, D. A General Reinforcement Learning Algorithm That Masters Chess, Shogi, and Go Through Self-Play. *Science* **2018**, *362*, 1140–1144. [CrossRef] [PubMed]
3. Huang, S.H.; Abdolmaleki, A.; Vezzani, G.; Brakel, P.; Mankowitz, D.J.; Neunert, M.; Bohez, S.; Tassa, Y.; Heess, N.; Riedmiller, M.A.; Hadsell, R. A Constrained Multi-Objective Reinforcement Learning Framework. In Proceedings of the 5th Conference on Robot Learning, London, UK, 8–11 November 2021; pp. 883–893.
4. Elshafei, M.M.K.; El-Sherberry, M.S. Interactive Bi-level Multiobjective Stochastic Integer Linear Programming Problem. *Trends Appl. Sci. Res.* **2008**, *3*, 154–164.
5. Kratochvíl, V.; Vomlel, J. Influence Diagrams for Speed Profile Optimization. *Int. J. Approx. Reason.* **2017**, *88*, 567–586. [CrossRef]

6. Gal, Y.; Pfeffer, A. Networks of Influence Diagrams: A Formalism for Representing Agents' Beliefs and Decision-Making Processes. *J. Artif. Intell. Res.* **2008**, *33*, 109–147. [CrossRef]
7. González-Ortega, J.; Insua, D.R.; Cano, J. Adversarial Risk Analysis for Bi-agent Influence Diagrams: An Algorithmic Approach. *Eur. J. Oper. Res.* **2019**, *273*, 1085–1096. [CrossRef]
8. Koller, D.; Milch, B. Multi-Agent Influence Diagrams for Representing and Solving Games. *Games Econ. Behav.* **2001**, *45*, 181–221. [CrossRef]
9. Polich, K.; Gmytrasiewicz, G. Interactive Dynamic Influence Diagrams. In Proceedings of the 6th International Joint Conference on Autonomous Agents and Multiagent Systems, Communications in Computer and Information Science, Honolulu, HI, USA, 14–18 May 2007; Volume 288, pp. 623–630.
10. Zhou, L.H.; Kevin, L.; Liu, W.Y. Game theory-based Influence Diagrams. *Expert Syst.* **2013**, *30*, 341–351. [CrossRef]
11. Diehl, M.; Haimes, Y. Influence Diagrams With Multiple Objectives and Tradeoff Analysis. *IEEE Trans. Syst. Man Cybern. Part A* **2004**, *34*, 293–304. [CrossRef]
12. Marinescu, R.; Razak, A.; Wilson, N. Multi-objective Influence Diagrams. In Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence, Catalina Island, CA, USA, 5–17 August 1 2012; pp. 574–583.
13. Jenzarli, A. Information/Relevance Influence Diagrams. In Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence (UAI), Montreal, QC, Canada, 18–20 August 1995; pp. 329–337.
14. Lauritzen, S.L.; Nilsson, D. Representing and Solving Decision Problems With Limited Information. *Manag. Sci.* **2001**, *47*, 1238–1251. [CrossRef]
15. Powell, W.B. *Reinforcement Learning and Stochastic Optimization: A Unified Framework for Sequential Decisions*; Wiley: Hoboken, NJ, USA, 2022.
16. Prestwich, S.D.; Toffano, F.; Wilson, N. A Probabilistic Programming Language for Influence Diagrams. In Proceedings of the 11th International Conference on Scalable Uncertainty Management, Granada, Spain, 4–6 October 2017.
17. Prestwich, S.D. Solving Mixed Influence Diagrams by Reinforcement Learning. In Proceedings of the 9th International Conference on Machine Learning, Optimization, and Data Science, Grasmere, UK, 22–26 September 2023.
18. Hayes, C.F.; Rădulescu, R.; Bargiacchi, E.; Källström, J.; Macfarlane, M.; Reymond, M.; Verstraeten, T.; Zintgraf, L.M.; Dazeley, R.; Heintz, F.; et al. A Practical Guide To Multi-Objective Reinforcement Learning and Planning. *Auton. Agent. Multi-Agent Syst.* **2022**, *36*, 26. [CrossRef]
19. van Moffaert, K.; Drugan, M.M.; Nowé, A. Scalarized Multi-Objective Reinforcement Learning: Novel Design Techniques. In Proceedings of the IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning, Singapore, 16–19 April 2013; pp. 191–199.
20. Zhang, K.; Yang, Z.; Başar, T. Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms. In *Handbook of Reinforcement Learning and Control*; Vamvoudakis, K.G., Wan, Y., Lewis, F.L., Cansever, D., Eds.; Studies in Systems, Decision and Control; Springer: Cham, Switzerland, 2021; Volume 325.
21. Rădulescu, R.; Mannion, P.; Roijers, D.M.; Nowé, A. Multi-Objective Multi-Agent Decision Making: A Utility-Based Analysis and Survey. *Auton. Agents Multi-Agent Syst.* **2020**, *34*, 10. [CrossRef]
22. Hu, T.; Luo, B.; Yang, C.; Huang, T. MO-MIX: Multi-Objective Multi-Agent Cooperative Decision-Making with Deep Reinforcement Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 12098–12112. [CrossRef] [PubMed]
23. Andrychowicz, M.; Wolski, F.; Ray, A.; Schneider, J.; Fong, R.; Welinder, P.; McGrew, B.; Tobin, J.; Abbeel, P.; Zaremba, W. Hindsight Experience Replay. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5055–5065.
24. Zobrist, A.L. A New Hashing Method with Application for Game Playing. *ICGA J.* **1990**, *13*, 69–73.
25. Hyatt, R.M.; Cozzie, A. The Effect of Hash Signature Collisions in a Chess Program. *ICGA J.* **2005**, *28*, 131–139. [CrossRef]
26. Skalse, J.; Hammond, L.; Griffin, C.; Abate, A. Lexicographic Multi-Objective Reinforcement Learning. In Proceedings of the 31st International Joint Conference on Artificial Intelligence, Vienna, Austria, 23–29 July 2022; pp. 3430–3436.
27. Gábor, Z.; Kalmár, Z.; Szepesvári, C. Multi-Criteria Reinforcement Learning. In Proceedings of the 15th International Conference on Machine Learning, Madison, WA, USA, 24–27 July 1998; pp. 197–205.
28. Giagkiozis, I.; Fleming, P.J. Methods for Multi-Objective Optimization: An Analysis. *Inf. Sci.* **2015**, *293*, 338–350. [CrossRef]
29. Gordon, A.D.; Henzinger, T.A.; Nori, A.V.; Rajamani, S.K. Probabilistic Programming. In Proceedings of the International Conference on Software Engineering, Hyderabad, India, 31 May–7 June 2014.
30. Pfeffer, A. *Practical Probabilistic Programming*; Manning Publications: Shelter Island, NY, USA, 2016.
31. Rossi, F.; van Beek, P.; Walsh, T. (Eds.) *Handbook of Constraint Programming*; Elsevier: Amsterdam, The Netherlands, 2006.
32. Crites, R.H.; Barto, A.G. Elevator Group Control Using Multiple Reinforcement Learning Agents. *Mach. Learn.* **1998**, *33*, 235–262. [CrossRef]
33. Useng, M.; Avdulrahman, S. A Survey on Distributed Reinforcement Learning. *Mesopotamian J. Big Data* **2022**, *2022*, 44–50. [CrossRef]
34. Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*; Series in Representation and Reasoning; Morgan Kaufman Publishers: Burlington, MA, USA, 1988.
35. Howard, R.A.; Matheson, J.E. Influence Diagrams. In *Readings in Decision Analysis*; Strategic Decisions Group: Menlo Park, CA, USA, 1981; Chapter 38, pp. 763–771.

36. Nease, R.F., Jr.; Owens, D.K. Use of Influence Diagrams to Structure Medical Decisions. *Med. Decis. Mak.* **1997**, *17*, 263–275. [CrossRef] [PubMed]
37. Chockalingam, S.; Maathuis, C. Influence Diagrams in Cyber Security: Conceptualization and Potential Applications. In Proceedings of the 22nd European Conference on Cyber Warfare and Security, Piraeus, Greece, 22–23 June 2023; Volume 22.
38. Carriger, J.F.; Newman, M.C. Influence Diagrams as Decision-Making Tools for Pesticide Risk Management. *Integr. Environ. Assess. Manag.* **2011**, *8*, 339–350. [CrossRef]
39. Dechter, R. A New Perspective on Algorithms for Optimizing Policies Under Uncertainty. In *Artificial Intelligence Planning Systems*; Association for the Advancement of Artificial Intelligence: Washington, DC, USA, 2000; pp. 72–81.
40. Jensen, F.; Jensen, V.; Dittmer, S. From Influence Diagrams to Junction Trees. In *Uncertainty in Artificial Intelligence*; Morgan Kaufmann: Burlington, MA, USA, 1994; pp. 367–373.
41. Shachter, R.D. Evaluating Influence Diagrams. *Oper. Res.* **1986**, *34*, 871–882. [CrossRef]
42. Shenoy, P. Valuation-Based Systems for Bayesian Decision Analysis. *Oper. Res.* **1992**, *40*, 463–484. [CrossRef]
43. Tatman, J.A.; Shachter, R.D. Dynamic Programming and Influence Diagrams. *IEEE Trans. Syst. Man Cybern.* **1990**, *20*, 365–379. [CrossRef]
44. Cano, A.; Gómez, M.; Moral, S. A Forward-Backward Monte Carlo Method for Solving Influence Diagrams. *Int. J. Approx. Reason.* **2006**, *42*, 119–135. [CrossRef]
45. Charnes, J.M.; Shenoy, P.P. Multistage Monte Carlo Method for Solving Influence Diagrams Using Local Computation. *Manag. Sci.* **2004**, *50*, 405–418. [CrossRef]
46. Marinescu, R.; Lee, J.; Dechter, R. A New Bounding Scheme for Influence Diagrams. In Proceedings of the 35th Conference on Artificial Intelligence, Vancouver, BC, Canada, 2–9 February 2021; Association for the Advancement of Artificial Intelligence: Washington, DC, USA, 2021; pp. 12158–12165.
47. Watthayu, W. Representing and Solving Influence Diagram in Multi-Criteria Decision Making: A Loopy Belief Propagation Method. In Proceedings of the International Symposium on Computer Science and its Applications, Hobart, TAS, Australia, 13–15 October 2008; pp. 118–125.
48. Yuan, C.; Wu, X. Solving Influence Diagrams Using Heuristic Search. In Proceedings of the International Symposium on Artificial Intelligence and Mathematics, Sanya, China, 23–24 October 2010.
49. Everitt, T.; Kumar, R.; Krakovna, V.; Legg, S. Modeling AGI Safety Frameworks with Causal Influence Diagrams. In Proceedings of the Workshop on Artificial Intelligence Safety, CEUR Workshop, Honolulu, HI, USA, 27 January 2019; Volume 2419.
50. Raiffa, H. *Decision Analysis*; Addison-Wesley: Reading, MA, USA, 1968.
51. Birge, J.R.; Louveaux, F.V. *Introduction to Stochastic Programming*; Springer: New York, NY, USA, 2011.
52. Ahmed, S.; Tawarmalani, M.; Sahinidis, N.V. A Finite Branch-and-Bound Algorithm for Two-Stage Stochastic Integer Programs. *Math. Program.* **2004**, *100*, 355–377. [CrossRef]
53. Charnes, A.; Cooper, W.W. Chance-Constrained Programming. *Manag. Sci.* **1959**, *6*, 73–79. [CrossRef]
54. García, J.; Fernández, F. A Comprehensive Survey on Safe Reinforcement Learning. *J. Mach. Learn. Res.* **2015**, *16*, 1437–1480.
55. Gu, S.; Yang, L.; Du, Y.; Chen, G.; Walter; Wang, J.; Yang, Y.; Knoll, A.C. A Review of Safe Reinforcement Learning: Methods, Theory and Applications. *arXiv* **2022**, arXiv:2205.10330.
56. Kovacs, A.; Kis, T. Constraint Programming Approach to a Bilevel Scheduling Problem. *Constraints* **2011**, *16*, 317–340. [CrossRef]
57. Arora, R.; Arora, S.R. An Algorithm for Non-Linear Multi-Level Integer Programming Problems. *Int. J. Comput. Sci. Math.* **2010**, *3*, 211–225. [CrossRef]
58. Mishra, S.; Verma, A.B. A Non-Differential Approach for Solving Tri-Level Programming Problems. *Am. Int. J. Res. Sci. Technol. Math.* **2015**.
59. Lu, J.; Han, J.; Hu, Y.; Zhang, G. Multilevel Decision-Making: A Survey. *Inf. Sci.* **2016**, *346–347*, 463–487. [CrossRef]
60. Alguacil, N.; Delgadillo, A.; Arroyo, J.M. A Trilevel Programming Approach for Electric Grid Defense Planning. *Comput. Oper. Res.* **2014**, *41*, 282–290. [CrossRef]
61. Harsanyi, J.C. Games with Incomplete Information Played by "Bayesian" Players, I–III: Part I. The Basic Model. *Manag. Sci.* **2004**, *50*, 1804–1817. [CrossRef]
62. Nagel, R. Unraveling in Guessing Games: An Experimental Study. *Am. Econ. Rev.* **1995**, *85*, 1313–1326.
63. Stahl, D.O.; Wilson, P.W. On Players' Models of Other Players: Theory and Experimental Evidence. *Games Econ. Behav.* **1995**, *10*, 218–254. [CrossRef]
64. Rothschild, C.; Albert, L.; Guikema, S. Adversarial Risk Analysis with Incomplete Information: A Level-k Approach. *Risk Anal.* **2011**, *32*, 1219–1231. [CrossRef]
65. Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; Klimov, O. Proximal Policy Optimization Algorithms. *arXiv* **2017**, arXiv:1707.06347.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

AppliedMath

MDPI

Article

Multi–Dimensional Data Analysis of Deep Language in J.R.R. Tolkien and C.S. Lewis Reveals Tight Mathematical Connections

Emilio Matricciani

Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB), Politecnico di Milano, 20133 Milan, Italy; emilio.matricciani@polimi.it

Abstract: Scholars of English Literature unanimously say that J.R.R. Tolkien influenced C.S. Lewis's writings. For the first time, we have investigated this issue mathematically by using an original multi-dimensional analysis of linguistic parameters, based on surface deep language variables and linguistic channels. To set our investigation in the framework of English Literature, we have considered some novels written by earlier authors, such as C. Dickens, G. MacDonald and others. The deep language variables and the linguistic channels, discussed in the paper, are likely due to writers' unconscious design and reveal connections between texts far beyond the writers' awareness. In summary, the capacity of the extended short-term memory required to readers, the universal readability index of texts, the geometrical representation of texts and the fine tuning of linguistic channels within texts—all tools largely discussed in the paper—revealed strong connections between *The Lord of the Rings* (Tolkien), *The Chronicles of Narnia*, *The Space Trilogy* (Lewis) and novels by MacDonald, therefore agreeing with what the scholars of English Literature say.

Keywords: alphabetical languages; extended short-term memory; human communication; human mind; sentences: mathematical modeling; universal readability index

check for updates

Citation: Matricciani, E. Multi–Dimensional Data Analysis of Deep Language in J.R.R. Tolkien and C.S. Lewis Reveals Tight Mathematical Connections. *AppliedMath* **2024**, *4*, 927–949. https://doi.org/10.3390/appliedmath4030050

Academic Editor: Libor Pekař

Received: 17 June 2024
Revised: 29 July 2024
Accepted: 31 July 2024
Published: 1 August 2024

Copyright: © 2024 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Unanimously, in a large number of papers—some of which are recalled here [1–8] from the vast literature on the topic—scholars of English Literature state that J.R.R. Tolkien influenced C.S. Lewis's writings. The purpose of the present paper is not to review the large wealth of literature based on the typical approach used by scholars of literature—which is not our specialty—but to investigate this issue mathematically and statistically—a study that has never been conducted before—by using recent methods devised by researching the impact of the surface deep language variables [9,10] and linguistic channels [11] in literary texts. Since scholars mention the influence of George MacDonald on both, we consider some novels written by this earlier author. To set all these novels in the framework of English Literature, we consider some novels written by other earlier authors, such as C. Dickens and others.

After this introduction, in Section 2, we introduce the literary texts (novels) considered. In Section 3, we report the series of words, sentences and interpunctions versus chapters for some novels, and define an index useful to synthetically describe regularity due to what we think is a conscious design by authors. In Section 4, we start exploring the four deep language variables; to avoid misunderstanding, these variables, and the linguistic channels derived from them, refer to the "surface" structure of texts, not to the "deep" structure mentioned in cognitive theory. In Section 5, we report results concerning the extended short-term memory and a universal readability index; both topics address human short-term memory buffers. In Section 6, we represent literary texts geometrically in the Cartesian plane by defining linear combinations of deep language variables and calculate the probability that a text can be confused with another. In Section 7, we show the linear relationships existing between linguistic variables in the novels considered. In Section 8,

we report the theory of linguistic channels. In Section 9, we apply it to the novels presently studied. Finally, in Section 10, we summarize the main findings and conclude. Several Appendices report numerical data.

2. Database of Literary Texts (Novels)

Let us first introduce the database of literary texts used in the present paper. Table 1 lists some basic statistics of the novels by Tolkien, Lewis and MacDonald. To set these texts in the framework of earlier English Literature, we consider novels by Charles Dickens (Table 2) and other authors (Table 3).

Table 1. Novels written by Tolkien, Lewis and MacDonald, with year of publication. Number of chapters (i.e., the number of samples considered in calculating the regression lines reported below), total number of characters contained in the words (C), total number of words W and sentences (S). Titles, footnotes and other extraneous material present in the digital texts have been deleted.

John R.R. Tolkien (1892–1973)	**Chapters (M)**	**Characters (C)**	**Words (W)**	**Sentences (S)**
The Hobbit (1937)	19	394,154	95,914	5890
The Lord of the Rings (1954–1955)	62	1,906,531	472,173	34,601
The Silmarillion (posthumous, 1977)	24	429,639	101,627	3346
Clive S. Lewis (1898–1963)				
The Screwtape Letters (1942)	31	135,204	31,040	1330
The Space Trilogy (1938–1945)	123	1,243,141	295,240	20,124
The Chronicles of Narnia (1950–1956)	110	1,318,482	322,544	23,515
George MacDonald (1824–1905)				
Phantastes: A Fairie Romance for Men and Women (1858)	25	283,676	67,551	3274
At the Back of the North Wind (1871)	38	349,041	90,697	5017
The Princess and the Goblin (1872)	32	208,325	51,090	3205
Lilith: A Romance (1895)	47	386,522	94,127	6271

Table 2. Novels by Charles Dickens, with year of publication. Number of chapters (M, i.e., the number of samples considered in calculating the regression lines reported below), total number of characters contained in the words ($\boldsymbol{C}$), total number of words W and sentences ($\boldsymbol{S}$).

Novel (Year of Publication)	**Chapters (M)**	**Characters (C)**	**Words (W)**	**Sentences (S)**
The Adventures of Oliver Twist (1837–1839)	53	679,008	160,604	9121
David Copperfield (1849–1850)	64	1,469,251	363,284	19,610
Bleak House (1852–1853)	64	1,480,523	350,020	20,967
A Tale of Two Cities (1859)	45	607,424	142,762	8098
Our Mutual Friend (1864–1865)	67	1,394,753	330,593	17,409

We have used the digital text of a novel (WinWord file) and counted, for each chapter, the number of characters, words, sentences and interpunctions (punctuation marks). Before doing so, we have deleted the titles, footnotes and other extraneous material present in the digital texts, a burdensome work. The count is very simple, although time-consuming. Winword directly provides the number of characters and words. The number of sentences was calculated by using WinWord to replace every full stop with a full stop: of course, this action does not change the text, but it gives the number of these substitutions and therefore the number of full stops. The same procedure was repeated for question marks and exclamation marks. The sum of the three totals gives the total number of sentences in the text analyzed. The same procedure gives the total number of commas, colons and semicolons. The sum of these latter values with the total number of sentences gives the total number of interpunctions.

Table 3. Novels by authors of English Literature, with year of publication. Number of chapters (***M***, i.e., the number of samples considered in calculating the regression lines reported below), total number of characters contained in the words (***C***), total number of words ***W*** and sentences (***S***).

Novel (Author, Year)	Chapters (*M*)	Characters (*C*)	Words (*W*)	Sentences (*S*)
Pride and Prejudice (J. Austen, 1813)	61	537,005	121,934	6013
Vanity Fair (W. Thackeray, 1847–1848)	66	1,285,688	277,716	13,007
Moby Dick (H. Melville, 1851)	132	92,2351	203,983	9582
The Mill On The Floss (G. Eliot, 1860)	57	888,867	207,358	9018
Alice's Adventures in Wonderland (L. Carroll, 1865)	12	107,452	27,170	1629
Adventures of Huckleberry Finn (M. Twain, 1884)	42	427,473	110,997	5887
The Jungle Book (R. Kipling, 1894)	9	209,935	51,090	3214
The War of the Worlds (H.G. Wells, 1897)	27	265,499	60,556	3306
The Wonderful Wizard of Oz (L.F. Baum, 1900)	22	156,973	39,074	2219
The Hound of The Baskervilles (A.C. Doyle, 1901–1902)	15	245,327	59,132	4080
Peter Pan (J.M. Barrie, 1902)	17	194,105	47,097	31,77
Martin Eden (J. London, 1908–1909)	45	601,672	139,281	9173

Some homogeneity can be noted in novels of the same author. The stories in *The Space Trilogy* and *The Chronicles of Narnia,* by Lewis are told with about the same number of chapters, words and sentences, as is also for a couple of MacDonald's novels, such as *At the Back of the North Wind* and *Lilith: A Romance*. Some homogeneity can be found in *David Copperfield, Bleak House* and *Our Mutual Friend* (by Dickens) and in *The Adventures of Oliver Twist* and *A Tale of Two Cities*. These numerical values, we think, are not due to chance but consciously managed by the authors, which is a topic we purse more in the next section.

3. Conscious Design of Texts: Words, Sentences and Interpunctions versus Chapters

First, we study the linguistic variables which we think the authors deliberately designed. In the specifics, we show the series of words, sentences and interpunctions versus chapter.

Let us consider a literary work (a novel) and its subdivision into disjointed blocks of text long enough to give reliable average values. Let n_S be the number of sentences contained in a text block, n_W the number of words contained in the n_S sentences, n_C the number of characters contained in the n_W words and n_I the number of punctuation marks (interpunctions) contained in the n_S sentences.

Figure 1 shows the series n_W versus the normalized chapter number for *The Lord of the Rings*, *The Chronicles of Narnia*, *The Space Trilogy*.

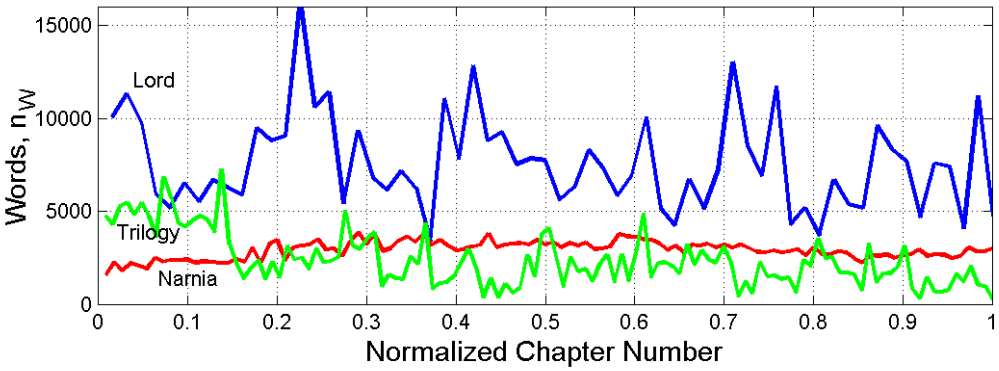


Figure 1. Series of words versus the normalized chapter number. Blue line: *The Lord of the Rings* (*Lord*); red line: *The Chronicles of Narnia* (*Narnia*); green line: *The Space Trilogy* (*Trilogy*).

For example, the normalized value of chapter 10 in *The Chronicles of Narnia*, is 10/110 = 0.09 in the x–scale of Figure 1. This normalization allows the synoptic showing of novels with a different number of chapters.

In *The Chronicles of Narnia* (in the following, *Narnia*, for brevity), we can notice a practically constant value n_W compared to *The Lord of the Rings* (*Lord*) and *The Space Trilogy* (*Trilogy*).

Let us define a synthetic index to describe the series drawn in Figure 1, namely the coefficient of variation δ, given by the standard deviation σ_{n_W} divided by the mean value $< n_W >$

$$\delta = \frac{\sigma_{n_W}}{< n_W >} \tag{1}$$

Tables 4 and 5 report δ for n_W, n_S and n_I. Since n_S and n_I are very well correlated with n_S, the three coefficients of dispersion are about the same.

Table 4. The coefficient of dispersion in the series of words, sentences and interpunctions in the indicated novels by Tolkien, Lewis and MacDonald.

Novel	Words	Sentences	Interpunctions	Average
The Hobbit	0.49	0.48	0.50	0.49
The Lord of the Rings	0.34	0.36	0.34	0.35
The Silmarillion	0.73	0.86	0.80	0.80
The Screwtape Letters	0.07	0.17	0.14	0.13
The Space Trilogy	0.60	0.61	0.58	0.59
The Chronicles of Narnia	0.16	0.20	0.20	0.19
At the Back of the North Wind	0.54	0.61	0.56	0.57
Phantastes: A Fairie Romance for Men and Women	0.66	0.73	0.63	0.67
Lilith: A Romance	0.43	0.53	0.46	0.47
The Princess and the Goblin	0.53	0.75	0.62	0.64

Table 5. The coefficient of dispersion in the series of words, sentences and interpunctions in the indicated novels.

Novel	Words	Sentences	Interpunctions	Average
The Adventures of Oliver Twist	0.31	0.33	0.32	0.32
David Copperfield	0.37	0.38	0.37	0.37
Bleak House	0.28	0.31	0.30	0.30
A Tale of Two Cities	0.52	0.57	0.52	0.54
Our Mutual Friend	0.26	0.29	0.27	0.27
Martin Eden	0.29	0.33	0.29	0.31
The Hound of The Baskervilles	0.26	0.29	0.25	0.27
Peter Pan	0.29	0.41	0.33	0.34

In Narnia $\delta = 0.16$, in Lord $\delta = 0.34$ and in Trilogy $\delta = 0.60$. Let us also notice the minimum value $\delta = 0.07$ in *The Screwtape Letters* (*Screwtape*).

The overall (words, sentences and interpunctions mixed together) mean value is $< \delta >= 0.44$ and the standard deviation $\sigma_\delta = 0.18$. Therefore, *Screwtape* is practically more than $2 \times \sigma_\delta$ from the mean, as also is *Silmarillion* on the other side, and *Narnia* is at about $1.5 \times \sigma_\delta$. In contrast, *Trilogy*, *Lord* and *The Hobbit* (*Hobbit*) are within $1 \times \sigma_\delta$.

From these results, it seems that Lewis designed the chapters of *Narnia* and *Screwtape* with an almost uniform distribution of words, sentences and interpunctions, very likely

because of the intended audience in *Narnia* (i.e., kids) and the "letters" fiction tool used in *Screwtape*. In *Trilogy* the design seems very different ($\delta = 0.60$, well within $1 \times \sigma_\delta$) likely due to the development of the science fiction story narrated.

Tolkien acted differently from Lewis, because he seems to have designed chapters more randomly and within $1 \times \sigma_\delta$, as *Hobbit* and *Lord* show. An exception is *The Silmarillion*, published posthumously, which is a text far from being a "novel".

Finally, notice that the novels by MacDonald show more homogeneous values, very similar to *Hobbit* and *Trilogy* and to the other novels listed in Table 5.

In conclusion, the analysis of series of words, sentences and interpunctions per chapter does not indicate likely connections between Tolkien, Lewis and MacDonald. Each author structured their use of words, sentences and punctuation according to distinct plans, which varied not only between authors but also between different novels by the same author.

There are, however, linguistic variables that—as we have reported for modern and ancient literary texts—are not consciously designed/managed by authors; therefore, these variables are the best candidates to reveal hidden mathematical/statistical connections between texts. In the next section, we start dealing with these variables, with the specific purpose of comparing Tolkien and Lewis, although this comparison is set in the more general framework of the authors mentioned in Section 2.

4. Surface Deep Language Variables

We start exploring the four stochastic variables we called deep language variables, following our general statistical theory on alphabetical languages [9–11]. To avoid possible misunderstandings, these variables, and the linguistic channels derived from them, refer to the "surface" structure of texts, not to the "deep" structure mentioned in cognitive theory.

Contrarily to the variables studied in Section 3, the deep language variables are likely due to unconscious design. As shown in [9–11], they reveal connections between texts far beyond writers' awareness; therefore, the geometrical representation of texts [10] and the fine tuning of linguistic channels [11] are tools better suited to reveal connections. They can also likely indicate the influence of an author on another.

We defined the number of characters per chapter n_C and the number of I_P's per chapter n_{I_P}, and the four deep language variables are [9] the number of characters C_P:

$$\boldsymbol{C_P = \frac{n_C}{n_W}} \tag{2}$$

the number of words per sentence P_F:

$$\boldsymbol{P_F = \frac{n_W}{n_S}} \tag{3}$$

the number of interpunctions per word, referred to as the word interval, I_P:

$$\boldsymbol{I_P = \frac{n_I}{n_W}} \tag{4}$$

the number of word intervals per sentence M_F:

$$\boldsymbol{M_F = \frac{n_{I_P}}{n_S}} \tag{5}$$

Equation (5) can be written also as $M_F = P_F / I_P$.

Tables 6–9 reports the mean and standard deviation of these variables. Notice that these values have been calculated by weighing each chapter with its number of words to avoid the short chapters weighing as much as long ones. For example, chapter 1 of *Lord* has 10097 words; therefore, its statistical weight is $10097/472173 \approx 0.021$, not $1/62 \approx 0.016$. Notice, also, that the coefficient of dispersion used in Section 2 was calculated by weighing each chapter $1/62$, not $10097/472173$, to visually agree with the series drawn in Figure 1.

Table 6. John R.R. Tolkien. Mean value and standard deviation (in parentheses) of $< C_P >$, $< P_F >$, $< I_P >$, $< M_F >$ in the indicated novels. Mean and standard deviation have been calculated by weighing each chapter with its number of words.

Novel	C_p	P_F	I_P	M_F
The Hobbit	4.11 (0.06)	16.54 (2.03)	7.93 (0.98)	2.09 (0.12)
The Lord of the Rings	4.04 (0.08)	13.92 (1.98)	6.68 (0.51)	2.08 (0.20)
The Silmarillion	4.23 (0.08)	31.21 (5.32)	8.58 (0.58)	3.62 (0.42)

Table 7. Clive S. Lewis. Mean value and standard deviation (in parentheses) of $< C_P >$, $< P_F >$, $< I_P >$, $< M_F >$ in the indicated novels. Mean and standard deviation have been calculated by weighing each chapter with its number of words.

Novel	C_p	P_F	I_P	M_F
The Screwtape Letters	4.36 (0.12)	23.95 (3.82)	9.72 (1.00)	2.47 (032)
The Space Trilogy	4.21 (0.16)	15.25 (3.05)	7.47 (0.98)	2.03 (0.22)
The Chronicles of Narnia	4.09 (0.09)	13.97 (1.94)	7.10 (0.89)	1.97 (0.15)

Table 8. George MacDonald. Mean value and standard deviation (in parentheses) of $< C_P >$, $< P_F >$, $< I_P >$, $< M_F >$ in the indicated novels. Mean and standard deviation have been calculated by weighing each chapter with its number of words.

At the Back of the North Wind	3.85 (0.11)	15.52 (3.34)	6.76 (0.77)	2.29 (0.35)
Phantastes: A Fairie Romance for Men and Women	4.20 (0.12)	21.15 (3.58)	6.43 (0.49)	3.28 (0.45)
Lilith: A Romance	4.11 (0.25)	15.87 (3.89)	6.43 (0.57)	2.45 (0.42)
The Princess and the Goblin	4.08 (0.14)	17.81 (4.05)	7.09 (1.22)	2.46 (0.53)

Table 9. Other authors. Mean value and standard deviation (in parentheses) of $< C_P >$, $< P_F >$, $< I_P >$, $< M_F >$ in the indicated novels. Mean and standard deviation have been calculated by weighing each chapter with its number of words.

Novel	$<C_P>$	$<P_F>$	$<I_P>$	$<M_F>$
The Adventures of Oliver Twist	4.23 (0.09)	18.04 (3.13)	5.70 (0.52)	3.16 (0.34)
David Copperfield	4.04 (0.12)	18.83 (2.50)	5.61 (0.30)	3.35 (0.33)
Bleak House	4.23 (0.14)	16.95 (2.21)	6.59 (0.49)	2.57 (0.21)
A Tale of Two Cities	4.26 (0.12)	18.27 (4.24)	6.19 (0.46)	2.93 (0.46)
Our Mutual Friend	4.22 (0.12)	16.46 (2.01)	6.03 (0.37)	2.73 (0.27)
Pride and Prejudice	4.40 (0.14)	21.31 (5.02)	7.16 (0.46)	2.95 (0.46)
Vanity Fair	4.63 (0.08)	21.95 (3.67)	6.73 (0.63)	3.25 (0.39)
Moby Dick	4.52 (0.16)	23.82 (7.44)	6.45 (0.99)	3.64 (0.80)
The Mill On The Floss	4.29 (0.13)	23.84 (4.99)	7.09 (0.69)	3.35 (0.48)
Alice's Adventures in Wonderland	3.96 (0.08)	17.19 (3.20)	5.79 (0.55)	2.95 (0.28)
Adventures of Huckleberry Finn	3.85 (0.10)	19.39 (3.12)	6.63 (0.67)	2.94 (0.48)
The Jungle Book	4.11 (0.09)	16.46 (3.09)	7.14 (0.53)	2.29 (0.30)
The War of the Worlds	4.38 (0.18)	19.22 (4.13)	7.67 (0.92)	2.48 (0.31)
The Wonderful Wizard of Oz	4.02 (0.10)	17.90 (2.23)	7.63 (0.64)	2.34 (0.15)
The Hound of The Baskervilles	4.15 (0.12)	15.07 (3.16)	7.83 (0.94)	1.91 (0.22)
Peter Pan	4.12 (0.09)	15.65 (3.98)	6.35 (0.92)	2.44 (0.37)
Martin Eden	4.32 (0.13)	15.61 (2.71)	6.76 (0.64)	2.30 (0.26)

Specifically, let M be the number of samples (i.e., chapters), then the mean value $< P_F >$ is given by

$$< \boldsymbol{P_F} >= \sum_{k=1}^{\boldsymbol{M}} \boldsymbol{P_{F,k}} \times \left(\boldsymbol{n_{W,k}} / \sum_{\boldsymbol{k=1}}^{\boldsymbol{M}} \boldsymbol{n_{W,k}} \right) \quad (6)$$

Therefore, notice, for not being misled, that $< P_F >\neq \frac{1}{M} \sum_{k=1}^{M} P_{F,k} \neq \sum_{k=1}^{M} n_{W,k} / \sum_{k=1}^{M} n_{S,k}$=$W/S$. In other words, $< P_F >$ is not given by the total number of words W divided by the total number of sentences S, or by assigning the weight $1/M$ to every chapter. The three values coincide only if all the text blocks contain the same number of words and the same number of sentences, which did not occur. The same observations apply to all other variables.

The following characteristics can be observed from Tables 6–9. *Lord* and *Narnia* share the same $< P_F >$. *Silmarillion* is distinctly different from *Lord* and *Hobbit*, which is in agreement with the different coefficient of dispersion. *Screwtape* is distinctly different from *Narnia* and *Trilogy.* There is a great homogeneity in Dicken's novels and a large homogeneity in $< C_P >$ in all novels.

In the next sections, we use $< P_F >$, $< I_P >$ and $< M_F >$ to calculate interesting indices connected to the short-term memory of readers.

5. Extended Short-Term Memory of Writers/Readers and Universal Readability Index

In this section, we deal with the linguistic variables that, very likely, are not consciously managed by writers who, of course, act also as readers of their own text. We first report findings concerning the extended short-term memory and then those concerning a universal readability index. Both topics address human short-term memory buffers.

5.1. Extended Short-Term Memory and Multiplicity Factor

In [12,13], we have conjectured that the human short-term memory is sensitive to two independent variables, which apparently engage two short-term memory buffers in series, constituents of what we have called the extended short-term memory (E–STM). The first buffer is modeled according to the number of words between two consecutive interpunctions, i.e., the variable I_P, the word interval, which follows Miller's 7 ± 2 law [14]; the second buffer is modeled according to the number of word intervals, I_P's, contained in a sentence—i.e., the variable M_F—ranging approximately from 1 to 7.

In [13], we studied the patterns (which depend on the size of the two buffers) that determine the number of sentences that theoretically can be recorded in the E–STM of a given capacity. These patterns were then compared with the number of sentences actually found in novels of Italian and English literature. We have found that most authors write for readers with short memory buffers and, consequently, are forced to reuse sentence patterns to convey multiple meanings. This behavior is quantified by the multiplicity factor α, defined as the ratio between the number of sentences in a novel and the number of sentences theoretically allowed by the two buffers, a function of I_P and M_F.

We found that $\alpha > 1$ is more likely than $\alpha < 1$ and often $\alpha \gg 1$. In the latter case, writers reuse many times the same pattern of number of words. Few novels show $\alpha < 1$; in this case, writers do not use some or most of them. The values of α found in the novels presently studied are reported in Tables 10 and 11.

Table 10. Multiplicity factor α, universal readability index $< G_U >$ and number of school years Y in the indicated novels by Tolkien, Lewis, MacDonald.

Novel	α	$<G_U>$	Y
The Hobbit	39.4	52.4	9.9
The Lord of the Rings	368.1	64.2	7.4
The Silmarillion	0.2	38.7	>13

Table 10. *Cont.*

Novel	α	$<G_U>$	Y
The Screwtape Letters	1.4	33.5	>13
The Space Trilogy	186.3	56.2	9.0
The Chronicles of Narnia	297.7	61.1	7.9
At the Back of the North Wind	26.3	63.9	7.5
Phantastes: A Fairie Romance for Men and Women	24.3	56.6	8.9
Lilith: A Romance	1.4	63.0	7.5
The Princess and the Goblin	8.5	58.2	8.4

Table 11. Multiplicity factor α, universal readability index $< G_U >$ and number of school years in the indicated novels of English Literature.

Novel	α	$<G_U>$	Y
The Adventures of Oliver Twist	9.46	63.19	7.5
David Copperfield	12.63	64.78	7.2
Bleak House	56.98	58.74	8.3
A Tale of Two Cities	11.89	59.91	8.0
Our Mutual Friend	43.41	62.68	7.6
Pride and Prejudice	5.20	50.33	10.4
Vanity Fair	5.26	49.74	10.5
Moby Dick	1.56	52.63	9.9
The Mill On The Floss	2.17	50.22	10.5
Alice's Adventures in Wonderland	2.90	59.27	8.1
Adventures of Huckleberry Finn	7.05	60.42	8.0
The Jungle Book	14.10	57.59	8.6
The War of the Worlds	6.72	49.05	10.8
The Wonderful Wizard of Oz	7.02	53.83	9.5
The Hound of The Baskervilles	43.87	54.87	9.2
Peter Pan	13.07	63.60	7.5
Martin Eden	46.33	58.53	8.2

5.2. Universal Readability Index

In Reference [14], we have proposed a universal readability index given by

$$G_U = 89\text{–}10kC_P + 300/P_F\text{–}6(I_P\text{–}6) \quad (7)$$

$$k = < C_{P,ITA} > / < C_{P,Eng} > \quad (8)$$

In Equation (8), $< C_{p,ITA} >= 4.48$, $< C_{p,ENG} >= 4.24$. By using Equations (7) and (8), the average value $< kC_P >$ of any language is forced to be equal to that found in Italian, namely 4.48. The rationale for this choice is that C_P is a parameter typical of a language which, if not scaled, would bias G_U without really quantifying the reading difficulty for readers, who in their language are used, on average, to reading shorter or longer words than in Italian. This scaling, therefore, avoids changing G_U for the only reason that a language has, on average, words shorter (as English) or longer than Italian. In any case, C_p affects Equation (7) much less than P_F or I_P.

The values of $< G_U >$—calculated as the other linguistic variables, i.e., by weighing chapters (samples) according to the number of words – are reported in Tables 10 and 11.

The reader may be tempted to calculate Equation (7) by introducing the mean values reported in Tables 6–9. This, of course, can be performed but it should be noted that the values so obtained are always less or equal (hence they are lower bounds) to the means calculated from the samples (see Appendix A). For example, for *Lord*, instead of 64.9, we would obtain 61.9.

It is interesting to "decode" these mean values into the minimum number of school years, Y necessary to make a novel "easy" to read, according to the Italian school system, which is assumed as the reference, see Figure 1 of [15]. The results are also listed in Tables 10 and 11.

5.3. Discussion

Several intriguing observations can be drawn from the results presented in the preceding subsections.

(a). *Silmarillion* with $\alpha = 0.2$ is quite diverse from other Tolkien's writings. Mathematically, this is due to its large $< M_F >= 3.62$ and $< I_P >= 8.58$. In practice, the number of theoretical sentences allowed by the E–STM to read this text is only $1/\alpha = 5$ times the number of sentence patterns actully used in the text. The reader needs a powerful E–STM and reading ability, since $G_U = 38.7$ and $Y > 13$. This does not occur for *Hobbit* ($\alpha = 39.4$, $G_U = 52.4$, Y =9.9) and *Lord* ($\alpha = 368.1, G_U = 64.2$, $Y = 7.4$) in which Tolkien reuses patterns many times, especially in *Lord*.

(b). *Lord* and *Narnia* show very large values, $\alpha = 368.1$ and $\alpha = 297.7$, and very similar G_U's and school years: $G_U = 64.2$, $Y = 7.4$ and $G_U = 61.1$, $Y = 7.9$, respectively. Sentence patterns are reused many times by Lewis in this novel, but not in *Screwtape* ($\alpha = 1.4$), which is more difficult to read ($G_U = 33.5$) and requires more years of schooling, $Y > 13$. Moreover, *Lord* and *Narnia* have practically the same $< P_F > \approx 14$.

(c). In general, *Narnia* is closer to *Lord* than to *Trilogy*, although the number of words and sentences in *Trilogy* and *Narnia* are quite similar (Table 1). This difference between *Trilogy* ($G_U = 56.2$, $Y = 9$) and *Narnia* ($G_U = 61.1$, $Y = 7.9$) might depend on the different readers addressed, kids for *Narnia* and adults for *Trilogy*, with different reading ability, as G_U indicates.

(d). The novels by MacDonald show values of α and G_U very similar to those of the other English novels.

(e). Notice the homogeneity in Dicken's novels, which require about $Y = 7 \sim 8$ years of school and readability index $< G_U >= 59 \sim 65$.

In conclusion, *Lord* and *Narnia* are the novels that address readers with very similar E–STM buffers, reuse sentence patterns in similar ways, contain the same number of words per sentence, and require the same reading ability and school years compared to other novels by Tolkien and Lewis. The mathematical connections between *Lord* and *Narnia* will be further pursued in the next section, where the four deep language parameters are used to represent texts geometrically.

6. Geometrical Representation of Texts

The mean values of Tables 6–9 can be used to assess how texts are "close", or mathematically similar, in the Cartesian coordinate plane, by defining linear combinations of deep–language variables. Texts are then modeled as vectors; the representation is discussed in detail in [9,10] and briefly recalled here. An extension of this geometrical representation of texts allows the calculation of the probability that a text may be confused with another one, an extension in two dimensions of the problem discussed in [16]. The values of the conditional probability between two texts (authors) can be considered an index indicating who influenced who.

6.1. Vector Representation of Texts

Let us consider the following six vectors of the indicated components of deep language variables: $\vec{R_1} = (< C_P >, < P_F >)$, $\vec{R_2} = (< M_F >, < P_F >)$, $\vec{R_3} = (< I_P >, < P_F >)$, $\vec{R_4} = (< C_P >, < M_F >)$, $\vec{R_5} = (< I_P >, < M_F >)$, $\vec{R_6} = (< I_P >, < C_P >)$ and their resulting vector sum:

$$\vec{R} = \sum_{k=1}^{6} \vec{R_k} = x\,\vec{i} + y\,\vec{j} \tag{9}$$

The choice of which parameter represents the component in the abscissa and ordinate axes is not important because, once the choice is made, the numerical results will depend on it, but not the relative comparisons and general conclusions.

In the first quadrant of the Cartesian coordinate plane, two texts are likely mathematically connected—they show close ending points of vector (9)—if their relative Pythagorean distance is small. A small distance means that texts share a similar mathematical structure, according to the four deep language variables.

By considering the vector components x and y of Equation (9), we obtain the scatterplot shown in Figure 2 where X and Y are normalized coordinates calculated by setting *Lord* at the origin $(X = 0,\ Y = 0)$ and *Silmarillion* at $(X = 1,\ Y = 1)$, according to the linear tranformations:

$$X = \frac{x - x_{Lord}}{x_{Silma} - x_{Lord}} \tag{10}$$

$$X = \frac{y - y_{Lord}}{y_{Silma} - y_{Lord}} \tag{11}$$

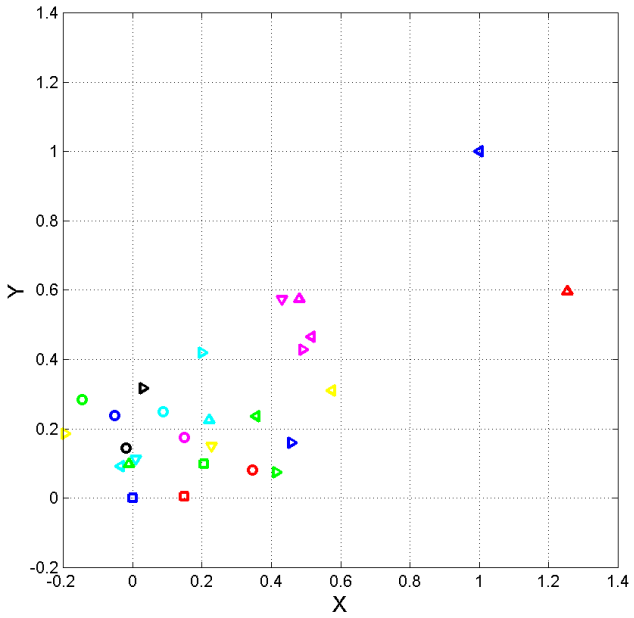


Figure 2. Normalized coordinates X and Y of the ending point of vector (5) such that *Lord*, blue square, is at (0,0) and *Silmarillion*, blue triangle pointing left, is (1,1). *Narnia*: red square; *Trilogy*: red circle; *Hobbit*: blue triangle pointing right; *Screwtape*: red triangle pointing upward; *Back*: cyan triangle pointing left; *Lilith*: cyan triangle pointing downward; *Back*: cyan triangle pointing left; *Phantastes*: cyan triangle pointing right; *Princess*: cyan triangle pointing upward; *Oliver*: blue circle; *David*: green circle; *Tale*: cyan circle; *Bleak*: magenta circle; *Mutual*: black circle; *Pride*: magenta triangle pointing right; *Vanity*: magenta triangle pointing left; *Moby*: magenta triangle pointing downward; *Mill*: magenta triangle pointing upward; *Alice*: yellow triangle pointing right; *Jungle*: yellow triangle pointing downward; *War*: yellow triangle pointing right; *Oz*: green triangle pointing left; *Bask*: green triangle pointing right; *Peter*: green triangle pointing upward; *Martin*: green square; *Finn*: black triangle pointing right.

From Figure 2, we can notice that *Silmarillion* and *Screwtape* are distinctly very far from all other texts examined, marking their striking diversity, as already remarked; therefore, in the following analyses, we neglect them. Moreover, *Pride*, *Vanity*, *Moby* and *Floss* are grouped together and far from *Trilogy*, *Narnia* and *Lord*; therefore, in the following analyses, we will not consider them further.

The complete set of the Pythagorean distance d between pairs of texts is reported in Appendix B. These data synthetically describe proximity of texts and may indicate to scholars of literature connections between texts not considered before.

Figure 3 shows example of these distances concerning *Lord*, *Narnia* and *Trilogy*. By referring to the cases in which $d < 0.2$, we can observe the following:

(a). The closest texts to *Lord* are *Narnia*, *Back*, *Lilith*, *Mutual* and *Peter*.
(b). The closest texts to *Narnia* are *Lord*, *Lilith*, *Bleak*, *Martin* and *Peter*.
(c). The closest texts to *Trilogy* are *Hobbit*, *Martin* and *Peter*.

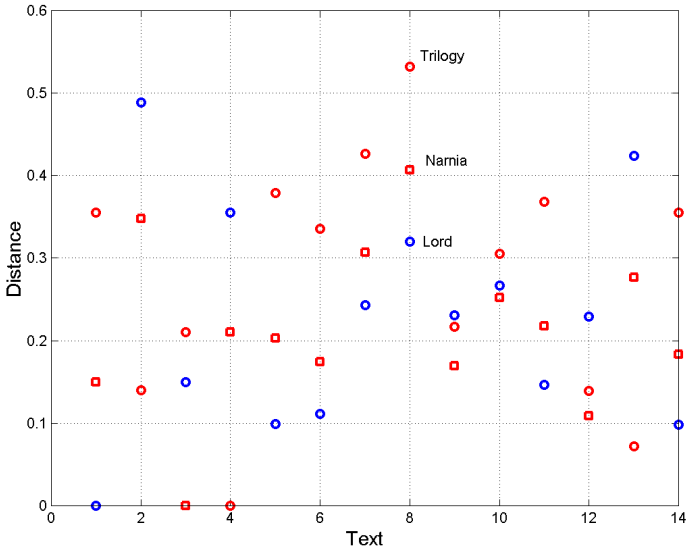


Figure 3. Pythagorean distance d between pairs of texts considering *Lord* (the distances referring to this case are labeled with blue circles), *Narnia* (red squares) and Trilogy (red circles). Key: *Lord* 1, *Hobbit* 2, *Narnia* 3, *Trilogy* 4, *Back* 5, *Lilith* 6, *Oliver* 7, *David* 8, *Bleak* 9, *Tale* 10, *Mutual* 11, *Martin* 12, *Bask* 13, *Peter* 14.

Besides the proximity with earlier novels, *Lord* and *Narnia* show close proximity with each other and with two novels by MacDonald.

These remarks, however, refer to the "average" display of vectors whose ending point depends only on mean values. The standard deviation of the four deep language variables, reported in Tables 6–9, do introduce data scattering; therefore, in the next subsection, we study and discuss this issue by calculating the probability (called "error" probability) that a text may be mathematically confused with another one.

6.2. Error Probability: An Index to Assess Who Influenced Who

Besides the vector $\vec{R}$ of Equation (9)—due to mean values—we can consider another vector $\vec{\rho}$, due to the standard deviation of the four deep language variables that adds to $\vec{R}$. In this case, the final random vector describing a text is given by

$$\vec{T} = \vec{R} + \vec{\rho} \tag{12}$$

Now, to obtain some insight into this new description, we consider the area of a circle centered at the ending point of $\vec{R}$.

We fix the magnitude (radius) ρ as follows. First, we add the variances of the deep language variables that determine the components x and y of $\vec{R}$, let them be σ_x^2, σ_y^2. Then, we calculate the average value $\sigma_\rho^2 = 0.5 \times \left(\sigma_x^2 + \sigma_y^2\right)$ and finally, we set

$$\rho = \sigma_\rho \tag{13}$$

Now, since in calculating the coordinates x and y of $\vec{R}$ a deep language variable can be summed twice or more, we add its standard deviation (referred to as sigma) twice or more times before squaring. For example, in the x–component, I_P appears three times; therefore, its contribution to the total variance in the x–axis is 9 times the variance calculated from the standard deviation reported in Tables 6–9. For *Lord*, for example, it is 9×0.51^2. After these calculations, the values of the 1–sigma circle are transformed into the normalized coordinates X, Y according to Equations (10) and (11).

Figure 4 shows a significant example involving *Lord*, *Narnia*, *Trilogy*, *Back* and *Peter*. We see that *Lord* can be almost fully confused with *Narnia*, and partially with *Trilogy*, but not vice versa. *Lord* can also be confused with *Peter* and *Back, therefore* indicating strong connections with these earlier novels.

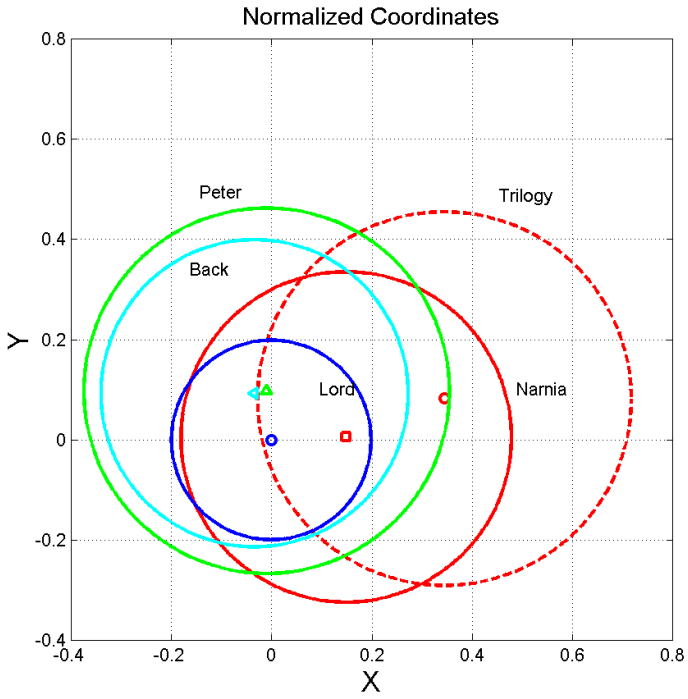


Figure 4. Normalized coordinates X and Y of the ending point of vector (5) and 1–sigma circles, such that *Lord*, blue square, is at (0,0) and *Silmarillion*, blue triangle pointing left, is (1,1). *Lord*: blue square (blue 1–sigma circle); *Narnia*: red square (red 1–sigma circle); *Trilogy*: red circle (dashed red 1–sigma circle); *Back*: cyan triangle pointing left (cyan 1–sigma circle); *Peter*: green triangle pointing upward (green 1–sigma circle).

Now, we can estimate the (conditional) probability that a text is confused with another by calculating the ratio of areas. This procedure is correct if we assume that the bivariate density of the normalized coordinates ρ_X, ρ_Y, centered at $\vec{R}$, is uniform. By assuming this hypothesis, we can calculate probabilities as the ratio of areas [17,18].

The hypothesis of substantial uniformity around $\vec{R}$ should be justified by noting that the coordinates X, Y are likely distributed according to a log-normal bivariate density because the logarithm of the four deep language variables, which combine in Equation (9) linearly, can be modeled as a Gaussian. For the central limit theorem, we should expect approximately a Gaussian model on the linear values, but with a significantly larger

standard deviation that that of the single variables. Therefore, in the area close to $\vec{R}$, the bivariate density function should not be peaked, hence the uniform density modeling.

Now, we can calculate the following probabilities. Let A be the common area of two 1–sigma circles (i.e., the area proportional to the joint probability of two texts), let A_1 be the area of 1–sigma circle of text 1 and A_2 the area of 1–sigma circle of text 2. Now, since probabilities are proportional to areas, we obtain the following relationships:

$$\frac{A}{A_1} = \frac{P(A_1, A_2)}{P(A_1)} = \frac{P(A_2/A_1)P(A_1)}{P(A_1)} = P(A_2/A_1) \tag{14}$$

$$\frac{A}{A_2} = \frac{P(A_1, A_2)}{P(A_2)} = \frac{P(A_1/A_2)P(A_2)}{P(A_2)} = P(A_1/A_2) \tag{15}$$

In other words, A/A_1 gives the conditional probability $P(A_2/A_1)$ that part of text 2 can be confused (or "contained") with text 1; A/A_2 gives the conditional probability $P(A_1/A_2)$ that part of text 1 can be confused with text 2. Notice that these conditional probabilities depend on the distance between two texts and on the 1–sigma radii (Appendix C).

Of course, these joint probabilities can be extended to three or more texts, e.g., in Figure 4 we could calculate the area shared by *Lord*, *Narnia* and *Trilogy* and the corresponding joint probability, which is not conducted in the present paper.

We think that the conditional probabilities and the visual display of 1–sigma circles give useful clues to establish possible hidden connections between texts and, maybe, even between authors, because the variables involved are not consciously managed by them.

In Table 12, the conditional probability $P(A_2/A_1)$ is reported in the columns; therefore, A_1 refers to the text indicated in the upper row. $P(A_1/A_2)$ is reported in the rows; therefore, A_2 refers to the text indicated in the left column.

Table 12. Conditional probability between the indicated novels. $P(A_2/A_1)$ is reported in the columns; therefore, A_1 refers to the text indicated in the upper row. $P(A_1/A_2)$ is reported in the rows; therefore, A_2 refers to the text indicated the left column. For example, assuming *Lord* as text 1 (column 1 of Table 12) and *Narnia* as text 2 (row 3), we find $P(A_2/A_1) = 0.974$ and vice versa. If we assume *Narnia* as text 1 (column 3) and *Lord* as text 2 (row 1), we find $P(A_2/A_1) = 0.356$.

	Novel	***Lord***	***Hobbit***	***Narnia***	***Trilogy***	***Back***	***Lilith***	***Oliver***	***David***	***Bleak***	***Tale***	***Mutual***	***Martin***	***Bask***	***Peter***
1	*Lord*	1	0.031	0.356	0.142	0.423	0.511	0.277	0.041	0.307	0.231	0.619	0.301	0.078	0.299
2	*Hobbit*	0.099	1	0.421	0.731	0.171	0.225	0.074	0	0.592	0.376	0.060	0.665	0.833	0.227
3	*Narnia*	0.974	0.354	1	0.550	0.647	0.781	0.489	0.166	0.927	0.625	0.886	0.949	0.462	0.602
4	*Trilogy*	0.498	0.786	0.704	1	0.400	0.510	0.297	0	0.921	0.608	0.473	0.978	0.908	0.421
5	*Back*	1	0.124	0.559	0.270	1	0.997	0.866	0.796	0.763	0.692	1	0.566	0.179	0.707
6	*Lililth*	0.891	0.121	0.498	0.254	0.735	1	0.741	0.559	0.762	0.661	1	0.546	0.169	0.521
7	*Oliver*	0.352	0.029	0.227	0.108	0.466	0.540	1	0.913	0.444	0.579	0.917	0.239	0.043	0.378
8	*David*	0.024	0	0.035	0	0.195	0.186	0.416	1	0.029	0.173	0.262	0.001	0	0.168
9	*Bleak*	0.307	0.182	0.339	0.264	0.323	0.437	0.350	0.051	1	0.598	0.526	0.558	0.208	0.296
10	*Tale*	0.330	0.165	0.327	0.248	0.417	0.541	0.650	0.427	0.852	1	0.774	0.484	0.176	0.385
11	*Mutual*	0.390	0.012	0.204	0.085	0.266	0.361	0.454	0.285	0.331	0.342	1	0.205	0.031	0.188
12	*Martin*	0.490	0.333	0.565	0.455	0.389	0.509	0.307	0.004	0.906	0.551	0.529	1	0.396	0.384
13	*Bask*	0.250	0.826	0.545	0.838	0.244	0.312	0.110	0	0.669	0.397	0.160	0.785	1	0.289
14	*Peter*	1	0.234	0.736	0.403	1	1	0.996	0.968	0.988	0.904	1	0.790	0.300	1

Notice that $P(A_2/A_1) = 1$ means $A = A_1$; therefore, text 1 can be fully confused with text 2. $P(A_1/A_2) = 1$ means $A = A_2$; therefore, text 2 can be fully confused with text 1.

For example, assuming *Lord* as text 1 (column 1 of Table 12) and *Narnia* as text 2 (row 3), we find $P(A_2/A_1) = 0.974$ and vice versa. If we assume *Narnia* as text 1 (column 3) and *Lord* as text 2 (row 1), we find $P(A_2/A_1) = 0.356$. These data indicate that *Lord* can be confused with *Narnia* with a probability close to 1, but not vice versa. In other words, in the data bank considered in this paper, if a machine randomly extracts a chapter from *Lord*, another machine, unaware of this choice, could attribute it to *Lord*, but also with decreasing probability to *Back*, *Peter, Narnia* and *Lilith.*

On the contrary, if the text is extracted from *Narnia*, then it is more likely attributed to *Peter* or *Trilogy* than to *Lord* or other texts.

We think that these conditional probabilities indicate who influenced who more. In other words, Tolkien influenced more Lewis that the opposite.

Now, we can define a synthetic parameter which highlights how much, on the average, two texts can be erroneously confused with each other. The parameter is the average conditional probability (see [16] for a similar problem):

$$p_e = P(A_2/A_1)P(A_1) + P(A_1/A_2)P(A_2) \tag{16}$$

Now, since in comparing two texts we can assume $P(A_1) = P(A_2) = 0.5$, we receive

$$p_e = 0.5 \times [P(A_2/A_1) + P(A_1/A_2)] \tag{17}$$

If $p_e = 0$, there is no intersection between the two 1–sigma circles. The two texts cannot be each other confused; therefore, there is no mathematical connection involving the deep language parameters (this happens for *Screwtape* and *Silmarillion*, which can be each other confused, but not with the other texts). If $p_e = 1$, the two texts can be totally confused, and the two 1–sigma circles coincide. Appendix D reports the values of p_e for all the pairs of novels.

Now, just to allow some rough analysis, it is reasonable to assume $p_e = 0.5$ as a reference threshold, i.e., the probability of obtaining heads or tails in flipping a fair coin. If $p_e > 0.5$, then two texts can be confused not by chance; if $p_e \leq 0.5$, then two texts cannot likely be confused.

To visualize p_e, Figure 5 draws p_e when text 1 is *Lord* (column 1 of Table 12), *Narnia* (column 3) or *Trilogy* (column 4). We notice that $p_e > 0.5$ in the following cases:

(a). *Lord* as text 1: *Narnia*, *Back*, *Lilith*, *Mutual*, *Peter*.
(b). *Narnia* as text 1: *Lord*, *Trilogy*, *Back*, *Lilith*, *Bleak*, *Mutual*, *Martin*, *Peter*.
(c). *Trilogy* as text 1: *Hobbit*, *Narnia*, *Bleak*, *Martin*, *Bask*.

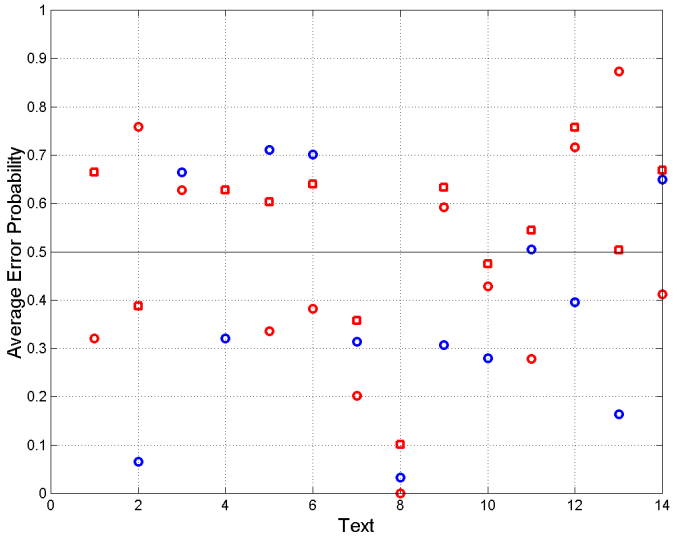


Figure 5. Error probability p_e versus text 2. *Lord* (the probabilities referring to this case are labeled with blue circles), *Narnia* (red squares) and *Trilogy* (red circles). Text key: *Lord* 1, *Hobbit* 2, *Narnia* 3, *Trilogy* 4, *Back* 5, *Lilith* 6, *Oliver* 7, *David* 8, *Bleak* 9, *Tale* 10, *Mutual* 11, *Martin* 12, *Bask* 13, *Peter* 14.

We can reiterate that Tolkien (*Lord*) appears significantly connected to Lewis (*Narnia*), to MacDonald (*Back*, *Lilith*) and Barrie (*Peter*), but not to Dicken's novels where, on the contrary, Lewis appears connected.

In the next section, the four deep language variables are singled out to consider linguistic channels existing in texts. This is the analysis we have called the "fine tuning" of texts [11].

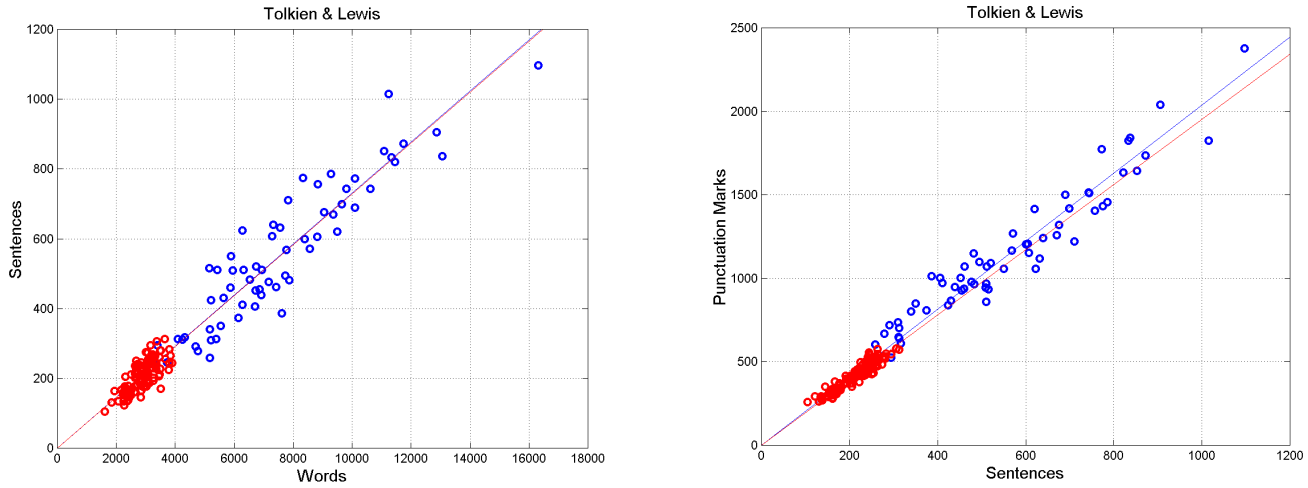
7. Linear Relationships in Literary Texts

The theory of linguistic channels, which we will be revisited in the next section, is based on the regression line between linguistic variables:

$$y = mx \tag{18}$$

Therefore, we show examples of these linear relationships found in *Lord* and *Narnia*.

Figure 6a shows the scatterplot of n_S versus n_W of *Lord* and *Narnia*. In *Narnia*, the slope of the regression line is $m = 0.0729$ and the correlation coefficient $r = 0.7610$. In *Lord*, $m = 0.0731$ and $r = 0.9199$. Since the average relationships—i.e., Equation (18)—are practically identical—see also the values of $< P_F >$ in Tables 6 and 7—while the correlation coefficients—i.e., the scattering of the data—are not, this fact will impact the sentence channel discussed in Section 9.



(**a**) (**b**)

Figure 6. (**a**) Scatterplot of n_S versus n_W in *Lord* (blue) and *Narnia* (red); (**b**) n_I versus n_S in *Lord* (blue) and *Narnia* (red).

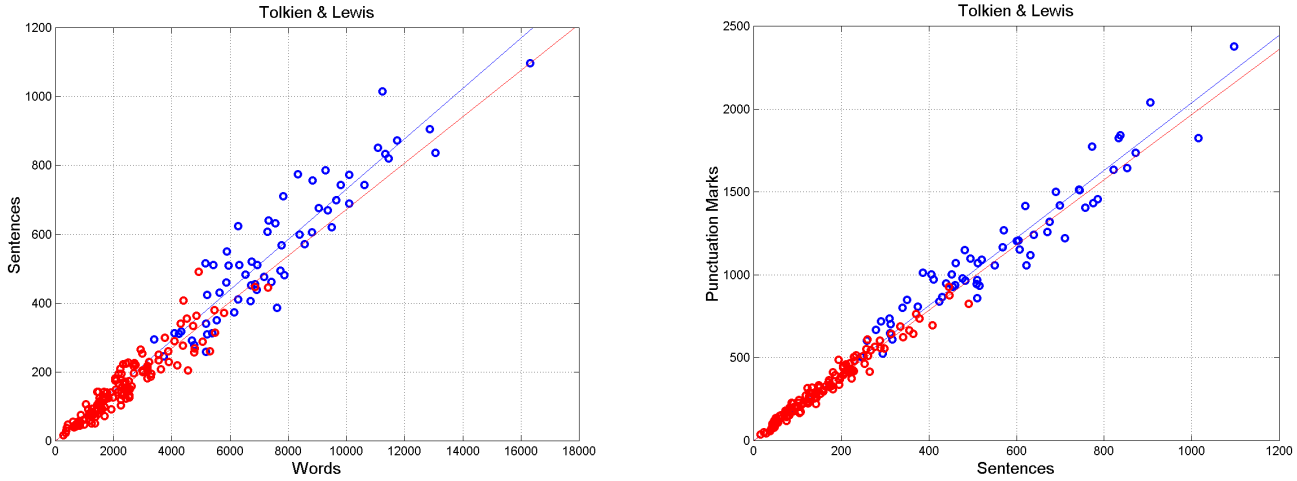
Similar observations can be carried out for Figure 6b, which shows n_I versus n_S in *Lord* and *Narnia*. We find $m = 2.0372$, $r = 0.9609$ in *Lord*, and $m = 1.9520$ and $r = 0.9384$ in *Narnia*. Appendix E reports the complete set of these parameters.

Figure 7 shows the scatterplots of *Lord* and *Trilogy*. In *Trilogy*, for n_S versus n_W m = 0.0672, $r = 0.9325$; for n_I versus n_S $m = 1.9664, r = 0.9830$.

Figure 8 shows the scatterplots for *Lord* and *Back* or *Lilith*. We see similar regression lines and data scattering. In *Back* (left panel), the regression line between n_S and n_W gives $m = 0.0681$, $r = 0.9416$; in *Lilith* (right panel), $m = 0.0676$, $r = 0.8890$. These results likely indicate the influence of MacDonald on Tolkien's writings because they are different from most other novels.

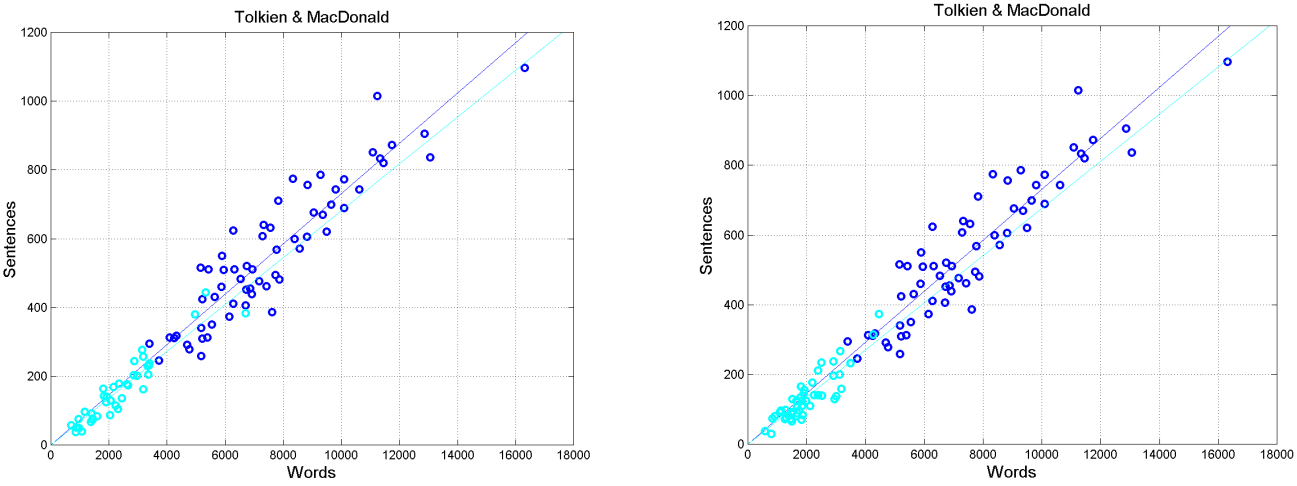
In conclusion, the regression lines of *Lord*, *Narnia* and *Trilogy* are very similar, but they can differ in the scattering of the data. Regression lines, however, describe only one aspect of the relationship, namely the relationship between conditional average values in Equation (18); they do not consider the other aspect of the relationship, namely the scattering of data, which may not be the same even when two regression lines almost coincide, as shown above. The theory of linguistic channels, discussed in the next section, on the contrary,

considers both slopes and correlation coefficients and provides a "fine tuning" tool to compare two sets of data by singling out each of the four deep language parameters.



(**a**) (**b**)

Figure 7. (**a**) Scatterplot of n_S versus n_W in *Lord* (blue) and *Trilogy* (red); (**b**) n_I versus n_S in in *Lord* (blue) and *Trilogy* (red).



(**a**) (**b**)

Figure 8. Scatterplot of the number of sentences n_S versus the number of words n_W: (**a**) *Lord* (blue) and *Back* (cyan); (**b**) *Lord* (blue) and *Lilith* (cyan).

8. Theory of Linguistic Channels

In this section, we recall the general theory of linguistic channels [11]. In a literary work, an independent (reference) variable x (e.g., n_W) and a dependent variable y (e.g., n_S) can be related by the regression line given by Equation (18).

Let us consider two different text blocks Y_k and Y_j, e.g., the chapters of work k and work j. Equation (18) does not give the full relationship between two variables because it links only conditional average values. We can write more general linear relationships, which take care of the scattering of the data—measured by the correlation coefficients r_k and r_j, respectively—around the average values (measured by the slopes m_k and m_j):

$$y_k = m_k x + n_k \tag{19}$$

$$y_j = m_j x + n_j \tag{20}$$

The linear models Equations (19) and (20) introduce additive "noise" through the stochastic variables n_k and n_j , with zero mean value [9,11,15]. The noise is due to the correlation coefficient $|r| \neq 1$.

We can compare two literary works by eliminating x; therefore, we compare the output variable y for the same number of the input variable x. For example, we can compare the number of sentences in two novels—for an equal number of words—by considering not only the average relationship, Equation (18), but also the scattering of the data, measured by the correlation coefficient, Equations (19) and (20). We refer to this communication channel as the "sentences channel", S–channel, and to this processing as "fine tuning" because it deepens the analysis of the data and can provide more insight into the relationship between two literary works or any other texts.

By eliminating x from Equations (19) and (20), we obtain the linear relationship between the input number of sentences in work Y_k (now the reference, input text) and the number of sentences in text Y_j (now the output text):

$$y_j = \frac{m_j}{m_k} y_k - \frac{m_j}{m_k} n_k + n_j \tag{21}$$

Compared to the new reference work Y_k, the slope m_{jk} is given by

$$\boldsymbol{m_{jk} = m_j / m_k} \tag{22}$$

The noise source that produces the correlation coefficient between Y_k and Y_j is given by

$$\boldsymbol{n_{jk} = -\frac{m_j}{m_k} n_k + n_j = -m_{jk} n_k + n_j} \tag{23}$$

The "regression noise–to–signal ratio", R_m, due to $\left|m_{jk}\right| \neq 1$, of the new channel is given by

$$\boldsymbol{R_m = \left(m_{jk} - 1\right)^2} \tag{24}$$

The unknown correlation coefficient r_{jk} between y_j and y_k is given by

$$\boldsymbol{r_{jk} = \cos\left|\arccos\left(r_j\right) - \arccos\left(r_k\right)\right|} \tag{25}$$

The "correlation noise–to–signal ratio", R_r, due to $\left|r_{jk}\right| < 1$, of the new channel from text Y_k to text Y_j is given by

$$\boldsymbol{R_r = \frac{1 - r_{jk}^2}{r_{jk}^2} m_{jk}^2} \tag{26}$$

Because the two noise sources are disjoint and additive, the total noise-to-signal ratio of the channel connecting text Y_k to text Y_j is given by

$$\boldsymbol{R = R_m + R_r} \tag{27}$$

Notice that Equation (27) can be represented graphically [10]. Finally, the total and the partial signal-to-noise ratios are given by

$$\Gamma_{dB} = -10 \times \log_{10} \mathrm{R} \tag{28}$$

$$\Gamma_{m,dB} = -10 \times \log_{10} R_m \tag{29}$$

$$\Gamma_{r,dB} = -10 \times \log_{10} R_r \tag{30}$$

Of course, we expect that no channel can yield $\left|r_{jk}\right| = 1$ and $\left|m_{jk}\right| = 1$; therefore, $\Gamma_{dB} = \infty$, a case referred to as the ideal channel, unless a text is compared with itself. In practice, we always find $\left|r_{jk}\right| < 1$ and $\left|m_{jk}\right| \neq 1$. The slope m_{jk} measures the multiplicative "bias" of the dependent variable compared to the independent variable; the correlation coefficient r_{jk} measures how "precise" the linear best fit is.

In conclusion, the slope m_{jk} is the source of the regression noise R_m, and the correlation coefficient r_{jk} is mostly the source of the correlation noise of the channel R_r.

9. Linguistic Channels

In long texts (such as novels, essays, etc.), we can define at least four linguistic linear channels [11], namely:

(a). Sentence channel (S–channel)
(b). Interpunctions channel (I–channel)
(c). Word interval channel, WI–channel
(d). Characters channel (C–channel).

In S–channels, the number of sentences of two texts is compared to the *same* number of words. These channels describe how many sentences the author of text j writes, compared to the writer of text k (reference text), by using the same number of words. Therefore, these channels are more linked to P_F than to other parameters. It is very likely they reflect the style of the writer.

In I–channels, the number of word intervals of two texts is compared for the *same* number of sentences. These channels describe how many short texts between two contiguous punctuation marks (of length I_P) two authors use; therefore, these channels are more linked to M_F than to other parameters. Since M_F is very likely connected with the E–STM, I–channels are more related to the second buffer of readers' E–STM than to the style of the writer.

In WI–channels, the number of words contained in a word interval (i.e., I_P) is compared for the *same* number of interpunctions. These channels are more linked to I_P than to other parameters. Since I_P is very likely connected with the E–STM, WI–channels are more related to the first buffer of readers' E–STM than to the style of the writer.

In C–channels, the number of characters of two texts is compared to the same number of words. They are more related to the language used, e.g., English, than to the other parameters, unless essays or scientific/academic texts are considered because these latter texts use, on average, longer words [9].

As an example, Table 13 reports the total and the partial signal-to-noise ratios Γ_{dB}, $\Gamma_{m,dB}$, $\Gamma_{r,dB}$ in the four channels by considering *Lord* as reference (input) text. In other words, text j is compared to text k (reference text, i.e., *Lord*).

Table 13. Total and the partial signal-to-noise ratios Γ_{dB}, $\Gamma_{m,dB}$, $\Gamma_{r,dB}$ in the four channels by considering *Lord* as reference (input) text.

	S–Channel			**I–Channel**			**WI–Channel**			**C–Channel**		
Novel	Γ_{dB}	$\Gamma_{m,dB}$	$\Gamma_{r,dB}$	Γ_{dB}	$\Gamma_{m,dB}$	$\Gamma_{r,dB}$	Γ_{dB}	$\Gamma_{m,dB}$	$\Gamma_{r,dB}$	Γ_{dB}	$\Gamma_{m,dB}$	$\Gamma_{r,dB}$
Hobbit	14.60	15.48	21.94	15.04	30.08	15.18	15.61	15.77	30.19	28.87	35.90	29.83
Narnia	10.12	51.26	10.12	21.83	27.57	23.18	7.94	27.08	7.99	23.14	37.47	23.30
Trilogy	21.27	21.86	30.24	20.07	29.18	20.64	18.74	19.21	28.63	27.61	27.85	40.30
Back	21.10	23.30	25.11	20.96	24.12	23.82	21.96	57.42	21.96	26.52	26.68	41.05
Lililth	19.92	22.47	23.44	18.47	19.15	26.87	20.86	26.28	22.33	19.17	33.98	19.31
Oliver	12.87	12.93	31.51	5.72	5.74	29.30	16.54	16.56	40.83	26.50	26.63	41.73
David	11.43	11.52	28.37	4.18	4.19	30.77	14.43	16.10	19.37	32.88	51.53	32.94
Bleak	14.91	14.93	38.01	12.10	12.29	25.80	29.00	36.33	29.88	21.76	26.69	23.45
Tale	12.66	13.31	21.25	8.26	8.56	19.99	16.13	22.76	17.20	25.88	25.89	55.01
Mutual	16.13	16.22	33.05	9.97	10.00	31.20	19.40	20.33	26.55	23.24	26.87	25.70
Martin	16.00	20.01	18.20	18.04	20.74	21.38	29.30	40.14	29.68	21.11	23.16	25.34
Bask	10.92	23.84	11.14	18.77	20.81	23.03	11.77	15.70	14.02	26.34	32.54	27.53
Peter	18.10	24.41	19.25	17.31	17.53	30.45	16.67	22.23	18.09	31.19	34.62	33.81

Appendix F reports Γ_{dB} for all novels considered in the paper.

Let us make some fundamental remarks on Table 13, applicable to whichever is the reference text. The signal-to-noise ratios of C–channels are practically the largest ones, ranging from 19.17 dB (*Lilith*) to 31.19 dB (*Back*). These results are simply saying that all authors use the same language and write texts of the same kind, which is novels, not essays

or scientific/academic papers. These channels are not apt to distinguish or assess large differences between texts or authors.

In the three other channels, we can notice that *Trilogy*, *Back* and *Lilith* have the largest signal-to-noise ratios, about $\sim$ 19 to $\sim$ 22 dB; therefore, these novels are very similar to *Lord*. In other words, these channels seem to confirm the likely influence by MacDonald on both *Lord* and *Trilogy* and the connection between *Lord* and *Trilogy*.

On the contrary, *Narnia* shows poor values in the S–Channel (10.12 dB) and WI–Channel (7.94 dB). These low values are determined by the correlation noise because $R = R_m + R_r \approx R_r$. If we consider only $\Gamma_{m,dB}$—i.e., only the regression line—then we notice a strong connection with *Lord* since $\Gamma_{m,dB} = 51.26$ dB. As we have already observed regarding Figure 6, the regression lines are practically identical but the spreading of the data is not. Lewis in *Narnia* is less "regular" than in *Trilogy* or Tolkien in *Lord* in shaping (unconsciously) these two linguistic channels.

10. Summary and Conclusions

Scholars of English Literature unanimously say that J.R.R. Tolkien influenced C.S. Lewis's writings. For the first time, we have investigated this issue mathematically by using an original multi-dimensional analysis of linguistic parameters, based on the surface deep language variables and linguistic channels.

To set our investigation in the framework of English Literature, we have also considered some novels written by earlier authors, such as Charles Dickens and others, including George MacDonald, because scholars mention his likely influence on Tolkien and Lewis.

In our multi-dimensional analysis, only the series of words, sentences and interpunctions per chapter, in our opinion, were consciously planned by the authors and, specifically, they do not indicate strong connections between Tolkien, Lewis and MacDonald. Each author distributed words, sentences and interpunctions differently from author to author and, sometimes, even from novel to novel by the same author.

On the contrary, the deep language variables and the linguistic channels, discussed in the paper, are likely due to unconscious design and can reveal connections between texts far beyond writers' awareness.

In summary, the buffers of the extended short-term memory required to readers, the universal readability index of texts, the geometrical representation of texts and the fine tuning of linguistic channels—all tools largely discussed in the paper—have revealed strong connections between *The Lord of the Rings* (Tolkien), *The Chronicles of Narnia* and *The Space Trilogy* (Lewis) on one side, and the strong connection also with some novels by MacDonald on the other side, therefore substantially agreeing with what scholars of English Literature say.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is contained within the article.

Acknowledgments: The author wishes to thank the many scholars who, with great care and love, maintain digital texts available to readers and scholars of different academic disciplines, such as Perseus Digital Library and Project Gutenberg.

Conflicts of Interest: The author declares no conflicts of interest.

Appendix A. Universal Readability Mean Index Lower Bound

The mean value of G_U is given by

$$< G_U >= 89\text{–}10k< C_P > +300 < 1/P_F >\text{–}6(< I_P > \text{–}6) \quad \text{(A1)}$$

The value calculated by introducing the mean of the variables is given by

$$G_{U,mean} = 89\text{–}10k < C_P > +300/ < P_F > \text{–}6(< I_P > \text{–}6) \quad \text{(A2)}$$

Therefore

$$< G_U > - G_{U,mean} = 300 \times \left(\frac{1}{< P_F >} - < \frac{1}{P_F} > \right) \quad \text{(A3)}$$

Now, it can be proved with the Cauchy–Schwarz inequality that $1/ < x > \geq < 1/x >$; therefore $< G_U > - G_{U,mean} \geq 0$; hence $< G_U > \geq G_{U,mean}$.

Appendix B. Pythagorean Distance *d* between Pairs of Texts

Table A1. Pythagorean distance *d* between pairs of texts.

Novel	*Lord*	*Hobbit*	*Narnia*	*Trilogy*	*Back*	*Lilith*	*Oliver*	*David*	*Bleak*	*Tale*	*Mutual*	*Martin*	*Bask*	*Peter*
Lord	0													
Hobbit	0.488	0												
Narnia	0.150	0.348	0											
Trilogy	0.355	0.140	0.211	0										
Back	0.099	0.499	0.203	0.379	0									
Lililth	0.112	0.453	0.174	0.336	0.048	0								
Oliver	0.244	0.518	0.307	0.426	0.146	0.141	0							
David	0.320	0.620	0.407	0.532	0.222	0.234	0.106	0						
Bleak	0.231	0.312	0.170	0.217	0.201	0.153	0.211	0.316	0					
Tale	0.267	0.381	0.252	0.305	0.201	0.161	0.143	0.239	0.096	0				
Mutual	0.146	0.479	0.218	0.369	0.055	0.044	0.099	0.190	0.170	0.151	0			
Martin	0.230	0.2612	0.109	0.139	0.241	0.197	0.294	0.400	0.096	0.192	0.230	0		
Bask	0.424	0.0964	0.277	0.072	0.451	0.408	0.496	0.602	0.286	0.371	0.441	0.211	0	
Peter	0.098	0.4744	0.183	0.355	0.025	0.024	0.146	0.232	0.177	0.182	0.048	0.217	0.427	0

Appendix C. Common Area between Circles

We list the Matlab code to calculate the common area between text, downloaded from https://it.mathworks.com/matlabcentral/answers/273066%E2%80%93overlapping%E2%80%93area%E2%80%93between%E2%80%93two%E2%80%93circles (accessed on 15 June 2024).

Let the distance between the centers of two circles be d and their two radii be r1 and r2. Then, the area, A, of the overlap region of the two circles can be calculated as follows using Matlab's a'tan2' function:

```
t = sqrt((d+r1+r2)*(d+r1-r2)*(d-r1+r2)*(-d+r1+r2));

A = r1^2*atan2(t,d^2+r1^2-r2^2)+r2^2*atan2(t,d^2-r1^2+r2^2)-t/2;
```

Appendix D. Conditional Error Probability

Table A2. Error probability between the indicated texts.

	Lord	*Hobbit*	*Narnia*	*Trilogy*	*Back*	*Lilith*	*Oliver*	*David*	*Bleak*	*Tale*	*Mutual*	*Martin*	*Bask*	*Peter*
Lord	—-													
Hobbit	0.065	—-												
Narnia	0.665	0.387	—-											
Trilogy	0.320	0.759	0.627	—-										
Back	0.711	0.148	0.603	0.335	—-									
Lililth	0.701	0.173	0.639	0.382	0.866	—-								
Oliver	0.314	0.051	0.358	0.203	0.666	0.641	—-							
David	0.033	0	0.101	0	0.496	0.372	0.664	—-						
Bleak	0.307	0.387	0.633	0.593	0.543	0.560	0.397	0.040	—-					
Tale	0.280	0.270	0.475	0.428	0.554	0.601	0.615	0.300	0.725	—-				
Mutual	0.505	0.036	0.545	0.279	0.633	0.681	0.686	0.273	0.428	0.558	—-			
Martin	0.395	0.499	0.757	0.716	0.478	0.528	0.273	0.003	0.732	0.517	0.367	—-		
Bask	0.164	0.830	0.504	0.873	0.211	0.241	0.077	0	0.439	0.286	0.095	0.591	—-	
Peter	0.649	0.231	0.669	0.412	0.853	0.761	0.687	0.568	0.642	0.644	0.594	0.587	0.294	—-

Appendix E. Slope and Correlation Coefficient of the Regression Lines

Table A3. Slope/correlation coefficient of the regression line $y = mx$, Equation (18), modeling the indicated variables (dependent/independent). We keep four digits because some novels differ only at the third and fourth digit.

Novel	Sentences/Words	Interpunctions/Sentences	Words/Interpunctions	Characters/Words
Lord	0.0731/0.9199	2.0372/0.9609	6.6134/0.9609	4.0367/0.9982
Hobbit	0.0608/0.9532	2.1010/0.9936	7.6902/0.9532	4.1014/0.9996
Narnia	0.0729/0.7610	1.9520/0.9384	6.9062/0.7991	4.0907/0.9919
Trilogy	0.0672/0.9325	1.9664/0.9830	7.3380/0.9696	4.2002/0.9976
Back	0.0681/0.9416	2.1640/0.9759	6.6045/0.9799	3.8496/0.9976
Lililth	0.0676/0.8890	2.2619/0.9488	6.2926/0.9800	4.1174/0.9863
Oliver	0.0566/0.9059	3.0893/0.9544	5.6302/0.9638	4.2248/0.9977
David	0.0537/0.9390	3.2949/0.9657	5.5775/0.9882	4.0474/0.9966
Bleak	0.0600/0.9258	2.5324/0.9715	6.51250.9694	4.2235/0.9923
Tale	0.0573/0.9574	2.7972/0.9785	6.1323/0.9912	4.2417/0.9983
Mutual	0.0618/0.9299	2.6814/0.9549	5.9766/0.9740	4.2197/0.9940
Martin	0.0658/0.8583	2.2243/0.9364	6.6785/0.9514	4.3171/0.9939
Bask	0.0684/0.7706	1.8517/0.9366	7.6984/0.9005	4.1320/0.9949
Peter	0.0687/0.8686	2.3080/0.9679	6.1018/0.9152	4.1117/0.9968

Appendix F. Total Signal-to-Noise Ratios Γ_{dB}, in the Four Linguistic Channels

Tables A4–A7 report the signal-to-noise ratio Γ_{dB} in the channels between the input text k (reference) reported in the first row, and the output text j reported in the left column. For example, in Table A4, if the input text is *Lord* and the output text is *Trilogy* then $\Gamma_{dB} = 21.27$ dB; vice versa, $\Gamma_{dB} = 20.44$. A slight asymmetry is typical of linguistic channels [12,15].

Table A4. Total signal-to-noise ratios Γ_{dB}, S–Channels.

Novel	*Lord*	*Hobbit*	*Narnia*	*Trilogy*	*Back*	*Lilith*	*Oliver*	*David*	*Bleak*	*Tale*	*Mutual*	*Martin*	*Bask*	*Peter*
Lord	∞	12.65	10.08	20.44	20.23	18.92	10.61	8.68	13.19	10.18	14.63	14.51	9.79	17.14
Hobbit	14.60	∞	8.21	19.11	19.02	14.75	15.97	17.00	21.64	24.02	23.05	12.73	8.50	13.08
Narnia	10.12	5.30	∞	8.21	7.70	11.54	6.81	4.19	6.85	4.15	7.11	13.32	23.39	13.54
Trilogy	21.27	18.00	9.58	∞	30.77	19.49	13.81	11.96	18.29	14.21	21.14	15.10	9.69	16.57
Back	21.10	17.94	8.87	30.56	∞	17.45	12.67	11.43	16.83	14.07	19.31	13.65	8.86	15.12
Lililth	19.92	13.16	12.79	19.39	17.58	∞	13.99	10.37	15.86	10.98	16.86	23.03	13.29	26.92
Oliver	12.87	17.07	10.19	15.51	14.60	15.61	∞	19.50	22.67	16.83	19.93	15.65	11.20	14.50
David	11.43	18.20	8.42	13.92	13.49	12.83	20.29	∞	19.17	21.60	17.53	12.38	9.09	11.86
Bleak	14.91	21.86	9.79	19.30	18.05	17.27	21.95	18.12	∞	19.17	30.17	15.67	10.43	15.34
Tale	12.66	24.57	7.84	15.85	15.69	13.21	16.61	20.78	19.89	∞	19.45	11.91	8.23	11.93
Mutual	16.13	22.78	9.70	21.87	20.24	18.07	18.92	16.27	29.88	18.43	∞	15.64	10.19	15.76
Martin	16.00	11.43	14.86	15.46	14.23	23.46	13.93	9.79	14.29	9.78	14.62	∞	16.34	26.65
Bask	10.92	6.53	23.97	9.38	8.78	13.09	8.49	5.56	8.33	5.38	8.51	15.69	∞	15.21
Peter	18.10	11.21	14.52	16.19	14.96	26.66	12.56	9.04	13.59	9.38	14.24	26.20	15.13	∞

Table A5. Total signal-to-noise ratios Γ_{dB}, I–Channels.

Novel	*Lord*	*Hobbit*	*Narnia*	*Trilogy*	*Back*	*Lilith*	*Oliver*	*David*	*Bleak*	*Tale*	*Mutual*	*Martin*	*Bask*	*Peter*
Lord	∞	15.57	21.18	19.50	21.74	19.50	9.35	8.36	14.05	11.16	12.37	19.14	17.61	18.44
Hobbit	15.04	∞	11.25	19.77	19.33	13.61	9.23	8.52	13.97	11.74	11.60	12.29	10.07	16.07
Narnia	21.83	12.48	∞	15.46	16.18	17.10	8.65	7.72	12.20	9.93	11.23	18.24	25.28	15.11
Trilogy	20.07	20.66	15.33	∞	20.28	15.03	8.61	7.83	12.86	10.53	11.05	14.25	14.16	15.95

Table A5. *Cont.*

Novel	*Lord*	*Hobbit*	*Narnia*	*Trilogy*	*Back*	*Lilith*	*Oliver*	*David*	*Bleak*	*Tale*	*Mutual*	*Martin*	*Bask*	*Peter*
Back	20.96	18.83	14.72	19.35	∞	19.45	10.31	9.26	16.69	12.90	13.81	17.18	12.61	23.09
Lililth	18.47	12.40	15.76	13.22	18.74	∞	11.43	10.00	17.75	13.45	16.06	27.65	12.92	23.22
Oliver	5.72	5.21	4.61	4.42	7.06	8.72	∞	22.73	12.64	16.56	16.36	8.04	3.42	9.25
David	4.18	4.22	3.04	3.25	5.57	6.66	22.01	∞	10.38	14.45	12.63	5.99	1.96	7.38
Bleak	12.10	11.84	9.57	10.56	15.30	16.42	14.53	12.69	∞	20.10	21.84	14.23	7.91	20.13
Tale	8.26	9.01	6.36	7.46	10.66	11.20	17.84	16.02	19.15	∞	19.40	9.88	5.04	13.18
Mutual	9.97	8.68	8.40	8.02	11.71	14.57	17.59	14.48	21.08	20.07	∞	13.28	6.83	15.34
Martin	18.04	11.32	17.10	12.46	16.71	27.92	10.97	9.59	15.87	12.48	15.04	∞	13.93	19.38
Bask	18.77	12.04	25.74	15.14	14.57	14.71	7.92	7.11	10.98	9.06	10.12	15.52	∞	13.36
Peter	17.31	14.69	13.28	14.34	22.40	22.88	11.85	10.47	20.95	14.94	16.77	18.77	11.10	∞

Table A6. Total signal-to-noise ratios Γ_{dB}, WI–Channels.

Novel	*Lord*	*Hobbit*	*Narnia*	*Trilogy*	*Back*	*Lilith*	*Oliver*	*David*	*Bleak*	*Tale*	*Mutual*	*Martin*	*Bask*	*Peter*
Lord	∞	16.96	8.69	19.72	21.94	20.11	15.14	12.42	28.75	14.96	18.33	29.47	13.83	15.46
Hobbit	15.61	∞	7.80	22.03	13.72	11.77	8.65	7.24	14.25	9.46	10.34	16.38	16.86	10.64
Narnia	7.94	9.59	∞	7.96	6.03	5.48	5.46	3.05	6.96	3.80	5.44	8.92	13.77	10.57
Trilogy	18.74	22.69	6.92	∞	18.24	15.14	10.32	9.40	17.94	12.40	12.81	18.24	13.92	10.84
Back	21.96	15.48	6.80	19.30	∞	26.10	14.37	14.32	26.03	19.46	19.22	19.02	11.69	12.07
Lililth	20.86	13.90	7.07	16.59	26.52	∞	17.01	17.17	24.86	22.59	24.34	18.32	11.15	12.89
Oliver	16.54	11.40	8.63	12.64	15.98	18.25	∞	18.55	17.28	16.43	23.12	15.86	10.46	16.18
David	14.43	10.55	6.50	12.03	15.89	18.35	18.72	∞	15.66	20.68	20.29	13.45	9.04	11.74
Bleak	29.00	15.86	7.97	18.98	26.27	24.36	15.99	13.95	∞	17.22	20.71	23.33	12.70	14.34
Tale	16.13	12.16	5.82	14.41	20.40	23.01	15.15	19.82	18.16	∞	19.83	14.60	9.60	10.71
Mutual	19.40	12.72	7.83	14.60	20.15	24.90	22.43	19.37	21.49	20.26	∞	17.75	10.94	14.62
Martin	29.30	17.61	9.50	19.33	18.82	17.40	14.28	11.19	22.92	13.25	16.43	∞	14.97	16.69
Bask	11.77	16.84	12.10	13.11	9.38	8.32	7.06	4.92	10.38	6.29	7.77	13.14	∞	11.52
Peter	16.67	13.01	12.48	13.17	13.36	13.42	14.94	10.26	15.40	10.80	14.27	17.92	13.59	∞

Table A7. Total signal-to-noise ratios Γ_{dB}, C–Channels.

Novel	*Lord*	*Hobbit*	*Narnia*	*Trilogy*	*Back*	*Lilith*	*Oliver*	*David*	*Bleak*	*Tale*	*Mutual*	*Martin*	*Bask*	*Peter*
Lord	∞	29.12	23.37	27.97	26.10	19.51	26.91	32.93	22.42	26.31	23.84	21.90	26.70	31.43
Hobbit	28.87	∞	20.03	26.66	22.08	17.21	26.33	24.95	20.20	27.13	21.51	20.64	22.77	25.73
Narnia	23.14	20.07	∞	24.11	21.11	28.25	23.63	26.63	30.01	22.43	29.10	25.22	31.08	26.48
Trilogy	27.61	26.30	23.69	∞	20.81	19.94	44.44	27.92	25.21	36.70	27.87	26.30	28.81	32.34
Back	26.52	22.80	21.89	21.57	∞	19.06	21.03	25.94	19.85	20.63	20.44	18.83	22.56	23.80
Lililth	19.17	17.15	28.14	20.28	18.09	∞	20.09	21.25	26.43	19.28	24.45	23.08	23.79	21.31
Oliver	26.50	25.91	23.13	44.38	20.22	19.65	∞	26.67	24.97	39.76	27.56	26.57	27.81	30.39
David	32.88	25.17	26.81	28.28	25.48	21.54	27.08	∞	24.77	25.80	26.30	23.38	31.25	36.01
Bleak	21.76	19.71	29.73	25.11	18.84	26.05	24.98	24.22	∞	23.63	36.70	31.88	29.79	25.53
Tale	25.88	26.71	21.87	36.57	19.78	18.78	39.69	25.31	23.56	∞	25.70	25.32	25.80	28.24
Mutual	23.24	21.04	28.76	27.79	19.51	24.05	27.59	25.83	36.72	25.79	∞	32.93	32.78	27.93
Martin	21.11	19.88	24.71	25.90	17.72	22.43	26.23	22.73	31.63	25.03	32.73	∞	26.77	24.53
Bask	26.34	22.64	30.92	29.07	21.84	23.73	28.13	30.99	30.08	26.19	32.99	27.17	∞	33.27
Peter	31.19	25.68	26.39	32.56	23.22	21.33	30.67	35.87	25.93	28.59	28.28	25.08	33.36	∞

References

1. Carpenter, H. *The Inklings: C. S. Lewis, J.R.R. Tolkien, Charles Williams, and Their Friends*; Houghton Mifflin: Boston, MA, USA, 1979.
2. Glyer, D.P. *The Company They Keep. C. S. Lewis and J. R. R. Tolkien as Writers in Community*; Kent State University Press: Kent, OH, USA, 2007.
3. Duriez, C.; Porter, D. *The Inklings Handbook: A Comprehensive Guide to the Lives, Thought, and Writings of C.S. Lewis, J.R.R. Tolkien, Charles Williams, Owen Barfield, and Their Friends*; Chalice Press: Nashville, TN, USA, 2001.
4. Isley, W.L.C.S. Lewis on Friendship. *Inklings Forever* **2008**, *6*, 9. Available online: https://pillars.taylor.edu/inklings_forever/vol6/iss1/9 (accessed on 30 May 2024).
5. Sammons, M.C. *War of the Fantasy Worlds: C.S. Lewis and J.R.R. Tolkien on Art and Imagination*; Praeger: Westport, CT, USA, 2010.
6. Duriez, C. *The Oxford Inklings. Lewis, Tolkien, and Their Circle*; Lion Books: Oxford, UK, 2015.

7. Hooper, W. The Inklings. In *C. S. Lewis and His Circle. Essays and Memoirs from the Oxford C. S. Lewis Society*; White, R., Wolfe, J., Wolfe, B.N., Eds.; Oxford University Press: New York, NY, USA; Oxford, UK, 2015; pp. 197–213.
8. Gokulapriya, T.J.R.R. Tolkien's Literary Works: A Review. *Int. Rev. Lit. Stud.* **2022**, *4*, 31–39.
9. Matricciani, E. Deep Language Statistics of Italian throughout Seven Centuries of Literature and Empirical Connections with Miller's 7 ∓ 2 Law and Short–Term Memory. *Open J. Stat.* **2019**, *9*, 373–406. [CrossRef]
10. Matricciani, E. A Statistical Theory of Language Translation Based on Communication Theory. *Open J. Stat.* **2020**, *10*, 936–997. [CrossRef]
11. Matricciani, E. Multiple Communication Channels in Literary Texts. *Open J. Stat.* **2022**, *12*, 486–520. [CrossRef]
12. Matricciani, E. Is Short–Term Memory Made of Two Processing Units? Clues from Italian and English Literatures down Several Centuries. *Information* **2024**, *15*, 6. [CrossRef]
13. Matricciani, E. A Mathematical Structure Underlying Sentences and Its Connection with Short-Term Memory. *Appl. Math* **2024**, *4*, 120–142. [CrossRef]
14. Miller, G.A. The Magical Number Seven, Plus or Minus Two. Some Limits on Our Capacity for Processing Information. *Psychol. Rev.* **1956**, *63*, 81–97. [CrossRef] [PubMed]
15. Matricciani, E. Readability Indices Do Not Say It All on a Text Readability. *Analytics* **2023**, *2*, 296–314. [CrossRef]
16. Matricciani, E. Linguistic Mathematical Relationships Saved or Lost in Translating Texts: Extension of the Statistical Theory of Translation and Its Application to the New Testament. *Information* **2022**, *13*, 20. [CrossRef]
17. Papoulis Papoulis, A. *Probability & Statistics*; Prentice Hall: Hoboken, NJ, USA, 1990.
18. Lindgren, B.W. *Statistical Theory*, 2nd ed.; MacMillan Company: New York, NY, USA, 1968.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

AppliedMath

MDPI

Article

Chebyshev Pseudospectral Method for Fractional Differential Equations in Non-Overlapping Partitioned Domains

Shina Daniel Oloniiju [1,*], Nancy Mukwevho [1,2], Yusuf Olatunji Tijani [1] and Olumuyiwa Otegbeye [3]

[1] Department of Mathematics, Rhodes University, Makhanda, P.O. Box 94, Grahamstown 6140, South Africa; nancy.mukwevho@univen.ac.za (N.M.); yusuf.tijani@ru.ac.za (Y.O.T.)
[2] Department of Mathematical and Computational Sciences, University of Venda, Private Bag X5050, Thohoyandou 0950, South Africa
[3] School of Computer Science and Applied Mathematics, University of the Witwatersrand, Private Bag 3, Braamfontein, Johannesburg 2050, South Africa; olumuyiwa.otegbeye@wits.ac.za
* Correspondence: s.oloniiju@ru.ac.za

Abstract: Fractional differential operators are inherently non-local, so global methods, such as spectral methods, are well suited for handling these non-local operators. Long-time integration of differential models such as chaotic dynamical systems poses specific challenges and considerations that make multi-domain numerical methods advantageous when dealing with such problems. This study proposes a novel multi-domain pseudospectral method based on the first kind of Chebyshev polynomials and the Gauss–Lobatto quadrature for fractional initial value problems.The proposed technique involves partitioning the problem's domain into non-overlapping sub-domains, calculating the fractional differential operator in each sub-domain as the sum of the 'local' and 'memory' parts and deriving the corresponding differentiation matrices to develop the numerical schemes. The linear stability analysis indicates that the numerical scheme is absolutely stable for certain values of arbitrary non-integer order and conditionally stable for others. Numerical examples, ranging from single linear equations to systems of non-linear equations, demonstrate that the multi-domain approach is more appropriate, efficient and accurate than the single-domain scheme, particularly for problems with long-term dynamics.

Keywords: multi-domain; pseudospectral method; fractional differential equations; Chebyshev polynomials; Gauss–Lobatto quadrature; stability analysis

check for updates

Citation: Oloniiju, S.D.; Mukwevho, N.; Tijani, Y.O.; Otegbeye, O. Chebyshev Pseudospectral Method for Fractional Differential Equations in Non-Overlapping Partitioned Domains. *AppliedMath* **2024**, *4*, 950–974. https://doi.org/10.3390/appliedmath4030051

Academic Editors: Renhai Wang and Junesang Choi

Received: 26 June 2024
Revised: 29 July 2024
Accepted: 30 July 2024
Published: 2 August 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Fractional differential equations (FDEs) are gaining significant attention in engineering, science, finance and other fields. Unlike classical differential equations, which are limited to integer-order derivatives, FDEs introduce a higher degree of complexity into differential equations by incorporating arbitrary non-integer-order derivatives. Fractional differential operators can be defined in various ways, with the most common definitions being those involving singular kernels, such as the Riemann–Liouville and Caputo fractional differential operators [1,2]. These definitions rely on integrals with power-law kernels, which can sometimes lead to singularities and complicate numerical approximations. These differential operators with power law kernels are, however, advantageous for developing spectral-based methods using basis functions such as the Chebyshev and Legendre polynomials, as these polynomials can be expressed in power series form. Therefore, the Caputo derivative will be used in the numerical method proposed in this study. In recent years, however, there has been a significant interest in non-singular kernels, which provide an alternative framework for defining fractional operators. Examples include the Caputo–Fabrizio and Atangana–Baleanu fractional operators, which use the exponential and Mittag-Leffler functions as kernels, respectively [3,4]. These non-singular kernels were introduced to broaden the scope and applicability of fractional calculus in various fields.

This added complexity allows fractional differential operators to capture the memory effects of a system or describe memory-inherent phenomena. As a result, FDEs offer a more robust framework for modeling the intricate dynamics of complex systems across various scientific disciplines [5,6]. Finding analytical solutions for fractional differential equations (FDEs) is particularly challenging; however, several studies have proposed analytical solutions and methods for FDEs, for instance, Ahmadian and Darvishi [7] derived solitons and traveling wave solutions using the tanh-coth, tan-cot, sech-csch and sec-csc expansion methods for the $(1+1)$-dimensional fractional Biswas–Milovic equation [8]. Kumar et al. [9] proposed a modified Laplace decomposition method, which hybridizes the Laplace transform and Adomian decomposition methods to analyze the fractional Navier–Stokes equation that models the unsteady flow of a Newtonian fluid in a tube. The fractional symmetric regularized long wave equation has been resolved analytically using methods such as the G'/G expansion method [10], the G'/G^2 expansion method [11], the extended complex method [11], the direct algebraic method [12] and the Jacobi elliptic function method [13]. Other analytical methods that have been used for FDEs include the Sumudu transform [14], the Kudryashov method [15] and various semi-analytical techniques such as the exp expansion method [16,17] and the sine-cosine method [18], among others. Numerical methods also play an important role in obtaining accurate solutions for FDEs, and they provide valuable insight into the dynamics of systems described by FDEs [19,20].

Many numerical methods for approximating the solutions of FDEs have been proposed. For example, Hattaf et al. [21] introduced a novel numerical approach for approximating the generalized Hattaf fractional derivative (GHF), which involves a fractional differential operator with a non-singular kernel. The numerical method developed was tested on FDEs with analytical solutions to verify its accuracy by comparing the approximate solutions to the exact solutions. Non-linear FDEs describing the dynamics of HIV infection were also solved to demonstrate the applicability of the method to real-world problems. The predictor–corrector methods (PCMs) for FDEs were proposed by Kumar et al. [22], Jhinga and Daftardar-Gejji [23] and Lee et al. [24]. These studies demonstrated that PCMs solve FDEs faster and drastically reduce computational time. Kumar et al. [25] introduced two computationally efficient numerical techniques based on the mid-point and Heun method for linear and non-linear fractional initial value problems (IVPs). The study concluded that the proposed methods offer significant advantages over the traditional and improved Euler methods.

The non-local nature of fractional differential operators sometimes makes analysis and finding solutions of FDEs herculean tasks [26]. Spectral methods have been a powerful numerical technique for approximating the solutions of FDEs [27–29]. Unlike other traditional numerical techniques, spectral-based methods are global and highly accurate, making them efficient for solving differential equations [30]. Numerous studies have leveraged the global nature of spectral-based methods to discretize fractional differential operators and obtain numerical solutions of FDEs [31]. Zayernouri and Karniadakis [32] and Zhao et al. [33] proposed an exponentially convergent Galerkin spectral method for FDEs. Zayernouri and Karniadakis [32] compared the performance of the developed spectral-based method and the established finite difference method (FDM) and concluded that the proposed method performs better than the FDM in terms of convergence rate. Khader and Sweilam [34] used the Legendre pseudospectral method to solve the fractional advection-dispersion equation with the Caputo fractional differential operator. The study analyzed the accuracy of the new discrete fractional differential operator by examining its convergence rate and estimating the maximum possible error, and the results were consistent with existing results from the literature. Stochastic fractional differential equations were solved by Cardone et al. [35] using the spectral collocation method. The method involves obtaining a fully discrete version of the original continuous operator and solving a system of non-linear algebraic equations. Tests on several problems demonstrated that the proposed method was effective for the chosen basis functions and collocation points.

Spectral methods exhibit exponential convergence for differential equations with sufficiently smooth solutions. This high level of accuracy has made spectral methods, particularly the Chebyshev pseudospectral method, a popular choice for solving a variety of fractional differential equations. For instance, Doha [36] used the Chebyshev pseudospectral method to approximate the solutions of fractional multi-term differential equations. Similarly, Sweilam [37] applied this method to numerically solve fractional integro-differential equations. Other studies that have used the Chebyshev pseudospectral method to approximate the solution of fractional differential equations include those of Dabiri [38], Oloniiju et al. [39,40,41], among others. However, this desirable property of spectral convergence can deteriorate with large time domains or non-smooth solutions [42]. The multi-domain approach tackles these challenges by offering an accurate and computationally efficient way to manage potentially non-smooth solutions. It can also improve numerical stability and accuracy by breaking down a problem's domain, especially time-dependent problems, into smaller sub-domains [43]. Samuel and Motsa [44] described a multi-domain spectral collocation method for solving hyperbolic partial differential equations (PDEs) with a large time domain. The study showed that dividing the computational domain into smaller sub-domains gives accurate results and reduces the computational time. Birem and Klein [45] presented a multi-domain spectral method for the Schrödinger equation. Jung and Olmos-Liceaga [46] introduced a non-conforming and non-overlapping multi-domain spectral method to simulate traveling wave solutions in the reaction-diffusion equations. The method has three key features: high-order accuracy and convergence rate, adaptive mesh and the multi-domain technique. Qin et al. [47] proposed a multi-domain Legendre–Galerkin least square method for linear differential equations with variable coefficients. The study established two key properties of the method: continuity, ensuring consistency and smoothness at the endpoints of each sub-domain, and coercivity, ensuring stability and convergence. Additionally, the study established the optimal error estimate in the H_1 norm.

This study explores the Chebyshev pseudospectral method in a multi-domain setting as a novel approach to solving FDEs. Here, the domain of the problem is divided into a finite number of non-overlapping sub-domains. The multi-domain approach for FDEs has been applied to problems in biology [48] and finance [49,50]. Maleki and Kajani [48] used the multi-domain Legendre–Gauss pseudospectral method to approximate solutions for the fractional Volterra population growth model. The study demonstrated the effectiveness of the proposed method in tackling a wide range of integro-differential equations, including those with integer and fractional orders. Bambe Moutsinga et al. [49] developed a spectral-based method for fractional differential equations which models valuation within an affine jump-diffusion framework, chaotic and hyper-chaotic systems and fractional pricing models for cryptocurrencies. The study established that the developed spectral method outperforms the MATLAB numerical packages based on the Euler method. Kharazmi [50] investigated a data-driven approach to solving stochastic fractional partial differential equations (SFPDEs) using multi-domain spectral methods. Zhao et al. [51] proposed a multi-domain spectral collocation method (MDSCM) for solving non-linear FDEs. The MDSCM exhibited a significant advantage over the single-domain spectral methods in achieving high accuracy for differential equations with low-regular solutions. Klein and Stoilov [52] proposed a numerical technique for the fractional Laplacian, based on the Riesz fractional operators, defined on $\mathbb{R}$. The integrands are transformed using the underlying Z_q curve to ensure accurate computation within each sub-interval. The transformed integrals within each sub-interval are then solved using the Clenshaw–Curtis algorithm. Tavassoli Kajani [53] proposed a modified Müntz–Legendre collocation technique on a single domain and in a multi-domain setting for the fractional pantograph equation. Xu and Hesthaven [54] proposed a highly accurate and stable multi-domain spectral method with a penalty method for fractional partial advection and diffusion equations. The study concluded that the proposed scheme's high accuracy makes it well suited

for long-term integration of FDEs. It was also noted that the convergence rate aligns with the theoretical analysis of the scheme.

The present study aims to introduce a novel numerical method, the non-overlapping multi-domain Chebyshev pseudospectral method, based on the first kind of Chebyshev polynomials and the Gauss–Lobatto quadrature for fractional differential equations. The choice of Chebyshev polynomials is due to their orthogonality property with respect to their weight function, which simplifies the projection of the approximate solution onto the polynomial space. Additionally, the computation of Chebyshev-Gauss–Lobatto quadrature weights and nodes is quite straightforward, and this reduces computational overhead compared to other orthogonal polynomials. The proposed method works by decomposing the domain of a fractional differential equation into smaller, non-overlapping sub-domains. The novelty of this study lies in its approach of partitioning the domain of a fractional initial value problem into non-overlapping sub-intervals and projecting the solution of the problem onto the space of Chebyshev orthogonal polynomials of the first kind. This technique involves separating the fractional differential operator into the sum of the 'local' and 'memory' components within each sub-domain. To demonstrate the versatility and efficiency of the proposed method, stability analysis and numerical experimentation on various types of fractional differential equations are carried out. Numerical error analysis is conducted to demonstrate the advantages of the domain decomposition approach over the single-domain Chebyshev pseudospectral method.

The rest of the article is organized as follows: Section 2 provides certain preliminary definitions which are necessary for developing the numerical schemes; Section 3 is devoted to developing and analyzing the numerical scheme for FDEs using the Chebyshev polynomials on the Gauss–Lobatto quadrature in non-overlapping partitioned domains; Section 4 demonstrates the accuracy and effectiveness of the numerical method by presenting some examples ranging from linear single equations to systems of non-linear equations that require long-term integration; and Section 5 ultimately concludes the study.

2. Preliminaries

This section introduces and defines the basic concepts and fractional operators required to develop the numerical scheme.

Definition 1. *Suppose $u : [a, b] \to \mathbb{R}$ is a continuous function, $\gamma > 0$ and the Riemann–Liouville fractional order integral of $u(t)$ is defined as [55]*

$$ {}^{RL}_{a}I^{\gamma}_{t}u(t) = \begin{cases} \frac{1}{\Gamma(\gamma)}\left(t^{\gamma-1} * u(t)\right) = \frac{1}{\Gamma(\gamma)}\int_{a}^{t}(t-\tau)^{\gamma-1}u(\tau)d\tau, & t > 0,\ \gamma > 0, \\ u(t), & \gamma = 0, \end{cases} \tag{1} $$

where Γ is the Euler gamma function and $\left(t^{\gamma-1} * u(t)\right)$ is the convolution product of the power function, $t^{\gamma-1}$, and the function $u(t)$.

Definition 2 (Fractional integral of $(t-a)^j$)**.** *For the power function $(t-a)^j$, the fractional integral is defined as [56]*

$$ {}^{RL}_{a}I^{\gamma}_{t}(t-a)^{j} = \frac{\Gamma(j+1)}{\Gamma(j+\gamma+1)}(t-a)^{j+\gamma}. \tag{2} $$

Definition 3. *The left-sided Caputo fractional differential operator is defined as [55]*

$$ {}^{C}_{a}D^{\gamma}_{t}u(t) = \frac{1}{\Gamma(\lceil\gamma\rceil-\gamma)}\left(t^{\lceil\gamma\rceil-\gamma-1} * u^{\lceil\gamma\rceil}(t)\right) = \frac{1}{\Gamma(\lceil\gamma\rceil-\gamma)}\int_{a}^{t}(t-\tau)^{\lceil\gamma\rceil-\gamma-1}u^{\lceil\gamma\rceil}(\tau)d\tau,\ \lceil\gamma\rceil-1 < \gamma \leq \lceil\gamma\rceil, \tag{3} $$

where $\lceil\gamma\rceil$ *is the ceiling of* γ, $\lceil\gamma\rceil \geq \gamma$ *and* $\left(t^{\lceil\gamma\rceil-\gamma-1} * u^{\lceil\gamma\rceil}(t)\right)$ *is the convolution product of the power function,* $t^{\lceil\gamma\rceil-\gamma-1}$, *and the* $\lceil\gamma\rceil$*-th integer-order derivative of the function* $u(t)$.

Definition 4 (Fractional derivative of $(t-a)^j$)**.** *The fractional derivative of* $(t-a)^j$ *is defined as [56]*

$$ {}_a^C D_t^\gamma (t-a)^j = \begin{cases} 0 & j \in \mathbb{N}_0, \quad j < \lceil\gamma\rceil \\ \frac{\Gamma(j+1)}{\Gamma(j+1-\gamma)}(t-a)^{j-\gamma} & j \in \mathbb{N}_0, \quad j \geq \lceil\gamma\rceil. \end{cases} \tag{4} $$

Definition 5. *The set of collocation points used in this study is that of the Chebyshev–Gauss–Lobatto points defined as* $t_r = -\cos\left(\frac{r\pi}{N}\right)$; $r = 0(1)N$ *[57].*

3. Numerical Discretization and Analysis via the Pseudospectral Method

This section introduces the numerical discretization of a fractional initial value problem using Chebyshev polynomials of the first kind and the Gauss–Lobatto quadrature. It covers the approximations of the fractional derivative of an arbitrary function, both in a single domain and in partitioned, non-overlapping domains. Firstly, the polynomial approximation of an arbitrary function $u(t)$, continuous on an interval $[0, h]$, is presented in terms of shifted first-kind Chebyshev polynomials. These polynomials are obtained through the following recurrence formula:

$$ \Pi_{h,0}(t) = 1, \quad \Pi_{h,1}(t) = \frac{2t}{h} - 1, \quad \Pi_{h,j+1}(t) = 2\Pi_{h,1}(t) \cdot \Pi_{h,j}(t) - \Pi_{h,j-1}(t), \quad j \geq 1. \tag{5} $$

The function $u(t)$ is represented as follows:

$$ u(t) \approxeq U_h(t) = \sum_{j=0}^{\infty} \hat{u}_j \Pi_{h,j}(t), \quad t \in [0, h], \tag{6} $$

where the coefficients $\hat{u}_j$ in Equation (6) satisfy the ℓ^2 orthogonality condition of the shifted first-kind Chebyshev polynomials, which are defined with respect to the inner product

$$ \hat{u}_j = \langle U_h(t), \Pi_{h,j}(t) \rangle = \frac{1}{\nu_j} \int_0^h \frac{U_h(t)\Pi_{h,j}(t)}{\sqrt{ht - t^2}} dt = \frac{1}{\nu_j} \sum_{i=0}^{\infty} U_h(t_i)\Pi_{h,j}(t_i)\varrho_i, \tag{7} $$

where

$$ \nu_j = \begin{cases} \pi & j = 0 \\ \frac{\pi}{2} & \text{otherwise.} \end{cases} \tag{8} $$

If a truncated N-th order approximation of the function is considered and collocated at $N+1$ distinct Chebyshev–Gauss–Lobatto (CGL) points, with $t_r \in [0, h]; r = 0(1)N$, then

$$ u(t) \approxeq \sum_{i=0}^{N} U_h(t_i)\varrho_i \sum_{j=0}^{N} \frac{1}{\nu_j} \Pi_{h,j}(t_i)\Pi_{h,j}(t_r), \qquad r = 0(1)N, \tag{9} $$

where ϱ_i is the Christoffel number associated with the Chebyshev–Gauss–Lobatto quadrature and is defined as follows:

$$ \varrho_i = \begin{cases} \frac{\pi}{2N} & I = 0, N \\ \frac{\pi}{N} & \text{otherwise.} \end{cases} \tag{10} $$

Using the following matrix representations

$$\boldsymbol{\rho} = \begin{bmatrix} \varrho_0 & & & \\ & \varrho_1 & & \\ & & \ddots & \\ & & & \varrho_N \end{bmatrix}, \quad \boldsymbol{\vartheta} = \begin{bmatrix} \frac{1}{\nu_0} & & & \\ & \frac{1}{\nu_1} & & \\ & & \ddots & \\ & & & \frac{1}{\nu_N} \end{bmatrix}, \quad \boldsymbol{\Pi}_h = \begin{bmatrix} \Pi_{h,0}(t_0) & \Pi_{h,0}(t_1) & \dots & \Pi_{h,0}(t_N) \\ \Pi_{h,1}(t_0) & \Pi_{h,1}(t_1) & \dots & \Pi_{h,1}(t_N) \\ \vdots & \vdots & \ddots & \vdots \\ \Pi_{h,N}(t_0) & \Pi_{h,N}(t_1) & \cdots & \Pi_{h,N}(t_N) \end{bmatrix}, \tag{11}$$

Equation (9) can be expressed as

$$u(t) \approxeq \sum_{i=0}^{N} U_h(t) E_i(t_r) = \mathbf{E}^{\mathsf{T}} \mathbf{U}_h, \tag{12}$$

where $\mathbf{E} = \boldsymbol{\rho}\, \boldsymbol{\Pi}_h^{\mathsf{T}}\, \boldsymbol{\vartheta}\, \boldsymbol{\Pi}_h$ and $\mathbf{U}_h = [U_h(t_0) \quad U_h(t_1) \quad \dots \quad U_h(t_N)]^{\mathsf{T}}$.

3.1. The Discretization of the Fractional Differential Operator in a Single Domain

Now, let us approximate the arbitrary non-integer order of the function $u(t)$. Assuming that $u(t) \in \mathcal{C}^p([0,h])$, where $p \geq N + \lceil \gamma \rceil + 1$, representing the space of continuously differentiable functions, $t > 0$ and $\gamma > 0$, then the approximation for the arbitrary non-integer derivative of $u(t)$ is given by

$${}_0^C D_t^{\gamma} u(t) \approxeq \sum_{i=0}^{N} U_h(t_i) \varrho_i \sum_{j=0}^{N} \frac{1}{\nu_j} \Pi_{h,j}(t_i) {}_0^C D_t^{\gamma} \Pi_{h,j}(t). \tag{13}$$

The fractional-order derivative of the Chebyshev polynomial ${}_0^C D_t^{\gamma} \Pi_{h,j}(t)$ is obtained by using the series form of the polynomial [56], as follows:

$${}_0^C D_t^{\gamma}\, \Pi_{h,j}(t) = j \sum_{k=0}^{j} \frac{(-1)^{j-k}(j+k-1)!2^{2k}}{(j-k)!(2k)!h^k} {}_0^C D_t^{\gamma} t^k = j \sum_{k=\lceil \gamma \rceil}^{j} \frac{(-1)^{j-k}(j+k-1)!2^{2k}}{(j-k)!(2k)!h^k} \frac{\Gamma(k+1)}{\Gamma(k-\gamma+1)} t^{k-\gamma}. \tag{14}$$

Proceeding to approximate $t^{k-\gamma}$ as a linear combination of shifted first-kind Chebyshev polynomials at the CGL points, Equation (14) can now be expressed as [39]:

$${}_0^C D_t^{\gamma}\, \Pi_{h,j}(t) \approxeq {}_h\mathcal{D}^{(\gamma)} \boldsymbol{\Pi}_h, \tag{15}$$

where ${}_h\mathcal{D}^{(\gamma)} = \boldsymbol{\Delta}\, \boldsymbol{\Omega}\, \boldsymbol{\Phi}\, \boldsymbol{\Delta}^{\mathsf{T}}\, \boldsymbol{\vartheta}$ and $\boldsymbol{\Delta}, \boldsymbol{\Omega}$ and $\boldsymbol{\Phi}$ are matrices whose entries are defined as follows:

$$\left.\begin{aligned} \Delta_{j,k} &= \begin{cases} 1 & j = 0, k = 0 \\ \frac{j(-1)^{j-k}(j+k-1)!2^{2j}}{(j-k)!(2k)!h^k} & j = 1(1)N,\ k = 0(1)j, \\ 0 & \text{otherwise} \end{cases} \qquad \Omega_{k,k} = \begin{cases} 0 & k < \lceil \gamma \rceil \\ \frac{\Gamma(k+1)}{\Gamma(k-\gamma+1)} & k = \lceil \gamma \rceil (1) N' \end{cases} \\ \Phi_{p,k} &= \begin{cases} 0 & k < \lceil \gamma \rceil \\ \frac{\sqrt{\pi} h^{k-\gamma+p} \Gamma(k-\gamma+p+\frac{1}{2})}{\Gamma(k-\gamma+p+1)} & k = \lceil \gamma \rceil (1) N,\ p = 0(1)s,\ s = 0(1)N. \end{cases} \end{aligned}\right\} \tag{16}$$

Therefore, ${}_0^C D_t^{\gamma} u(t)$ is now approximated as

$${}_0^C D_t^{\gamma} u(t) \approxeq ({}_{0,h}\mathbf{D}^{\gamma})^{\mathsf{T}}\, \mathbf{U}_h, \tag{17}$$

where ${}_{0,h}\mathbf{D}^{\gamma} = \boldsymbol{\rho}\, \boldsymbol{\Pi}_h^{\mathsf{T}}\, \boldsymbol{\vartheta}\, {}_h\mathcal{D}^{(\gamma)}\, \boldsymbol{\Pi}_h$. We refer to $({}_{0,h}\mathbf{D}^{\gamma})^{\mathsf{T}}$ as the fractional differentiation matrix in the domain $[0,h]$.

3.2. Discretization of Fractional Differential Operator in Non-Overlapping Partitioned Domains

Assume that $u(t)$ is defined in $t \in [0, T]$, partitioned into q non-overlapping sub-intervals, such that $0 = t_0 < t_1 < \ldots < t_q = T$ and $h_l = t_l - t_{l-1}$, $l = 1(1)q$. Suppose $t \in [t_l, t_{l+1}]$ and $\lceil\gamma\rceil - 1 < \gamma \leq \lceil\gamma\rceil$; then,

$$\left.\begin{aligned} {}_0^C D_t^\gamma u(t) &= \frac{1}{\Gamma(\lceil\gamma\rceil-\gamma)} \int_0^t (t-\tau)^{\lceil\gamma\rceil-\gamma-1} u^{\lceil\gamma\rceil}(\tau) d\tau \\ &= \sum_{e=1}^{l} \left[{}_{t_{e-1}}^C D_{t_e}^\gamma u(t) \right] + {}_{t_l}^C D_t^\gamma u(t) \\ &= \sum_{e=1}^{l} \left[\frac{1}{\Gamma(\lceil\gamma\rceil-\gamma)} \int_{t_{e-1}}^{t_e} (t-\tau)^{\lceil\gamma\rceil-\gamma-1} u^{\lceil\gamma\rceil}(\tau) d\tau \right] + \frac{1}{\Gamma(\lceil\gamma\rceil-\gamma)} \int_{t_l}^{t} (t-\tau)^{\lceil\gamma\rceil-\gamma-1} u^{\lceil\gamma\rceil}(\tau) d\tau. \end{aligned}\right\} \tag{18}$$

The first term in Equation (18) is the 'memory', and the latter term of the equation is the local arbitrary-order derivative of $u(t)$ with respect to $t \in [t_l, t_{l+1}]$. The arbitrary-order derivative in the first sub-domain is local and calculated with no memory term. The two terms in Equation (18) are approximated separately. If the length, h_l, of each sub-domain, $[t_l, t_{l+1}]$, is equivalent to h (from discretization in a single domain) in Section 3.1, and we collocate at $N+1$ distinct CGL points, $t_r^l \in [t_l, t_{l+1}]$, with $t_N^l \equiv t_0^{l+1}$, then the second term in Equation (18) is approximated as follows:

$${}_{t_l}^C D_t^\gamma u(t) \approxeq ({}_{t_l,h_l}\mathbf{D}_l^\gamma)^\mathsf{T} \mathbf{U}_{h_l,l}, \quad t \in [t_l, t_{l+1}]. \tag{19}$$

The first term in Equation (18), the 'memory', which is the arbitrary non-integer-order derivative, ${}_{t_{e-1}}^C D_{t_e}^\gamma u(t)$, is approximated as

$${}_{t_{e-1}}^C D_{t_e}^\gamma u(t) \approxeq \sum_{i=0}^{N} U_h(t_i^e) \varrho_i \sum_{j=0}^{N} \Pi_{h_l,j}(t_i^e) \, {}_{t_{e-1}}^C D_{t_e}^\gamma \Pi_{h_l,j}(t), \quad t_i^e \in [t_{e-1}, t_e]. \tag{20}$$

The derivative ${}_{t_{e-1}}^C D_{t_e}^\gamma \Pi_{h_l,j}(t)$ at the CGL points t_r^l is approximated as ${}_{t_{e-1}}^C D_{t_e}^\gamma \Pi_{h_l,j}(t_r^l) \approxeq \boldsymbol{\Delta}\,\boldsymbol{\Theta}$, where the entries of $\boldsymbol{\Theta}$ are defined as

$$\Theta_{j,r} = \begin{cases} 0 & j < \lceil\gamma\rceil \\ \frac{j!}{\Gamma(\lceil\gamma\rceil-\gamma)(j-\lceil\gamma\rceil)!} \int_{t_{e-1}}^{t_e} (t_r^l - \tau)^{\lceil\gamma\rceil-1-\gamma} (\tau - t_{e-1})^{j-\lceil\gamma\rceil} d\tau, & \substack{t_r^l \in [t_l, t_{l+1}] \\ j = \lceil\gamma\rceil(1)N,\ r = 0(1)N.} \end{cases} \tag{21}$$

The integral in Equation (21) can be computed using any quadrature formula, such as Newton–Cotes or Gaussian quadrature formulas. Therefore, the first term of Equation (18) is then calculated as

$$\sum_{e=1}^{l} {}_{t_{e-1}}^C D_{t_e}^\gamma u(t) \approxeq \sum_{e=1}^{l} ({}_{t_{e-1},t_e}\mathbf{D}_e^\gamma)^\mathsf{T} \mathbf{U}_{h_l,e}, \tag{22}$$

where $\mathbf{U}_{h_l,e} = [U_{h_l}(t_0^e) \quad U_{h_l}(t_1^e) \quad \ldots \quad U_{h_l}(t_N^e)]^\mathsf{T}$, t_i^e are the CGL points mapped onto $[t_{e-1}, t_e]$ and ${}_{t_{e-1},t_e}\mathbf{D}_e^\gamma = \boldsymbol{\rho}\,\boldsymbol{\Pi}_{h_l}^\mathsf{T}\,\boldsymbol{\vartheta}\,\boldsymbol{\Delta}\,\boldsymbol{\Theta}$.

3.3. Pseudospectral Discretization of FIVPs in Non-Overlapping Partitioned Domain

Consider the fractional initial value problem (FIVP)

$$\begin{cases} {}_0^C D_t^\gamma u(t) = \kappa u(t) + f(t), & t \in (0, T], \\ u(0) = u_0, \end{cases} \tag{23}$$

where $\kappa \in \mathbb{C}$ and $\gamma \in (0, 1)$. Suppose the domain $[0, T]$ is partitioned into q non-overlapping domains, such that $0 = t_0 < t_1 < t_2 < \ldots < t_q = T$, and as before, we define $h_l = t_l - t_{l-1}$,

$l = 1(1)q$. To obtain the solution of Equation (23) on an arbitrary domain $[t_l, t_{l+1}]$, we rewrite the equation as

$$\begin{cases} \sum_{e=1}^{l} {}_{t_{e-1}}^{C}D_{t_e}^{\gamma}u(t) + {}_{t_l}^{C}D_t^{\gamma}u(t) = \kappa u(t) + f(t), & t \in [t_l, t_{l+1}], \quad l = 1(1)q-1, \\ u(0) = u_0. \end{cases} \tag{24}$$

Assume that the solution $u(t)$ of Equation (24) can be approximated as a linear combination of functions which are in the space of shifted Chebyshev polynomials of at most degree N on the CGL quadrature. Then, using the discretizations of the differential operator given in Sections 3.1 and 3.2, the FIVP (24) is now equivalent to the following algebraic systems:

$$\left.\begin{aligned} ({}_{0,h_l}\mathbf{D}_1^{\gamma})^{\mathsf{T}}\mathbf{U}_{h_l,1} &= \mathbf{E}_1^{\mathsf{T}}\mathbf{f} \\ U_{h_l}(t_0^1) &= u_0, \quad t \in [t_0, t_1 \equiv t_0 + h_l], \end{aligned}\right\} \tag{25}$$

and

$$\left.\begin{aligned} &\sum_{e=1}^{l}({}_{t_{e-1},t_e}\mathbf{D}_e^{\gamma})^{\mathsf{T}}\mathbf{U}_{h_l,e} + ({}_{t_l,h_l}\mathbf{D}_l^{\gamma})^{\mathsf{T}}\mathbf{U}_{h_l,l} = \kappa\mathbf{E}_l^{\mathsf{T}}\mathbf{U}_{h_l,l} + \mathbf{E}_l^{\mathsf{T}}\mathbf{f} \\ &U_{h_l}(t_0^l) = U_{h_l}(t_N^{l-1}), \quad t \in [t_l, t_{l+1} \equiv t_l + h_l], \quad l = 2(1)q-1, \end{aligned}\right\} \tag{26}$$

where $\mathbf{f} = [f(t_0^l), f(t_1^l), \ldots, f(t_N^l)]^{\mathsf{T}}$ and $l = 1(1)q$.

3.4. The Numerical Stability of the Scheme

The stability of the proposed numerical method is demonstrated by considering the following linear fractional initial value problem

$${}_0^{C}D_t^{\gamma}u(t) = \kappa u(t), \quad \kappa \in \mathbb{C} \quad s.t. \quad Re(\kappa) \leq 0, \quad \gamma \in (0,1), \quad t \in [0,h], \quad h \in \mathbb{R}_+, \tag{27}$$

whose exact solution is $u(t) = u_0 E_{\gamma,1}(\kappa t^{\gamma})$, where $E_{\gamma,1}(\cdot)$ is the two-parameter Mittag-Leffler function. If we consider a partition of the domain $[0,h]$ into $q = h$ equal sub-domains, the discretization of the model Equation (27) through the proposed numerical method in an arbitrary sub-domain $[t_l, t_{l+1}] \equiv [t_l, t_l + h_l]$ is obtained as

$$\left[\frac{1}{h_l^{\gamma}}({}_{0,h_l}\mathbf{D}_l^{\gamma})^{\mathsf{T}} - \kappa\mathbf{E}_l^{\mathsf{T}}\right]\mathbf{U}_{h_l,l} = -\sum_{e=1}^{l}({}_{t_{e-1},t_e}\mathbf{D}_e^{\gamma})^{\mathsf{T}}\mathbf{U}_{h_l,e}. \tag{28}$$

Here, $({}_{t_{e-1},t_e}\mathbf{D}_e^{\gamma})^{\mathsf{T}}$ is the memory differentiation matrix as in Equation (22), $({}_{0,h_l}\mathbf{D}_l^{\gamma})^{\mathsf{T}}$ is the fractional differentiation matrix in a single domain $[0,h_l]$ and $({}_{t_l,h_l}\mathbf{D}_l^{\gamma})^{\mathsf{T}} \equiv \frac{1}{h_l^{\gamma}}({}_{0,h_l}\mathbf{D}_l^{\gamma})^{\mathsf{T}}$. To study the stability of the proposed pseudospectral numerical scheme, we define the stability matrix

$$S(w) = [P - wQ]^{-1}R, \tag{29}$$

where $w = \kappa h_l^{\gamma}$ and P, Q and R are the matrices defined, respectively, as

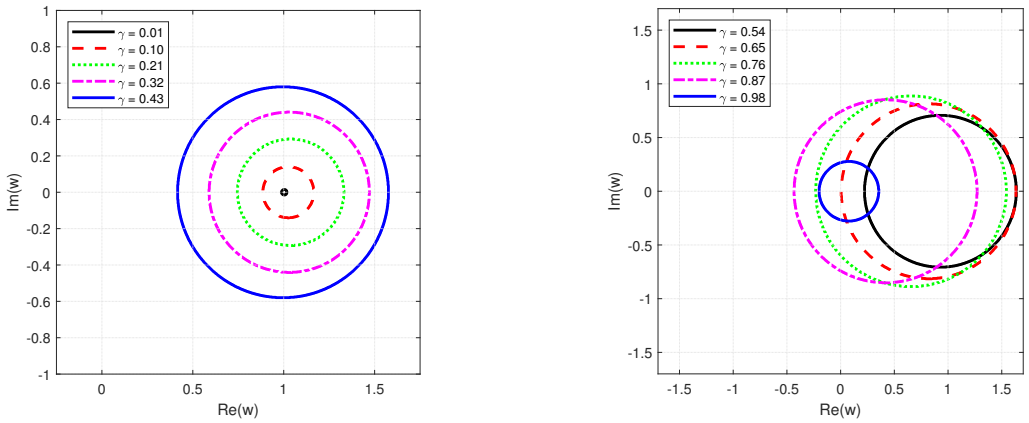
$$P = \begin{bmatrix} ({}_{0,h_l}\mathbf{D}_2^{\gamma})^{\mathsf{T}} & 0 & \dots & 0 \\ 0 & ({}_{0,h_l}\mathbf{D}_3^{\gamma})^{\mathsf{T}} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & ({}_{0,h_l}\mathbf{D}_q^{\gamma})^{\mathsf{T}} \end{bmatrix}, \quad Q = \begin{bmatrix} \mathbf{E}_2^{\mathsf{T}} & 0 & \dots & 0 \\ 0 & \mathbf{E}_3^{\mathsf{T}} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{E}_q^{\mathsf{T}} \end{bmatrix},$$
$$R = \begin{bmatrix} -h_l^{\gamma}({}_{t_0,t_1}\mathbf{D}_1^{\gamma})^{\mathsf{T}} & 0 & \dots & 0 \\ -h_l^{\gamma}({}_{t_0,t_1}\mathbf{D}_2^{\gamma})^{\mathsf{T}} & -h_l^{\gamma}({}_{t_1,t_2}\mathbf{D}_2^{\gamma})^{\mathsf{T}} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -h_l^{\gamma}({}_{t_0,t_1}\mathbf{D}_{q-1}^{\gamma})^{\mathsf{T}} & -h_l^{\gamma}({}_{t_1,t_2}\mathbf{D}_{q-1}^{\gamma})^{\mathsf{T}} & \dots & -h_l^{\gamma}({}_{t_{q-2},t_{q-1}}\mathbf{D}_{q-1}^{\gamma})^{\mathsf{T}} \end{bmatrix}. \tag{30}$$

The stability of the numerical scheme is determined by the eigenvalues of $S(w)$. The stability region of the numerical scheme (28) is typically defined by the set of points in the $\mathbb{C}$-plane for which the spectral radius or the modulus of dominant eigenvalue is less than or equal to 1. If there is more than one non-zero eigenvalue with the potential to cross the stability threshold of 1, the stability region is then determined by considering these eigenvalues and the regions where these eigenvalues have a magnitude less than or equal to 1. The overall stability region of the numerical scheme would then be the intersection of these regions. If the stability region includes $Re(w) \leq 0$ (the entire left half of the $\mathbb{C}$-plane), then the numerical scheme is said to be absolutely stable (A-stable) [58]. If, in addition to A-stability, the limit of the dominant eigenvalue tends to zero as w tends to $-\infty$, then the numerical scheme is said to be Lubich-stable (L-stable) [59]. The L-stability property means that any components that may cause instability in the numerical scheme are damped as the length of the sub-domain, h_l, increases, which is particularly important when handling long-term integration of stiff differential equations. To illustrate, Table 1 shows the dominant eigenvalue, stability region and stability property for a linear discretization ($N = 1$) of Equation (28) with two sub-domains ($q = 2$) for some selected values of fractional order γ. As shown in the table, the stability region of the numerical scheme is the exterior of some closed disks in $\mathbb{C}$; the boundaries of these disks are illustrated in Figure 1. As indicated by the figure, the stability regions for $\gamma = 0.01, 0.10, 0.21, 0.32, 0.43, 0.54, 0.65$ satisfy the conditions for absolute and Lubich stability, and so the numerical scheme for these values of γ is absolutely and Lubich-stable. However, the numerical schemes for $\gamma = 0.76, 0.87, 0.98$ are not absolutely stable and, by extension, not L-stable.

In Figure 2, the stability regions for a quadratic approximation with $q = 3$ and $q = 5$ for $\gamma = 0.21, 0.54, 0.98$ are presented. The stability region in the figure is the area outside the elliptical disks, which shows that the stability regions for $q = 3$ and $q = 5$ sub-domains are the same. However, in both cases, a varying number of eigenvalues with the potential to cross the stability threshold are obtained. For $q = 3$, there is one non-zero dominant eigenvalue for $\gamma = 0.21$ and 0.54, and there are four non-zero eigenvalues for $\gamma = 0.98$ with the potential to cross the stability threshold of 1. For $q = 5$, $\gamma = 0.21, 0.54, 0.98$ all have eight non-zero eigenvalues that can cross the stability threshold. However, when the intersection of the resulting stability regions from these eigenvalues is taken, the same stability region with $q = 3$ is obtained for each illustrated γ value, as shown in Figure 2. The figure shows that the numerical schemes for $\gamma = 0.21$ and $\gamma = 0.54$ are absolutely stable, and the scheme for $\gamma = 0.98$ is not absolutely stable (conditional stability). Although a general conclusion about the stability of the numerical scheme cannot be made, it can be inferred from this analysis that whether the numerical scheme is absolutely or conditionally stable depends on the value of the arbitrary non-integer order of the fractional differential equation.

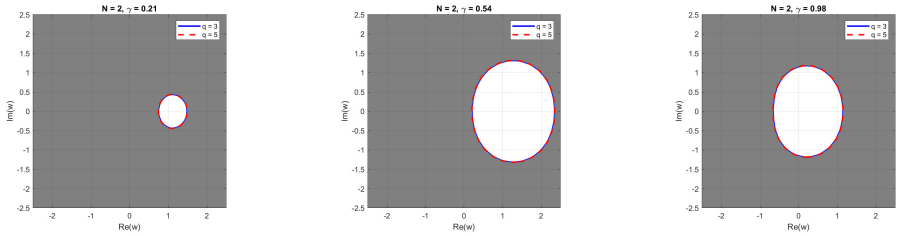
Table 1. The dominant eigenvalue, stability region and stability property for some selected values of arbitrary order, γ, for a linear approximation ($N = 1$) with $q = 2$.

γ	Dominant Eigenvalue	Stability Region	Stability Property
0.01	$\frac{-124958025860977}{9007199254740992w - 9034579827725136}$	$\left\|w - \frac{9034579827725136}{9007199254740992}\right\| \geq \frac{124958025860977}{9007199254740992}$	A–stable L–stable
0.10	$\frac{-1254334826011059}{9007199254740992w - 9235870117935490}$	$\left\|w - \frac{9235870117935490}{9007199254740992}\right\| \geq \frac{1254334826011059}{9007199254740992}$	A–stable L–stable
0.21	$\frac{-2627472107360489}{9007199254740992w - 9350065552559051}$	$\left\|w - \frac{9350065552559051}{9007199254740992}\right\| \geq \frac{2627472107360489}{9007199254740992}$	A–stable L–stable
0.32	$\frac{-3959902260456486}{9007199254740992w - 9279409270142741}$	$\left\|w - \frac{9279409270142741}{9007199254740992}\right\| \geq \frac{3959902260456486}{9007199254740992}$	A–stable L–stable
0.43	$\frac{-10428492653806604}{18014398509481984w - 17946001659717557}$	$\left\|w - \frac{17946001659717557}{18014398509481984}\right\| \geq \frac{10428492653806604}{18014398509481984}$	A–stable L–stable
0.54	$\frac{-12700697864360454}{18014398509481984w - 16735571476860095}$	$\left\|w - \frac{16735571476860095}{18014398509481984}\right\| \geq \frac{12700697864360454}{18014398509481984}$	A–stable L–stable
0.65	$\frac{-7316166143825643}{9007199254740992w - 7382436898791357}$	$\left\|w - \frac{7382436898791357}{9007199254740992}\right\| \geq \frac{7316166143825643}{9007199254740992}$	A–stable L–stable
0.76	$\frac{-15952263845377693}{18014398509481984w - 11810807580156978}$	$\left\|w - \frac{11810807580156978}{18014398509481984}\right\| \geq \frac{15952263845377693}{18014398509481984}$	
0.87	$\frac{-30618033637964479}{36028797018963968w - 15079171789616800}$	$\left\|w - \frac{15079171789616800}{36028797018963968}\right\| \geq \frac{30618033637964479}{36028797018963968}$	
0.98	$\frac{-9934148207590571}{36028797018963968w - 2781302084999484}$	$\left\|w - \frac{2781302084999484}{36028797018963968}\right\| \geq \frac{9934148207590571}{36028797018963968}$	



(**a**) (**b**)

Figure 1. The boundaries of the closed disks in Table 1. The stability region of the linear discretization with two sub-domains for the indicated values of arbitrary non-integer order, (**a**) $\gamma = 0.01, 0.10, 0.21, 0.32, 0.43$ and (**b**) $\gamma = 0.54, 0.65, 0.76, 0.87, 0.98$, lies outside these circles.



(**a**) (**b**) (**c**)

Figure 2. The stability regions of quadratic approximations with $q = 3$ and $q = 5$ for (**a**) $\gamma = 0.21$, (**b**) $\gamma = 0.54$ and (**c**) $\gamma = 0.98$. The stability regions are the shaded areas located outside the elliptical disks.

4. Numerical Experimentation and Results

This section presents a series of numerical examples, ranging from single linear equations to systems of non-linear equations, along with the results. Detailed analyses of the results for each example are provided to illustrate key observations. For a comprehensive understanding of the numerical methods used in these examples, refer to the description provided in Section 3.

Example 1. *Consider the linear fractional differential equation [60]*

$$ {}_0^C D_t^\gamma u(t) + u(t) = \frac{2t^{2-\gamma}}{\Gamma(3-\gamma)} - \frac{t^{1-\gamma}}{\Gamma(2-\gamma)} + t^2 - t, \quad 0 < \gamma \leq 1, \quad t \in [0, T], \tag{31} $$

with initial condition $u(0) = 0$ *and exact solution* $u(t) = t^2 - t$.

Table 2 presents the absolute error between the approximate and exact solutions using the non-overlapping multi-domain pseudospectral method with the CGL quadrature and the Chebyshev polynomials as the basis functions. The table also includes numerical results obtained without decomposing the domain. This comparative analysis highlights the accuracy and effectiveness of the multi-domain approach in reducing errors and improving the accuracy of the numerical solution. From the table, it can be seen that, for the fractional order of $\gamma = 0.5$, although the maximum absolute error in the multi-domain setup does not decrease in terms of significant zeros, it reduces as q increases. This indicates that, although minimal, there is an advantage in using more sub-domains. Comparing the error to the single-domain result, there is an additional significant zero in the error of the multi-domain approach, which indicates that the multi-domain approach is more effective for this problem.

Table 2. Maximum absolute error for Example 1 in single-domain and multi-domain setups.

	$N = 10,\ \gamma = 0.5$			$N = 10,\ T = 6.0$		
q	$T = 1.0$	$T = 3.0$	$T = 6.0$	$\gamma = 0.2$	$\gamma = 0.6$	$\gamma = 0.8$
5	7.7426×10^{-5}	6.2253×10^{-4}	2.7903×10^{-3}	1.3625×10^{-3}	3.6161×10^{-3}	5.2468×10^{-3}
10	4.0800×10^{-5}	5.2107×10^{-4}	2.2192×10^{-3}	6.7562×10^{-4}	1.9440×10^{-3}	2.8829×10^{-3}
20	2.3538×10^{-5}	4.4940×10^{-4}	1.8850×10^{-3}	4.4076×10^{-4}	1.5245×10^{-3}	2.1135×10^{-3}
50	2.0489×10^{-5}	4.1298×10^{-4}	1.6596×10^{-3}	3.8984×10^{-4}	1.2966×10^{-3}	1.5794×10^{-3}
100	2.0286×10^{-5}	4.0279×10^{-4}	1.6132×10^{-3}	3.8131×10^{-4}	1.2746×10^{-3}	1.4322×10^{-3}
200	2.0374×10^{-5}	4.0190×10^{-4}	1.6047×10^{-3}	3.8224×10^{-4}	1.2946×10^{-3}	1.3898×10^{-3}
			Single domain			
	3.3518×10^{-4}	1.0362×10^{-3}	2.1415×10^{-3}	9.0794×10^{-4}	2.5580×10^{-3}	2.9839×10^{-3}

From Table 2, the effectiveness of the proposed numerical technique in obtaining an accurate approximate solution to a linear FDE is evident, so in the examples that follow, we consider non-linear FDEs to further validate the numerical method.

Example 2. *Consider the non-linear fractional differential equation [60]*

$$ {}_0^C D_t^\gamma u(t) - u^2(t) = -\frac{2}{(t+1)^2}, \quad 0 < \gamma \leq 1, \quad t \in [0, T], \tag{32} $$

where the initial condition is $u(0) = -2$ *and the exact solution when* $\gamma = 1$ *is given by* $u(t) = -\dfrac{2}{(t+1)}$.

Table 3 shows the approximate solution obtained using two different sub-domains in the proposed method. The solutions in these sub-domains match the numerical values reported in Odibat and Momani [60]. It should be noted that closed-form solutions for $\gamma = 0.2$ and $\gamma = 0.4$ are not available. However, for $\gamma = 1$, the accuracy of the proposed method is unquestionable. To further give credence to the accuracy of the proposed method, Table 4 demonstrates that the residual error is consistently close to zero for various values of γ. This suggests that the proposed method effectively resolves the problem in Example 2.

Table 3. Approximate solution of Example 2 using two different sub-domains, q, and the numerical values reported in Odibat and Momani [60].

t	$\gamma = 0.2$			$\gamma = 0.4$			$\gamma = 1.0$		
	$q = 10$	$q = 20$	Reported [60]	$q = 10$	$q = 20$	Reported [60]	$q = 10$	$q = 20$	Exact
0.0	−2.0000	−2.0000	−2.0000	−2.0000	−2.0000	−2.0000	−2.0000	−2.0000	−2.0000
0.2	−1.4454	−1.4455	−1.4864	−1.4874	−1.4874	−1.4899	−1.6666	−1.6666	−1.6666
0.4	−1.3295	−1.3296	−1.3665	−1.3506	−1.3506	−1.3525	−1.4285	−1.4285	−1.4285
0.6	−1.2477	−1.2478	−1.2822	−1.2525	−1.2526	−1.2541	−1.2500	−1.2500	−1.2500
0.8	−1.1867	−1.1868	−1.2192	−1.1774	−1.1774	−1.1787	−1.1111	−1.1111	−1.1111
1.0	−1.1396	−1.1397	−1.1704	−1.1177	−1.1178	−1.1189	−0.9999	−1.0000	−1.0000

Table 4. The residual errors of Example 2 at selected points using two different sub-domains, q, for different values of fractional order γ.

t	$\gamma = 0.2$		$\gamma = 0.4$		$\gamma = 0.6$	
	$q = 10$	$q = 20$	$q = 10$	$q = 20$	$q = 10$	$q = 20$
0.2	4.4230×10^{-9}	4.8561×10^{-13}	9.7841×10^{-11}	9.2539×10^{-12}	3.2469×10^{-10}	5.2865×10^{-11}
0.4	9.3414×10^{-13}	2.1760×10^{-13}	1.6637×10^{-11}	3.8897×10^{-12}	9.3717×10^{-11}	2.0417×10^{-11}
0.6	4.9305×10^{-13}	1.2423×10^{-13}	8.9744×10^{-12}	2.1446×10^{-12}	4.9785×10^{-11}	1.1396×10^{-11}
0.8	2.8477×10^{-13}	7.1831×10^{-14}	5.4355×10^{-12}	1.3228×10^{-12}	3.0330×10^{-11}	7.1187×10^{-12}
1.0	1.7807×10^{-13}	4.6629×10^{-14}	3.5078×10^{-12}	8.6397×10^{-13}	1.9900×10^{-11}	4.6949×10^{-12}

Example 3. *Next, consider the non-linear fractional differential equation [61]*

$$ {}_0^C D_t^\gamma u(t) + u(t)^2 = \frac{2t^{2-\gamma}}{\Gamma(3-\gamma)} - \frac{t^{1-\gamma}}{\Gamma(2-\gamma)} + (t^2 - t + 1)^2, \quad 0 < \gamma \leq 1, \quad t \in [0, T], \quad (33) $$

with the initial condition $u(0) = 1$ *and whose exact solution is* $u(t) = t^2 - t + 1$.

Table 5 presents the maximum absolute error of the difference between the exact and approximate solutions obtained using the multi-domain and single-domain methods. It is observed from the table that there is conformity with the trend observed in Table 2. As the domain of the FDE becomes larger, the error in the single domain scheme deteriorates; however, in comparison, the magnitude of the errors is better for the multi-domain scheme. We note that the magnitude of the error increases as γ increases, similar to the trend observed in Table 2.

Table 5. The maximum absolute error of the difference between the exact and approximate solutions of Example 3 for different numbers of sub-domains, q.

	$N = 10,\ \gamma = 0.5$			$N = 10,\ T = 6.0$		
q	$T = 1.0$	$T = 3.0$	$T = 6.0$	$\gamma = 0.2$	$\gamma = 0.6$	$\gamma = 0.8$
5	6.5833×10^{-5}	4.5300×10^{-4}	7.7010×10^{-4}	1.6897×10^{-4}	9.9126×10^{-4}	1.0256×10^{-3}
10	3.6216×10^{-5}	2.6020×10^{-4}	4.7125×10^{-4}	1.0248×10^{-4}	6.3577×10^{-4}	7.4647×10^{-4}
20	1.9347×10^{-5}	1.3492×10^{-4}	2.8073×10^{-4}	5.5876×10^{-5}	3.9608×10^{-4}	5.1975×10^{-4}
50	1.6865×10^{-5}	8.2634×10^{-5}	1.3621×10^{-4}	2.3143×10^{-5}	2.0636×10^{-4}	3.2248×10^{-4}
100	1.6547×10^{-5}	6.7163×10^{-5}	8.2558×10^{-5}	1.0361×10^{-5}	1.1932×10^{-4}	2.1211×10^{-4}
200	1.6576×10^{-5}	5.9548×10^{-5}	6.7326×10^{-5}	7.5375×10^{-6}	7.6523×10^{-5}	1.3480×10^{-4}
			Single domain			
	3.1847×10^{-4}	9.0415×10^{-4}	1.7639×10^{-3}	6.1773×10^{-4}	1.9684×10^{-3}	2.2669×10^{-3}

Example 4. *Consider the initial value fractional homogeneous Bagley–Torvik differential equation*

$$ {}_0^C D_t^{\gamma} u(t) + u''(t) + u(t) = 0, \quad \text{for} \quad t \in (0, T] \quad \text{and} \quad \gamma \in (1, 2), \tag{34} $$

subject to the initial conditions $u(0) = 1.0$ *and* $u'(0) = 0$.

If we replace the right-hand side of Equation (34) with $2 + 4\sqrt{\frac{t}{\pi}} + t^2$ and set $t \in (0, 1]$, subject to the starting conditions $u(0) = 0.0$ and $u'(0) = 0$, the closed-form solution is given as $u(t) = t^2$ when $\gamma = 1.5$. Table 6 presents the errors obtained by taking the absolute value of the difference between the numerical solutions obtained through the multi-domain technique and the exact solutions, and these are compared with the errors obtained through the evolutionary artificial neural networks method in the study of Raja et al. [62].

Table 6. The absolute errors of the difference between the exact and numerical solutions of Example 4 at different t values using $N = 12$ and $q = 400$ in comparison with the errors reported in the work of Raja et al. [62].

t	Exact Solution	Present Result	Present Error	Reported Error [62]
0.1	0.01	0.01000	1.83183×10^{-7}	5.48×10^{-5}
0.3	0.09	0.09005	1.26966×10^{-6}	2.66×10^{-3}
0.5	0.25	0.25014	6.69388×10^{-6}	1.67×10^{-2}
0.7	0.49	0.49026	3.40281×10^{-5}	5.80×10^{-2}
0.9	0.81	0.81040	1.25968×10^{-4}	1.50×10^{-1}
1.0	1.00	1.00048	2.12565×10^{-4}	2.25×10^{-1}

Table 6 compares the exact and approximate solutions obtained using the multi-domain method at various t values and the associated errors at each t. The errors are compared with those obtained in Raja et al. [62]. It is observed from the table that the approximate solutions match the exact solutions better than the results in Raja et al. [62], and this validates the efficiency of the proposed method as it outperforms a neural network approach known to be efficient and reliable.

To further validate the accuracy of the proposed multi-domain method, the infinity norm of the residual for different numbers of sub-domains is presented. Table 7 gives the infinity norm of the residual errors for Example 4. The residual error is obtained through the discrete equivalent of Equation (34), defined by

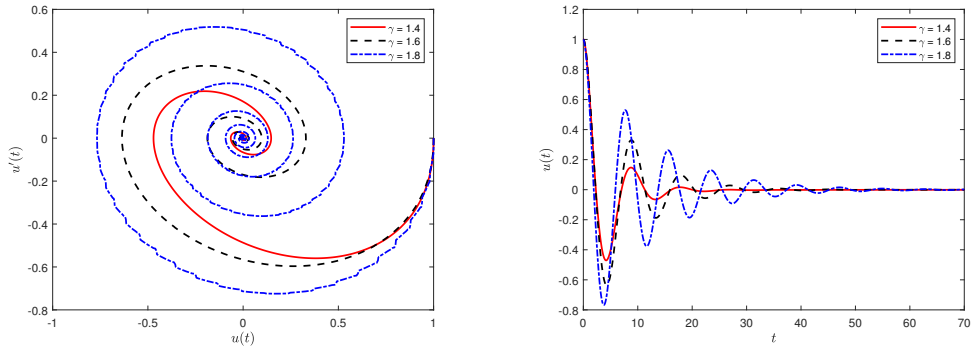
$$ \text{Residual} = \sum_{e=1}^{l} ({}_{t_{e-1}, t_e}\mathbf{D}_e^{\gamma})^{\mathsf{T}} \mathbf{U}_{h_l, e} + ({}_{t_l, h_l}\mathbf{D}_l^{\gamma})^{\mathsf{T}} \mathbf{U}_{h_l, l} + ({}_{t_l, h_l}\mathbf{D}_l^{2})^{\mathsf{T}} \mathbf{U}_{h_l, l} + \mathbf{E}_l^{\mathsf{T}} \mathbf{U}_{h_l, l}. \tag{35} $$

Table 7 shows that the method is efficient in solving fractional-order differential equations of varying values of arbitrary non-integer order, as the residual error remains very small for various fractional orders, demonstrating the effectiveness of the multi-domain technique.

Table 7. The maximum residual errors for Example 4 for different T values, fractional orders, and different numbers of multi-domains, q.

	$N = 10,\ \gamma = 1.5$			$N = 10,\ T = 6.0$		
q	$T = 1.0$	$T = 3.0$	$T = 6.0$	$\gamma = 1.2$	$\gamma = 1.6$	$\gamma = 1.8$
10	1.115×10^{-10}	1.158×10^{-11}	1.933×10^{-12}	1.031×10^{-12}	1.870×10^{-12}	3.445×10^{-12}
20	7.346×10^{-10}	6.618×10^{-11}	1.158×10^{-11}	1.078×10^{-11}	2.104×10^{-11}	2.765×10^{-11}

To further validate the results obtained, we reproduce the solution profiles obtained by Zafar et al. [63] and observe excellent agreement in the result, as shown in Figure 3.



(**a**) (**b**)

Figure 3. The dynamics of the Bagley–Torvik Equation (34) for various values of γ with $T = 70$, $N = 10$ and $q = 400$: (**a**) the phase portrait for the $u - u'$ plane and (**b**) the $u(t)$ profile.

In the remaining examples in this section, we consider systems of coupled fractional initial value problems that describe chaotic systems. These equations often exhibit long-term dynamics and, invariably, require long-term integration to fully understand the underlying dynamics involved. Therefore, a numerical technique, such as the multi-domain Chebyshev pseudospectral method, that can handle long-term integration is a perfect fit for solving such systems, as corroborated by the results in the next four examples.

Example 5. *Consider the fractional chaotic system of differential equations which describe a memory-inherent electrical system [64]*

$$\left.\begin{aligned} {}_0^C D_t^\gamma u_1(t) &= \frac{u_2(t)}{c}, \\ {}_0^C D_t^\gamma u_2(t) &= -\frac{1}{L}\Big(u_1(t) + \beta u_2(t)(u_3(t))^2 - \beta u_2(t)\Big), \\ {}_0^C D_t^\gamma u_3(t) &= -u_2(t) - \alpha u_3(t) + u_2(t)u_3(t), \qquad t \in (0, T], \quad \gamma \in (0, 1], \end{aligned}\right\} \tag{36}$$

subject to the following initial state, $(0.1, 0.0, 0.1)$. *Here,* $c = 0.70$, $L = 3.00$, $\beta = 1.10$ *and* $\alpha = 0.60$. *The voltage across the capacitor, c, is represented by the variable* $u_1(t)$, *and the current through the inductor, L, is represented by the variable* $u_2(t)$, *and the internal state of the memristive system is denoted by the variable* $u_3(t)$.

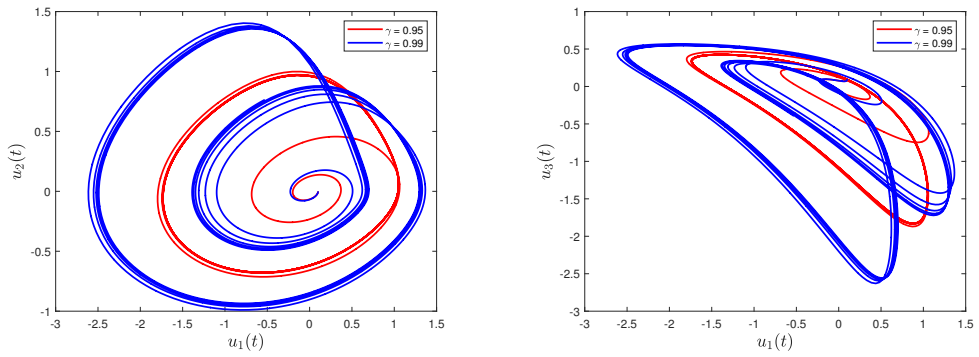
The average maximum residual error for the system (36) is given in Table 8. We remark here that the maximum residual errors presented in Table 8 and subsequent tables in this

section are obtained as the average of the maximum residuals of the equations in the system. The residuals are calculated similarly to Equation (35). It can be seen from Table 8 that for various fractional orders and seemingly large domains, $[0, T]$, a high level of accuracy is obtained when the multi-domain Chebyshev pseudospectral method is used. Due to the inherently chaotic nature of the problem, the dynamics of the system typically require a large physical domain to fully develop, and to effectively capture these complex long-term behaviors, the physical domain is partitioned into a large number of smaller sub-domains. The table shows that the average of the maximum residual errors decreases marginally as the number of sub-domains increases. This inverse relationship between the number of sub-domains and the residual error highlights the applicability of the multi-domain technique when dealing with this kind of system that requires long-term integration. This is particularly evident when the error is compared to the error associated with the single-domain method. As already observed in the previous examples and now corroborated by Table 8, the error associated with the single-domain method is worse than the error obtained when the multi-domain approach is used.

Table 8. The residual errors for Example 5 for two values of T using $N = 10$ and for different numbers of sub-domains q, with $c = 0.70$, $L = 3.00$, $\beta = 1.10$ and $\alpha = 0.60$.

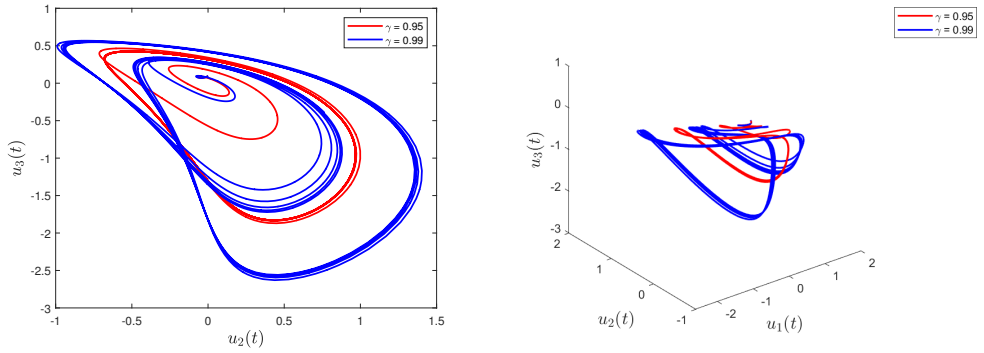
	$T = 20.0$			$T = 50.0$		
q	$\gamma = 0.7$	$\gamma = 0.8$	$\gamma = 0.9$	$\gamma = 0.7$	$\gamma = 0.8$	$\gamma = 0.9$
100	2.515×10^{-11}	7.810×10^{-12}	8.503×10^{-12}	8.781×10^{-11}	3.336×10^{-11}	8.054×10^{-10}
200	9.780×10^{-12}	2.590×10^{-12}	2.144×10^{-12}	3.434×10^{-11}	1.113×10^{-11}	1.935×10^{-10}
300	5.544×10^{-12}	1.365×10^{-12}	9.795×10^{-13}	1.973×10^{-11}	5.817×10^{-12}	8.468×10^{-11}
400	3.717×10^{-12}	8.632×10^{-13}	5.718×10^{-13}	1.319×10^{-11}	3.701×10^{-12}	4.723×10^{-11}
500	2.720×10^{-12}	6.125×10^{-13}	4.095×10^{-13}	9.769×10^{-12}	2.592×10^{-12}	3.004×10^{-11}
			Single domain			
	2.397×10^{-9}	2.439×10^{-9}	1.920×10^{-7}	3.101×10^{-8}	4.108×10^{-7}	5.427×10^{-7}

To lend further credibility to the numerical results of this example, we reproduce the dynamics, in the form of phase portraits, of the memristive chaotic electrical system as reported in the work of Sene [64]. As shown in Figure 4, there is an excellent agreement between these phase portraits and those in the work of Sene [64].



(**a**) (**b**)

Figure 4. *Cont.*



(**c**) (**d**)

Figure 4. The phase portraits of the chaotic systems (36) for the (**a**) $u_1 - u_2$ plane, (**b**) $u_1 - u_3$ plane, (**c**) $u_2 - u_3$ plane and (**d**) $u_1 - u_2 - u_3$ surface with $\gamma = 0.95$ and 0.99, $T = 250$, $N = 13$ and $q = 850$.

Example 6. *Consider the fractional modified stretch-twist-fold differential equations [65]*

$$\left.\begin{aligned} {}_0^C D_t^\gamma u_1(t) &= \Omega_1 u_3(t) - 8u_1(t)u_2(t), \quad t \in (0,T], \quad \gamma \in (0,1], \\ {}_0^C D_t^\gamma u_2(t) &= 11(u_1(t))^2 + 3(u_2(t))^2 + (u_3(t))^2 + \Omega_2 u_1(t)u_3(t) - 3\Omega_3^2, \\ {}_0^C D_t^\gamma u_3(t) &= 2\big(1 - \Omega_4\big)u_2(t)u_3(t) - \Omega_1 u_1(t) - \Omega_2 u_1(t)u_2(t), \end{aligned}\right\} \tag{37}$$

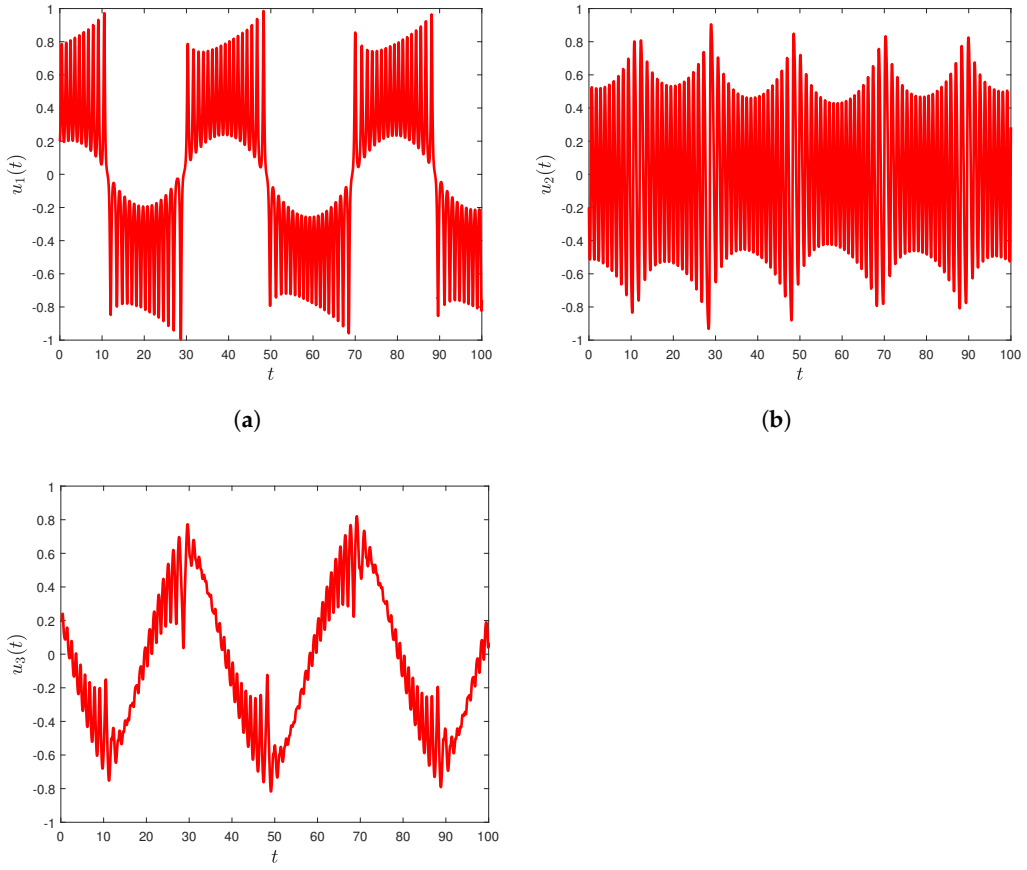
under the initial state $(0.2, -0.2, 0.2)$, *with* $\gamma = 1.00$, $\Omega_1 = 0.2$, $\Omega_2 = 2.0$, $\Omega_3 = 1.0$ *and* $\Omega_4 = 0.1$.

Similar to the previous example, this system typically requires a large physical domain to fully stabilize and develop, and when solved on a single domain, the numerical solution is unstable, as evidenced by the average residual errors in a single domain presented in Table 9. We present the average maximum residual errors for various fractional orders and relatively large domains, $[0, T]$, for different relatively large numbers of sub-domains in Table 9. There is an agreement between the level of accuracy observed for the results in Table 9 and those obtained in the previous examples. The error obtained for the various sub-domain numbers used is shown to be small, indicating that the proposed multi-domain Chebyshev pseudospectral method efficiently solves this system of FDEs. As q increases, it is observed that the average residual error improves, resulting in higher accuracy. The superior effectiveness of the multi-domain approach over the single-domain numerical scheme becomes marginally clearer in this example. In this case, the single-domain numerical scheme is observed to produce quite large numerical errors, whereas the multi-domain technique can achieve a higher accuracy level. This disparity in the level of performance highlights the limitation in the scope of the single-domain scheme, while the multi-domain approach demonstrates robustness and versatility, especially for complex systems like the one considered in this example.

Figure 5 gives the behavior of the solution profiles as t evolves. The results are as expected and in complete agreement with the result of Azam et al. [65]. It is also observed in Figure 6 that the chaotic system flows through the various states by turning and twisting about the quadrants. This dynamics was also captured by Azam et al. [65] and validates the numerical results being put forward in this study.

Table 9. The average maximum residual errors for Example 6 for two values of T and different numbers of sub-domains q using $N = 10$. The following parameter values are used: $\Omega_1 = 0.2$, $\Omega_2 = 2.0$, $\Omega_3 = 1.0$ and $\Omega_4 = 0.1$.

	$T = 20.0$			$T = 50.0$		
q	$\gamma = 0.7$	$\gamma = 0.8$	$\gamma = 0.9$	$\gamma = 0.7$	$\gamma = 0.8$	$\gamma = 0.9$
100	1.292×10^{-8}	4.793×10^{-8}	2.171×10^{-8}	1.364×10^{-10}	1.059×10^{-8}	4.952×10^{-8}
200	1.883×10^{-8}	1.192×10^{-8}	6.393×10^{-9}	1.060×10^{-8}	2.928×10^{-8}	4.607×10^{-8}
300	8.108×10^{-9}	5.539×10^{-9}	3.267×10^{-9}	1.826×10^{-8}	3.502×10^{-8}	1.144×10^{-8}
400	4.546×10^{-9}	3.223×10^{-9}	1.790×10^{-9}	2.750×10^{-8}	1.284×10^{-8}	1.065×10^{-8}
500	3.082×10^{-9}	2.098×10^{-9}	1.182×10^{-9}	1.883×10^{-8}	1.194×10^{-8}	6.393×10^{-9}
			Single domain			
	1.730×10^{1}	1.999×10^{1}	2.419×10^{1}	8.388×10^{1}	2.195×10^{-1}	4.131×10^{1}



(**a**) (**b**)

(**c**)

Figure 5. The time evolution of $u_1(t)$, $u_2(t)$, $u_3(t)$ of the chaotic system (37) with $\gamma = 1.00$, $\Omega_1 = 0.2$, $T = 100$, $N = 13$ and $q = 350$.

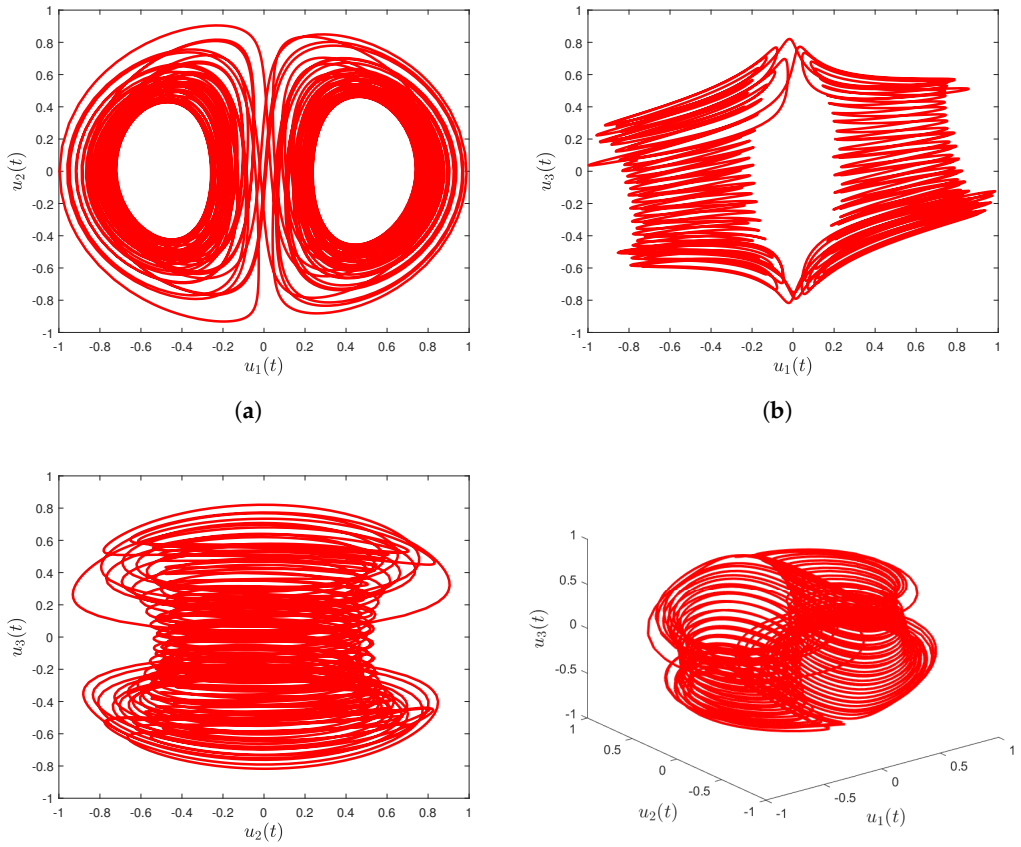


Figure 6. The phase portraits of the chaotic system (37) for the (**a**) $u_1 - u_2$ plane, (**b**) $u_1 - u_3$ plane, (**c**) $u_2 - u_3$ plane and (**d**) $u_1 - u_2 - u_3$ surface with $\gamma = 1.00$, $\Omega_1 = 0.2$, $T = 100$, $N = 13$ and $q = 350$.

Example 7. *Consider the fractional Rossler differential equations [66]*

$$\left.\begin{aligned} {}_0^C D_t^\gamma u_1(t) &= -u_2(t) - u_3(t), \quad t \in (0, T], \quad \gamma \in (0, 1], \\ {}_0^C D_t^\gamma u_2(t) &= u_1(t) + \Psi_1 u_2(t), \\ {}_0^C D_t^\gamma u_3(t) &= 0.2 + u_1(t)u_3(t) - 10u_3(t), \end{aligned}\right\} \tag{38}$$

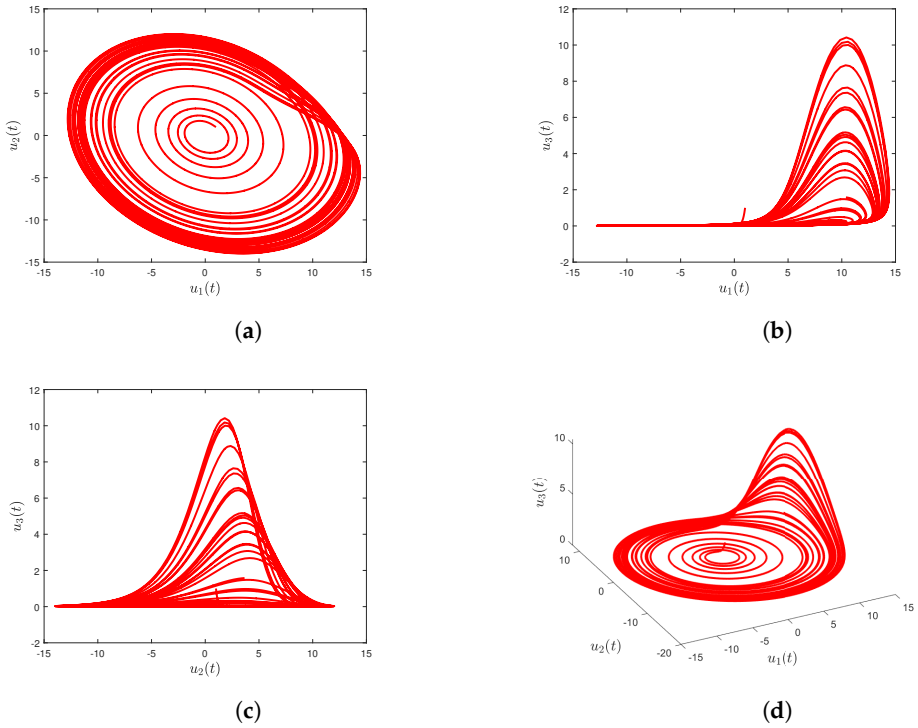
with the following initial conditions: $(1, 1, 1)$, *with* $\Psi_1 = 0.4$ *and* $\gamma = 0.9$.

Table 10 shows the average maximum residual errors for Equation (38) for selected values of arbitrary non-integer order, γ, and for two domains, $(0, T]$. The results are presented for different numbers of domain decomposition. From Table 10, we see that the residual errors are small, indicating that the approximate solutions are accurate. As has been validated several times in this study, an increase in the number of sub-domains used results in more accurate solutions. It is also shown that using the single-domain approach results in less accuracy in the approximate solutions obtained.

Table 10. The residual errors for Example 7 at two final times T using the number of grid points $N = 10$ for different numbers of sub-domains q and $\Psi_1 = 0.40$.

	$T = 20.0$			$T = 50.0$		
q	$\gamma = 0.7$	$\gamma = 0.8$	$\gamma = 0.9$	$\gamma = 0.7$	$\gamma = 0.8$	$\gamma = 0.9$
100	1.073×10^{-7}	4.973×10^{-8}	1.171×10^{-8}	1.469×10^{-7}	1.469×10^{-7}	1.893×10^{-7}
200	5.544×10^{-8}	1.857×10^{-8}	3.569×10^{-9}	1.269×10^{-7}	6.666×10^{-8}	4.696×10^{-8}
300	3.479×10^{-8}	1.015×10^{-8}	1.740×10^{-9}	9.241×10^{-8}	3.853×10^{-8}	3.299×10^{-8}
400	2.459×10^{-8}	6.520×10^{-9}	1.042×10^{-9}	6.979×10^{-8}	2.574×10^{-8}	1.933×10^{-8}
500	1.859×10^{-8}	4.653×10^{-9}	7.023×10^{-10}	5.544×10^{-8}	1.857×10^{-8}	1.189×10^{-8}
			Single domain			
	1.594×10^{-4}	2.218×10^{-4}	6.733×10^{-5}	1.725×10^{-4}	2.518×10^{-4}	3.726×10^{-4}

Figure 7 displays the phase portraits for the chaotic system, and the result agrees with the results obtained in Li et al. [66], thereby validating the present numerical solutions.



(**a**) (**b**) (**c**) (**d**)

Figure 7. The phase portraits of the chaotic systems (38) for the (**a**) u_1-u_2 plane, (**b**) $u_1 - u_3$ plane, (**c**) $u_2 - u_3$ plane and (**d**) $u_1 - u_2 - u_3$ surface with $\gamma = 0.90$, $\Psi_1 = 0.4$, $T = 200$, $N = 10$ and $q = 900$.

Example 8. *Consider the following non-linear fractional chaotic system [67]*

$$\left.\begin{aligned} {}_0^C D_t^{\gamma} u_1(t) &= \Psi_1 u_1(t) + u_2(t) + u_2(t)u_3(t), \quad t \in (0, T], \quad \gamma \in (0, 1], \\ {}_0^C D_t^{\gamma} u_2(t) &= -u_1(t)u_3(t) + u_2(t)u_3(t), \\ {}_0^C D_t^{\gamma} u_3(t) &= -\Psi_2 u_3(t) - \Psi_3 u_1(t)u_2(t) - m_1, \end{aligned}\right\} \tag{39}$$

with the initial state given as $(0.2, 0.2, 0.2)$, $\Psi_1 = \Psi_2 = \Psi_3 = 1.0$, $m_1 = 0.0$ *and* $\gamma = 0.9$.

In Table 11, the average maximum residual errors obtained for Equation (39) are presented. The table shows that an increase in the number of partitioned domains results in higher accuracy of approximate solutions obtained for the selected fractional orders and domains. The average maximum residual errors are observed to be small, confirming that

the proposed multi-domain method is highly efficient for this fractional differential model, unlike the single-domain approach, which is not as accurate.

Table 11. The average maximum residual errors for Example 8 in two different domains $(0, T]$ with $N = 10$ for different numbers of sub-domains q, $\Psi_1 = \Psi_2 = \Psi_3 = 1.0$ and $m_1 = 0.0$.

	$T = 20.0$			$T = 50.0$		
q	$\gamma = 0.7$	$\gamma = 0.8$	$\gamma = 0.9$	$\gamma = 0.7$	$\gamma = 0.8$	$\gamma = 0.9$
100	4.114×10^{-8}	6.012×10^{-8}	5.802×10^{-8}	2.142×10^{-7}	1.226×10^{-7}	2.711×10^{-7}
200	1.305×10^{-8}	1.673×10^{-8}	1.559×10^{-8}	5.335×10^{-8}	6.845×10^{-8}	7.236×10^{-8}
300	6.155×10^{-9}	7.361×10^{-9}	7.034×10^{-9}	3.422×10^{-8}	4.357×10^{-8}	4.049×10^{-8}
400	3.398×10^{-9}	4.196×10^{-9}	3.935×10^{-9}	2.050×10^{-8}	2.575×10^{-8}	2.436×10^{-8}
500	2.195×10^{-9}	2.720×10^{-9}	2.540×10^{-9}	1.304×10^{-8}	1.676×10^{-8}	1.559×10^{-8}
			Single domain			
	3.055×10^{-2}	1.225×10^{-2}	9.667×10^{-1}	2.279×10^{-2}	6.453×10^{1}	4.708×10^{1}

We present the phase portraits of the chaotic system in Figure 8. The figure validates the dynamics reported by Sene [67] and is in excellent agreement with the previously published result. The agreement between the phase portraits presented here and those documented in the study by Sene [67] serves to further corroborate the accuracy and reliability of the current numerical method.

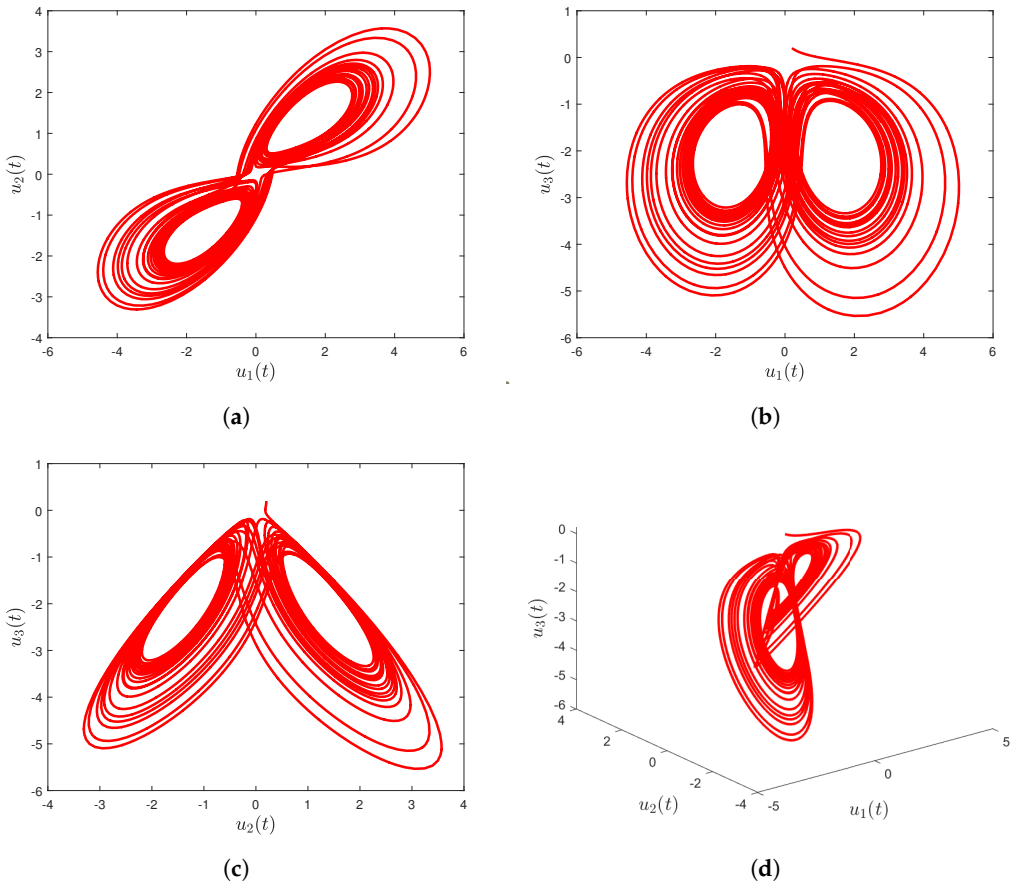


Figure 8. The phase portraits of the chaotic systems (39) for the (**a**) $u_1 - u_2$ plane, (**b**) $u_1 - u_3$ plane, (**c**) $u_2 - u_3$ plane and (**d**) $u_1 - u_2 - u_3$ surface with $\gamma = 0.95$, $T = 150$, $N = 10$ and $q = 800$.

Figure 9 shows the phase portraits of the chaotic system in Equation (39) using two different γ values. The figure justifies that the dynamics of the chaotic system depend on the arbitrary real order.

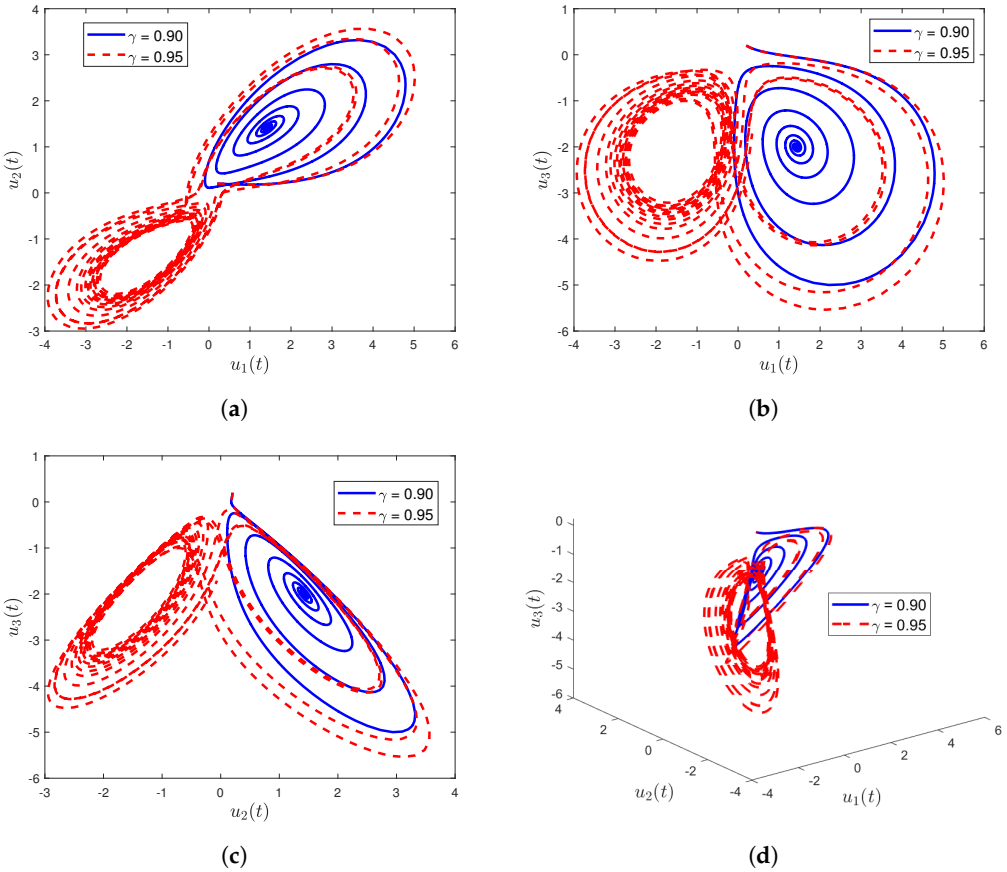


Figure 9. The phase portraits of the chaotic systems (39) for the (**a**) $u_1 - u_2$ plane, (**b**) $u_1 - u_3$ plane, (**c**) $u_2 - u_3$ plane and (**d**) $u_1 - u_2 - u_3$ surface with $\gamma = 0.90$ and 0.95, $T = 80$, $N = 10$ and $q = 350$.

Figure 10 displays the accuracy of the numerical results (the maximum absolute residual in each sub-domain) of each equation in the chaotic systems presented in Equations (36)–(39). The figures show that the accuracy of the solutions obtained through the multi-domain Chebyshev pseudospectral method is high and sometimes reaches errors smaller than those reported so far in the tables. As can be clearly observed, as the number of sub-domains in the computational domain increases, the residual error decreases correspondingly. In some instances, the residual error is reduced to the order of 10^{-16}. However, the residual errors with a small number of sub-domains already show good accuracy, thereby showing that using the multi-domain approach proposed in this study is very efficient in solving fractional differential equations that typically require long-term integration.

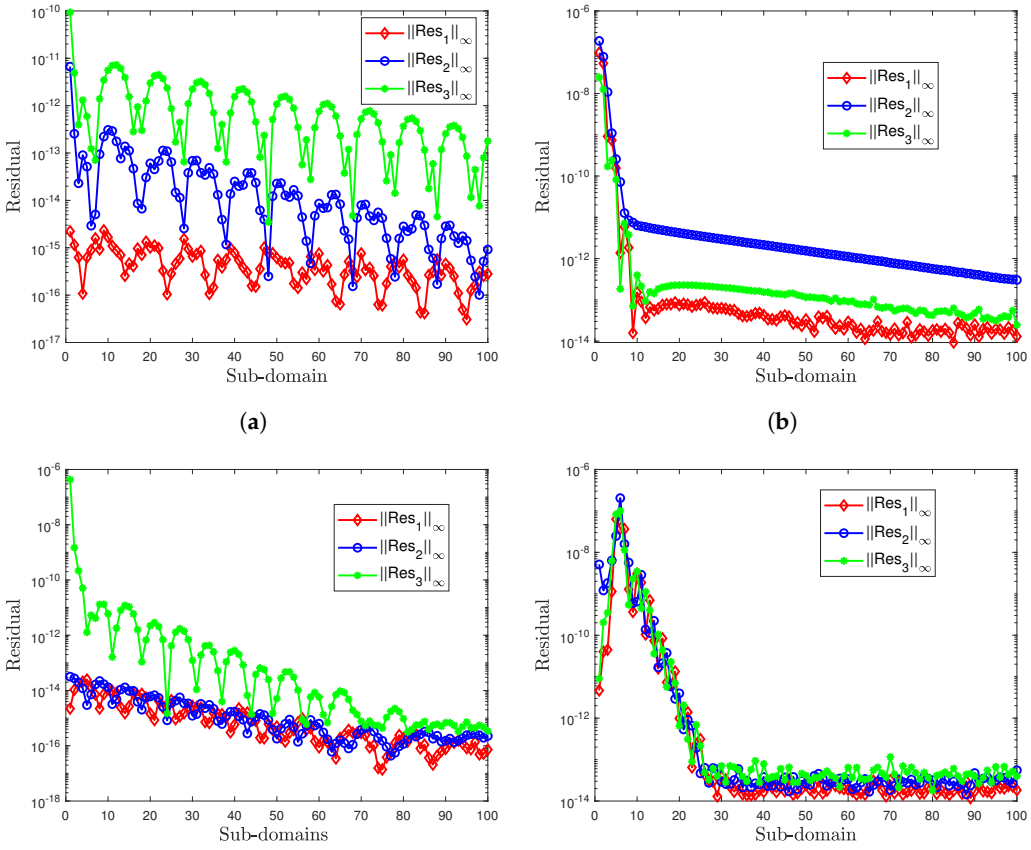


Figure 10. The maximum residual error in each domain for (**a**) Equation (36) (**b**) Equation (37) (**c**) Equation (38) and (**d**) Equation (39) using $\gamma = 0.8,\ T = 50,\ N = 10$ and $q = 100$.

5. Conclusions

This study introduced a novel approach, the multi-domain Chebyshev pseudospectral method for solving fractional differential equations. The novelty of the method stemmed from the use of non-overlapping multi-domains, Chebyshev polynomials of the first kind and Gauss–Lobatto collocation points in developing the scheme, which was specifically adapted to handle fractional differential equations. The method leverages the global properties of spectral-based methods for discretizing non-local operators such as fractional differential operators, the exponential convergence of the Chebyshev polynomial-based pseudospectral method and the multi-domain approach's ability to maintain the accuracy of solutions to problems with long-term dynamics. The stability region of the multi-domain scheme was established through linear stability analysis. The stability analysis showed that obtaining an absolute and Lubich-stable numerical scheme for some values of arbitrary non-integer order is possible. Several numerical experiments involving linear and non-linear single FDEs and systems of non-linear FDEs were conducted. The examples were carefully chosen to span varying complexity levels, including fractional initial value problems with long-term dynamics, such as chaotic systems, which were shown to significantly benefit from a multi-domain numerical scheme. These examples have established that the proposed multi-domain numerical method gives highly accurate results. To validate the method, comparisons of the numerical results obtained in this study were made with analytical solutions, where available, and approximate solutions from the literature. It was found that the proposed method gave more accurate solutions than the approximate solutions from the literature. Where an exact solution was unavailable, residual error analysis was conducted to establish the accuracy of the proposed method. In all the examples considered, the method was found to produce comparative, highly accurate solutions. In

future studies, we will explore the applicability of the Chebyshev pseudospectral method in non-overlapping or overlapping partitioned domains to fractional boundary value problems with non-linear dynamics.

Author Contributions: S.D.O.: Conceptualization, Investigation, Methodology, Software, Writing—Original Draft, Reviewing and Editing, Supervision. N.M.: Investigation, Methodology, Writing—Original Draft, Reviewing and Editing. Y.O.T.: Investigation, Methodology, Writing—Original Draft, Reviewing and Editing. O.O.: Conceptualization, Methodology, Writing—Original Draft, Reviewing and Editing, Supervision. All authors have read and agreed to the published version of the manuscript.

Funding: This work is based on the research supported in part by the National Research Foundation of South Africa (Ref Number TTK2204163593). The APC was funded by the University of Witwatersrand.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data generated are contained within the article.

Acknowledgments: Yusuf Olatunji Tijani acknowledges the support of Rhodes University.

Conflicts of Interest: The authors declare that there are no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CGL	Chebyshev–Gauss–Lobatto
FDEs	Fractional differential equations
MSCM	Multi-domain spectral collocation method

References

1. Podlubny, I. *Fractional Differential Equations: An Introduction to Fractional Derivatives, Fractional Differential Equations, to Methods of Their Solution and Some of Their Applications*; Elsevier: Amsterdam, The Netherlands, 1998.
2. Oldham, K.; Spanier, J. *The Fractional Calculus Theory and Applications of Differentiation and Integration to Arbitrary Order*; Elsevier: Amsterdam, The Netherlands, 1974.
3. Caputo, M.; Fabrizio, M. A new definition of fractional derivative without singular kernel. *Prog. Fract. Differ. Appl.* **2015**, *1*, 73–85.
4. Atangana, A.; Baleanu, D. New fractional derivatives with nonlocal and non-singular kernel: Theory and application to heat transfer model. *arXiv* **2016**, arXiv:1602.03408.
5. Javadi, R.; Mesgarani, H.; Nikan, O.; Avazzadeh, Z. Solving fractional order differential equations by using fractional radial basis function neural network. *Symmetry* **2023**, *15*, 1275. [CrossRef]
6. Benchochra, M.; Karapinar, E.; Lazreg, J.E.; Salim, A. *Fractional Differential Equations: New Advancements for Generalized Fractional Derivatives*; Springer: Cham, Switzerland, 2023.
7. Ahmadian, S.; Darvishi, M. Fractional version of (1+1)-dimensional Biswas–Milovic equation and its solutions. *Optik* **2016**, *127*, 10135–10147. [CrossRef]
8. Manafian, J.; Lakestani, M. A new analytical approach to solve some of the fractional-order partial differential equations. *Indian J. Phys.* **2017**, *91*, 243–258. [CrossRef]
9. Kumar, S.; Kumar, D.; Abbasbandy, S.; Rashidi, M. Analytical solution of fractional Navier–Stokes equation by using modified Laplace decomposition method. *Ain Shams Eng. J.* **2014**, *5*, 569–574. [CrossRef]
10. Shakeel, M.; Mohyud-Din, S.T. A novel (G′/G)-expansion method and its application to the space-time fractional symmetric regularized long wave (SRLW) equation. *Adv. Trends Math.* **2015**, *2*, 1–16. [CrossRef]
11. Zhu, Q.; Qi, J. Exact Solutions of the Nonlinear Space-Time Fractional Partial Differential Symmetric Regularized Long Wave (SRLW) Equation by Employing Two Methods. *Adv. Math. Phys.* **2022**, *2022*, 8062119. [CrossRef]
12. Senol, M. New analytical solutions of fractional symmetric regularized-long-wave equation. *Rev. Mex. De Física* **2020**, *66*, 297–307. [CrossRef]
13. Ünal, S.Ç.; Daşcıoğlu, A.; Bayram, D.V. Jacobi elliptic function solutions of space-time fractional symmetric regularized long wave equation. *Math. Sci. Appl. E-Notes* **2021**, *9*, 53–63. [CrossRef]
14. Tuluce Demiray, S.; Bulut, H.; Belgacem, F.B.M. Sumudu transform method for analytical solutions of fractional type ordinary differential equations. *Math. Probl. Eng.* **2015**, *2015*, 131690. [CrossRef]
15. Kumar, D.; Darvishi, M.; Joardar, A. Modified Kudryashov method and its application to the fractional version of the variety of Boussinesq-like equations in shallow water. *Opt. Quantum Electron.* **2018**, *50*, 128. [CrossRef]

16. He, J.H. Exp-function method for fractional differential equations. *Int. J. Nonlinear Sci. Numer. Simul.* **2013**, *14*, 363–366. [CrossRef]
17. Ahmadian, S.; Darvishi, M. A new fractional Biswas–Milovic model with its periodic soliton solutions. *Optik* **2016**, *127*, 7694–7703. [CrossRef]
18. Turan, D.A. A sine-cosine wavelet method for the approximation solutions of the fractional Bagley-Torvik equation. *Sigma J. Eng. Nat. Sci. Ve Fen Bilim. Derg.* **2022**, *40*, 150–154.
19. Khan, A.; Shah, K.; Luo, D. Numerical Solutions to Some Families of Fractional Order Differential Equations by Laguerre Polynomials. In *Nonlinear Systems-Theoretical Aspects and Recent Applications*; IntechOpen: London, UK, 2020.
20. Sun, Z.Z.; Gao, G.H. *Fractional Differential Equations: Finite Difference Methods*; Walter de Gruyter GmbH & Co KG: Berlin, Germany, 2020.
21. Hattaf, K.; Hajhouji, Z.; Ait Ichou, M.; Yousfi, N. A numerical method for fractional differential equations with new generalized Hattaf fractional derivative. *Math. Probl. Eng.* **2022**, *2022*, 3358071. [CrossRef]
22. Kumar, M.; Jhinga, A.; Daftardar-Gejji, V. Higher order numerical methods for fractional delay differential equations. *Indian J. Pure Appl. Math.* **2024**, 1–22. [CrossRef]
23. Jhinga, A.; Daftardar-Gejji, V. A new numerical method for solving fractional delay differential equations. *Comput. Appl. Math.* **2019**, *38*, 166. [CrossRef]
24. Lee, S.; Kim, H.; Jang, B. A Novel Numerical Method for Solving Nonlinear Fractional-Order Differential Equations and Its Applications. *Fractal Fract.* **2024**, *8*, 65. [CrossRef]
25. Kumar, S.; Shaw, P.K.; Abdel-Aty, A.H.; Mahmoud, E.E. A numerical study on fractional differential equation with population growth model. *Numer. Methods Partial Differ. Equations* **2024**, *40*, e22684. [CrossRef]
26. Bhrawy, A.H.; Taha, T.M.; Machado, J.A.T. A review of operational matrices and spectral techniques for fractional calculus. *Nonlinear Dyn.* **2015**, *81*, 1023–1052. [CrossRef]
27. Baleanu, D.; Bhrawy, A.; Taha, T. A modified generalized Laguerre spectral method for fractional differential equations on the half line. In *Proceedings of the Abstract and Applied Analysis*; Wiley Online Library: Hoboken, NJ, USA, 2013; Volume 2013, p. 413529.
28. Delkhosh, M.; Parand, K. A new computational method based on fractional Lagrange functions to solve multi-term fractional differential equations. *Numer. Algorithms* **2021**, *88*, 729–766. [CrossRef]
29. Xu, Q.; Zheng, Z. Spectral collocation method for fractional differential/integral equations with generalized fractional operator. *Int. J. Differ. Equations* **2019**, *2019*, 3734617. [CrossRef]
30. Raslan, K.R.; Abd El Salam, M.A.; Ali, K.K.; Mohamed, E.M. Spectral Tau method for solving general fractional order differential equations with linear functional argument. *J. Egypt. Math. Soc.* **2019**, *27*, 33. [CrossRef]
31. Shen, J.; Sheng, C. Spectral methods for fractional differential equations using generalized Jacobi functions. *Handb. Fract. Calc. Appl. Numer. Methods* **2019**, *3*, 127–155.
32. Zayernouri, M.; Karniadakis, G.E. Exponentially accurate spectral and spectral element methods for fractional ODEs. *J. Comput. Phys.* **2014**, *257*, 460–480. [CrossRef]
33. Zhao, J.; Zhang, Y.; Xu, Y. Implicit Runge–Kutta and spectral Galerkin methods for Riesz space fractional/distributed-order diffusion equation. *Comput. Appl. Math.* **2020**, *39*, 47. [CrossRef]
34. Khader, M.; Sweilam, N. Approximate solutions for the fractional advection–dispersion equation using Legendre pseudo-spectral method. *Comput. Appl. Math.* **2014**, *33*, 739–750. [CrossRef]
35. Cardone, A.; D'Ambrosio, R.; Paternoster, B. A spectral method for stochastic fractional differential equations. *Appl. Numer. Math.* **2019**, *139*, 115–119. [CrossRef]
36. Doha, E.H.; Bhrawy, A.H.; Ezz-Eldien, S.S. Efficient Chebyshev spectral methods for solving multi-term fractional orders differential equations. *Appl. Math. Model.* **2011**, *35*, 5662–5672. [CrossRef]
37. Sweilam, N.H.; Khader, M.M. A Chebyshev pseudo-spectral method for solving fractional-order integro-differential equations. *ANZIAM J.* **2010**, *51*, 464–475. [CrossRef]
38. Dabiri, A.; Karimi, L. A Chebyshev pseudospectral method for solving fractional-order optimal control problems. In Proceedings of the 2019 American Control Conference, Philadelphia, PA, USA, 10–12 July 2019.
39. Oloniiju, S.D.; Goqo, S.P.; Sibanda, P. A geometrically convergent pseudo-spectral method for multi-dimensional two-sided space fractional partial differential equations. *J. Appl. Anal. Comput.* **2021**, *11*, 1699–1717. [CrossRef] [PubMed]
40. Oloniiju, S.D.; Goqo, S.P.; Sibanda, P. A Chebyshev pseudo-spectral method for the numerical solutions of distributed order fractional ordinary differential equations. *Appl. Math. E-Notes* **2022**, *22*, 132–141.
41. Oloniiju, S.D.; Goqo, S.P.; Sibanda, P. A pseudo-spectral Method for time distributed order two-sided space fractional differential equations. *Taiwan. J. Math.* **2021**, *25*, 959–979. [CrossRef]
42. Adel, M.; Khader, M.M.; Assiri, T.A.; Kallel, W. Numerical simulation for COVID-19 model using a multidomain spectral relaxation technique. *Symmetry* **2023**, *15*, 931. [CrossRef]
43. Mkhatshwa, M.; Khumalo, M.; Dlamini, P. Multi-domain multivariate spectral collocation method for (2+1) dimensional nonlinear partial differential equations. *Partial Differ. Equations Appl. Math.* **2022**, *6*, 100440. [CrossRef]
44. Samuel, F.; Motsa, S. Solving hyperbolic partial differential equations using a highly accurate multidomain bivariate spectral collocation method. *Wave Motion* **2019**, *88*, 57–72. [CrossRef]
45. Birem, M.; Klein, C. Multidomain spectral method for Schrödinger equations. *Adv. Comput. Math.* **2016**, *42*, 395–423. [CrossRef]

46. Jung, J.H.; Olmos-Liceaga, D. A Chebyshev multidomain adaptive mesh method for reaction-diffusion equations. *Appl. Numer. Math.* **2023**, *190*, 283–302. [CrossRef]
47. Qin, Y.; Chen, Y.; Huang, Y.; Ma, H. Multidomain Legendre-Galerkin least-squares method for linear differential equations with variable coefficients. *Numer. Math. Theory Methods Appl.* **2020**, *13*, 665–688.
48. Maleki, M.; Kajani, M.T. Numerical approximations for Volterra's population growth model with fractional order via a multi-domain pseudospectral method. *Appl. Math. Model.* **2015**, *39*, 4300–4308. [CrossRef]
49. Bambe Moutsinga, C.R. Robust Time Spectral Methods for Solving Fractional Differential Equations in Finance. Ph.D. Thesis, University of Pretoria, Pretoria, South Africa, 2021.
50. Kharazmi, E. Data-Driven Single-/Multi-Domain Spectral Methods for Stochastic Fractional PDEs. Ph.D. Thesis, Michigan State University, East Lansing, MI, USA, 2018.
51. Zhao, T.; Mao, Z.; Karniadakis, G.E. Multi-domain spectral collocation method for variable-order nonlinear fractional differential equations. *Comput. Methods Appl. Mech. Eng.* **2019**, *348*, 377–395. [CrossRef]
52. Klein, C.; Stoilov, N. Multidomain spectral approach to rational-order fractional derivatives. *Stud. Appl. Math.* **2024**, *152*, 1110–1132. [CrossRef]
53. Tavassoli Kajani, M. Numerical solution of fractional pantograph equations via Müntz–Legendre polynomials. *Math. Sci.* **2023**, *18*, 387–395. [CrossRef]
54. Xu, Q.; Hesthaven, J.S. Stable multi-domain spectral penalty methods for fractional partial differential equations. *J. Comput. Phys.* **2014**, *257*, 241–258. [CrossRef]
55. Hayat, U.; Kamran, A.; Ambreen, B.; Yildirim, A.; Mohyud-din, S.T. On system of time-fractional partial differential equations. *Walailak J. Sci. Technol. (WJST)* **2013**, *10*, 437–448.
56. Oloniiju, S.; Nkomo, N.; Goqo, S.; Sibanda, P. Shifted Chebyshev spectral method for two-dimensional space-time fractional partial differential equations. *Preprints* **2022**, 2022050138. [CrossRef]
57. Trefethen, L.N. *Finite Difference and Spectral Methods for Ordinary and Partial Differential Equations*; Cornell University: New York, USA, 1996.
58. Dahlquist, G.G. A special stability problem for linear multistep methods. *BIT Numer. Math.* **1963**, *3*, 27–43. [CrossRef]
59. Ehle, B.L. On Padé Approximations to the Exponential Function and A-Stable Methods for the Numerical Solution of Initial Value Problems. Ph.D. Thesis, University of Waterloo, Waterloo, ON, USA, 1969.
60. Odibat, Z.M.; Momani, S. An algorithm for the numerical solution of differential equations of fractional order. *J. Appl. Math. Informatics* **2008**, *26*, 15–27.
61. Li, C.; Yi, Q.; Chen, A. Finite difference methods with non-uniform meshes for nonlinear fractional differential equations. *J. Comput. Phys.* **2016**, *316*, 614–631. [CrossRef]
62. Raja, M.A.Z.; Khan, J.A.; Qureshi, I.M. Solution of fractional order system of Bagley-Torvik equation using evolutionary computational intelligence. *Math. Probl. Eng.* **2011**, *2011*. [CrossRef]
63. Zafar, A.A.; Kudra, G.; Awrejcewicz, J. An investigation of fractional Bagley–Torvik equation. *Entropy* **2019**, *22*, 28. [CrossRef] [PubMed]
64. Sene, N. Introduction to the fractional-order chaotic system under fractional operator in Caputo sense. *Alex. Eng. J.* **2021**, *60*, 3997–4014. [CrossRef]
65. Azam, A.; Aqeel, M.; Ahmad, S.; Ahmad, F. Chaotic behavior of modified stretch-twist-fold (STF) flow with fractal property. *Nonlinear Dyn.* **2017**, *90*, 1–12. [CrossRef]
66. Li, C.; Liao, X.; Yu, J. Synchronization of fractional order chaotic systems. *Phys. Rev. E* **2003**, *68*, 067203. [CrossRef]
67. Sene, N. Study of a Fractional-Order Chaotic System Represented by the Caputo Operator. *Complexity* **2021**, *2021*, 5534872. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.