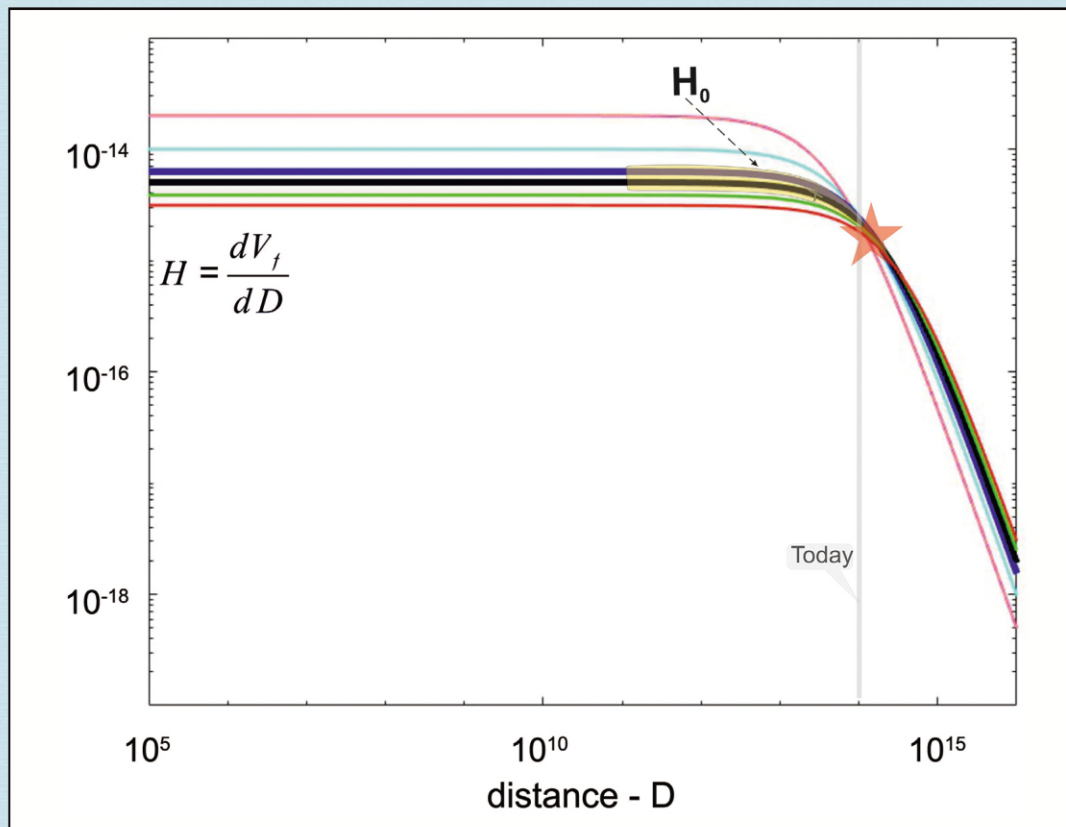


Journal of Modern Physics



ISSN: 2153-1196



Journal Editorial Board

ISSN: 2153-1196 (Print) ISSN: 2153-120X (Online)

<https://www.scirp.org/journal/jmp>

Editorial Board

Prof. Nikolai A. Sobolev	Universidade de Aveiro, Portugal
Prof. Mohamed Abu-Shady	Menoufia University, Egypt
Dr. Hamid Alemohammad	Advanced Test and Automation Inc., Canada
Prof. Emad K. Al-Shakarchi	Al-Nahrain University, Iraq
Dr. Francesco Bajardi	Scuola Superiore Meridionale, Italy
Prof. Antony J. Bourdillon	UHRL, USA
Dr. Swarniv Chandra	Government General Degree College, India
Prof. Tsao Chang	Fudan University, China
Prof. Wan Ki Chow	The Hong Kong Polytechnic University, China
Prof. Jean Cleymans	University of Cape Town, South Africa
Prof. Stephen Robert Cotanch	NC State University, USA
Prof. Claude Daviau	Ministry of National Education, France
Prof. Rami Ahmad El-Nabulsi	Chiang Mai University, Thailand
Prof. Peter Chin Wan Fung	The University of Hong Kong, China
Prof. Ju Gao	The University of Hong Kong, China
Prof. Robert Golub	North Carolina State University, USA
Dr. Sachin Goyal	University of California, USA
Dr. Wei Guo	Florida State University, USA
Prof. Karl Hess	University of Illinois, USA
Prof. Peter Otto Hess	Universidad Nacional Autónoma de México, Mexico
Prof. Ahmad A. Hujeirat	University of Heidelberg, Germany
Prof. Haikel Jelassi	National Center for Nuclear Science and Technology, Tunisia
Prof. Magd Elias Kahil	October University for Modern Sciences and Arts (MSA), Egypt
Prof. Santosh Kumar Karn	Dr. APJ Abdul Kalam Technical University, India
Prof. Sanjeev Kumar	Dr. Bhimrao Ambedkar University, India
Dr. Giuseppe Levi	Bologna University, Italy
Prof. Yu-Xian Li	Hebei Normal University, China
Prof. Anton A. Lipovka	Sonora University, Mexico
Prof. Wu-Ming Liu	Chinese Academy of Sciences, China
Dr. Ludi Miao	Cornell University, USA
Dr. Grégory Moreau	Paris-Saclay University, France
Prof. Christophe J. Muller	University of Provence, France
Dr. Rada Novakovic	National Research Council, Italy
Dr. Vasilis Oikonomou	Aristotle University of Thessaloniki, Greece
Prof. Vinod Prasad	Swami Shirdhanand College Delhi, India
Prof. Tongfei Qi	University of Kentucky, USA
Prof. Mohammad Mehdi Rashidi	University of Birmingham, UK
Prof. Haiduke Sarafian	The Pennsylvania State University, USA
Prof. Kunnat J. Sebastian	University of Massachusetts, USA
Dr. Ramesh C. Sharma	Ministry of Defense, India
Dr. Reinoud Jan Slagter	Astronomisch Fysisch Onderzoek Nederland, The Netherlands
Dr. Giorgio Sonnino	Université Libre de Bruxelles, Belgium
Prof. Yogi Srivastava	Northeastern University, USA
Dr. Mitko Stoev	South-West University “Neofit Rilski”, Bulgaria
Dr. A. L. Roy Vellaisamy	City University of Hong Kong, China
Prof. Lev Zalman Vilenchik	Felicitex Therapeutics, USA
Prof. Anzhong Wang	Baylor University, USA
Prof. Cong Wang	Beihang University, China
Prof. Yuan Wang	University of California, Berkeley, USA
Prof. Peter H. Yoon	University of Maryland, USA
Prof. Meishan Zhao	University of Chicago, USA
Prof. Pavel Zhuravlev	University of Maryland at College Park, USA

Table of Contents

Volume 14 Number 3

February 2023

Hubble Tension versus the Cosmic Evolution of Hubble Parameter in the Unicentric Model of the Observable Universe

A. Hujeirat.....183

Electron Shape Calculated for the Dual-Charge Dual-Mass Model

A. Young.....198

Average Luminous Mass of Early Galaxies at $z = 10 - 20$ Predicted as $\sim 10^9$ Solar Masses

T. R. Mongan.....208

An Outline of the Grand Unified Theory of Gauge Fields

B. Qiao.....212

Disjointed Equivalence of Gravitational and Inertial Mass

D. J. Pons.....237

A Physical Explanation for the Formation of Auroras

W. H. Qian.....271

The Speed of Light Is Not Constant in Basic Big Bang Theory

J. Van Royen.....287

Photon Structure and Wave Function from the Vector Potential Quantization

C. Meis.....311

Inadequacies in the Current Definition of the Meter and Ramifications Affecting Einstein's Second Postulate of Relativity

S. D. Deines.....330

Mathematical Aspects of $SU(2)$ and $SO(3, \mathbb{R})$ Derived from Two-Mode Realization in Coordinate-Invariant Form

A. Wünsche.....361

Journal of Modern Physics (JMP)

Journal Information

SUBSCRIPTIONS

The *Journal of Modern Physics* (Online at Scientific Research Publishing, <https://www.scirp.org/>) is published monthly by Scientific Research Publishing, Inc., USA.

Subscription rates:

Print: \$89 per issue.

To subscribe, please contact Journals Subscriptions Department, E-mail: sub@scirp.org

SERVICES

Advertisements

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

Reprints (minimum quantity 100 copies)

Reprints Co-ordinator, Scientific Research Publishing, Inc., USA.

E-mail: sub@scirp.org

COPYRIGHT

Copyright and reuse rights for the front matter of the journal:

Copyright © 2023 by Scientific Research Publishing Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY).

<http://creativecommons.org/licenses/by/4.0/>

Copyright for individual papers of the journal:

Copyright © 2023 by author(s) and Scientific Research Publishing Inc.

Reuse rights for individual papers:

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

Disclaimer of liability

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

PRODUCTION INFORMATION

For manuscripts that have been accepted for publication, please contact:

E-mail: jmp@scirp.org

Hubble Tension versus the Cosmic Evolution of Hubble Parameter in the Unicentric Model of the Observable Universe

Ahmad Hujeirat

IWR, Heidelberg University, Heidelberg, Germany

Email: AHujeirat@iwr.uni-heidelberg.de

How to cite this paper: Hujeirat, A. (2023) Hubble Tension versus the Cosmic Evolution of Hubble Parameter in the Unicentric Model of the Observable Universe. *Journal of Modern Physics*, 14, 183-197.

<https://doi.org/10.4236/jmp.2023.143013>

Received: January 1, 2023

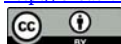
Accepted: February 4, 2023

Published: February 7, 2023

Copyright © 2023 by author(s) and Scientific Research Publishing Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

Recently, a unicentric model of the observable universe (UNIMOUN) was proposed. Accordingly, big bangs are common events in our infinitely large, flat, homogeneous and isotropic parent universe. Their progenitors are clusters of cosmically dead and massive neutron stars that merged after reaching the ultimate lowest quantum energy state, where the matter is in an incompressible superconducting gluon-quark superfluid state and zero-entropy, hence granting the resulting progenitors measurable sizes and immunity to collapsing into black holes. Our big bang happened to occur in our neighbourhood, thereby enduing the universe, the observed homogeneity and isotropy. As the enclosed mass of the progenitor was finite, the dynamically expanding curved spacetimes embedded the fireball started flattening to finally diffuse into the flat spacetime of the parent universe. By means of general relativistic numerical hydrodynamical calculations, we use the $\mathcal{H}$ -metric to follow the time-evolution of the spacetime embedding the progenitor during the hadronization and the immediately following epochs. Based thereon, we find that the kinetic energy of newly created normal matter increases with distance in a self-similar manner, imitating thereby outflows of nearly non-interacting particles. On cosmic time scales, this behaviour yields a Hubble parameter, $H(t)$, which decreases slowly with the distance from the big bang event. Given the sensitivity of the data of the Cosmic Microwave Background (CMB) from Planck to the underlying cosmological model, we conclude that UNIMOUN is a viable alternative to Λ CMD-cosmologies.

Keywords

General Relativity; Big Bang, Black Holes, Quasars, Neutron Stars, Quantum Chromodynamics, Condensed Matter, Incompressibility, Superfluidity, Super-Conductivity

1. Introduction

The seminal paper of Edwin Hubble in 1929 about the relation between velocity and distance among extra-galactic objects changed our understanding of the universe entirely and caught Einstein by surprise. He then retracted his static model of the universe and removed the famous cosmological constant Λ . This relation is now known as the Hubble law, and it indicates that distant galaxies recede from each other at much higher rates than their nearby counterparts. Over the last 100 years, the value of the Hubble constant dropped from its initial value of roughly $500 \text{ km}\cdot\text{s}^{-1}\cdot\text{Mpc}^{-1}$ down to approximately $70 \text{ km}\cdot\text{s}^{-1}\cdot\text{Mpc}^{-1}$, thanks to the continuous improvement of instrumentations in observational astronomy (see [1] [2] [3], and the references therein).

On the other hand, the rapid improvements in instrumentations in observational astronomy have recently highlighted discrepancies among key cosmological parameters. While the Planck data of the CMB radiation reveals that the Hubble constant is $H_0 = 67.27 \pm 0.66 \text{ km/s/Mps}$, the Type Ia supernova measurements yield $H_0 = 73.2 \pm 1.3 \text{ km/s/Mps}$ [4] [5]. The precisions of these predictions are, unfortunately, not model-independent, but certain a priori assumptions related to the topology of spacetime and to Λ CDM-cosmologies¹ are inherently included. These in turn rely directly or indirectly on components that are poorly understood, such as inflation, dark matter, and dark energy [6] [7] [8] [9], for which there are no remedies in sight. Continuing to rely on these models may delay the appearance of new consistent models of the universe.

Recently UNIMOUN has been presented. It is a general relativistic (-GR) model and relies neither on inflation nor on dark matter and/or dark energy. It revives Einstein's objection to the existence of black holes. Still, it incorporates components from quantum field theory (QFT) and chromodynamics, such as incompressibility and superfluidity of gluon-quark matter [10] [11] [12]. Also, it hypothesizes a connection between general relativity and QFT that spacetimes embedding SuSu-cores must be conformally flat. Here SuSu-matter is the state of supranuclear dense matter with zero entropy, where the matter is said to be in incompressible superconducting gluon-quark superfluid phase [13].

The motivations underlying this hypothesis may be summarized as follows:

- The well-observed glitch phenomena in pulsars [14] [15] and the merger of the neutron stars in GW170817 [16] suggest that massive pulsars may be born with embryonic cores that are made of incompressible superconducting gluon-quark superfluids with zero entropy [12], (*i.e.* SuSu-matter). The density here is equal to the maximum possible density permitted in our universe [10], ρ_{max}^{uni} , which is predicted to be $[3\rho_0 \leq \rho_{max}^{uni} \leq 5\rho_0]$. These embryonic cores serve as seeds in a universal life cycle that leads to the formation of the next generations of big bangs (see Figure 1).
- The GR field equations are incapable of treating incompressible SuSu-matter self-consistently [17]. In fact, the causality condition requires the pressure in

¹CDM stands for Cold Dark Matter.

SuSu-matter to be upper-limited by the constant energy density, thereby blocking the regime of ultrabaric state of matter [18]. However, as long as the embedding spacetime is not flat, this causality condition under these circumstances will be violated. In this case, each side of the GR field equations must vanish at every event of the spacetime embedding the core, *i.e.*

$G_{\mu\nu} = T_{\mu\nu} = 0$. A reliable astrophysical situation requires a new type of matter, such as the incompressible SuSu-matter, which sets to evolve in space and time as dictated by an intrinsically time-dependent metric.

Indeed, the matter inside the cores of glitching pulsars, e.g. the Crab and Villa pulsars, are expected to obey these conditions nicely [19].

- Geometrical locations in incompressible SuSu-cores are ill-defined [20]; the interiors of these cores are completely uniform, homogeneous and isotropic. Unlike normal stars, where the gradient of any physical quantity has to fulfil the regularity condition at the centre serve, in SuSu-cores however, these conditions are fulfilled at every event, save the confining surface, where the coupling information between all its constituents are mirrored and encoded in a holographic manner as proposed by the holographic principle [21].
- Both the homogeneity and isotropy of both the observable universe and the interior of SuSu-cores imply that the decaying front must have started at the surface and not in the interior of SuSu-cores. This also may explain why the progenitor of our BB escaped the collapse into a hypermassive black hole (BH).

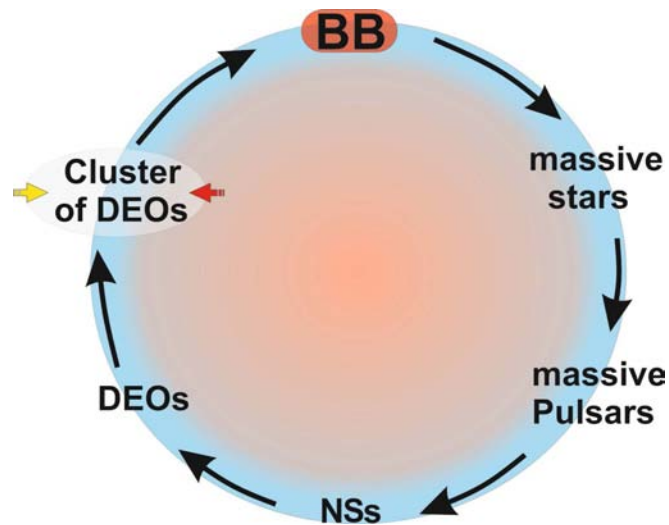


Figure 1. A schematic description of the life cycle of an arbitrary unicentric universe: It starts with a big bang explosion (**BB**) through which the first generation of massive stars are formed and subsequently collapse to generate the first generation of massive pulsars, that evolve toward neutron stars (**NSs**). As these degenerate stars cool down on cosmic time scales, the masses and dimensions of their SuSu-cores grow to finally turn into invisible dark energy objects (**DEOs**). These objects start conglomerating on relatively long time scales to form tight clusters of DEOs, some of which may originate from the ambient parent universe. Once merged, they create the progenitors for the next generation of big bangs.

When applying these arguments to the evolution of our universe, the following universal life cycle emerges: It starts with a big bang explosion triggered by the hadronization process of the progenitor, which is expected to last for roughly one hour or less, then the jettisoned normal matter in the ambient space cools down, contracts and subsequently forms the first generation of massive stars. These stars are predicated to form from the gravitationally bound normal matter inside the local universe after the hadronization process has been completed, which implies that the epoch of formation of the first generation of stars may have been completed much earlier than in the models based on Λ CDM-cosmologies. However, once they are formed, these objects are short-living and eventually doomed to collapse and form the first generation of massive pulsars that subsequently evolve into massive NSs. These, in turn, continue to liberate their rest energies whilst their SuSu-cores grow in mass and dimension to finally metamorphose into invisible DEOs [Dark energy in DEOs and in cosmology are different by construction. In the former case, the pressure is not a local thermodynamical quantity (see [10], and the references therein)]. In the following epoch, DEOs start conglomerating toward forming tight clusters that subsequently merge and form the progenitors for the next generations of big bangs. In this picture, the DEOs that formed the progenitor may not necessarily originate from the same perturbed sub-universe containing the BB but may originate from multiple sub-universes as depicted in **Figure 1**. The origin of our observable homogenous and isotropic universe is a big bang that happened in our neighbourhood, hence the model's name. The main properties of UNIMOUN may be summarised as follows:

- Our observable universe is a submanifold of our infinitely large and flat parent universe.
- Big bangs are frequent events through which limited topological deformations of the embedding spacetime are triggered but ultimately diffuse out into the infinitely large and flat parent universe.
- The occurrences of BBs are self-sustained processes (see **Figure 1**), that evolve without invoking exotic and/or external forces/phenomena, such as inflation [9], dark matter [22] and dark energy [7].
- The BB-progenitors are made of SuSu-matter, which enables the progenitors to be of finite, measurable sizes and prevent their collapse into hypermassive black holes.
- Our BB must have occurred in our neighbourhood, and therefore may explain the observed homogeneity of the universe. However, the isotropic character has been retained after billions of years, when the initially curved spacetime finally diffused into the infinitely large and flat parent universe.
- The observed receding velocities of high redshift galaxies are consequences of extensive matter and momentum transfer from the fireball-matter into the then old and inactive ones that populated the ambient spacetime.

2. The Analytical and Numerical Approach

Our analytical and numerical approaches are based on solving the GR time-

dependent field equations using the $[4 \times 4]$ diagonal $\mathcal{H}$ -metric [23]:

$$ds_H^2 = \sum g_{\mu\nu} dx^\mu dx^\nu, \quad (1)$$

where the non-zero diagonal elements of the metric g , read:

$$\begin{aligned} g_{00} &= e^{2\mathcal{V}(r,t)} = 1 - \chi(r,t) \\ g_{11} &= e^{2\lambda(r,t)} = -\frac{1}{1 - \chi(r,t)} \\ g_{22} &= -r^2, \quad g_{33} = -r^2 \sin^2 \theta, \end{aligned} \quad (2)$$

where $\mathcal{V}, \lambda$ are functions of the distance r and time with respect to $\mathcal{H}_0$, respectively. Here both indices μ and ν run from 0 to 3, and $\chi(r,t)$ is the curvature function defined as:

$$\chi(r,t) = \alpha_{bb} \frac{m_n(r,t)}{r}. \quad (3)$$

$m_n(r,t)$ here corresponds to the enclosed mass of normal matter only, and α_{bb} is the so-called compactness parameter.

In this model, only normal matter is set to interact with the embedding spacetime and affect its topology. As the progenitor is made of incompressible Su-Su-matter with vanishing entropy, $m_n(r, t \leq 0) = 0$ is set to vanish. Hence prior to hadronization, the topology of the spacetime embedded the progenitor was indistinguishable from that of the parent spacetime. At $t = 0$ the hadronization front formed at the surface of the progenitor and started propagating inwards, converting SuSu-matter into normal matter. During this hadronization phase, $m_n(r,t)$ set to increase continuously, thereby enhancing the curvature of the embedding spacetime.

The topological changes of the spacetime are calculated using the following equation (see [20] [23], for the derivation):

$$\begin{aligned} & \frac{1}{2r} \left[\frac{\partial}{\partial t} (e^{-2\mathcal{V}}) + \frac{e^{-2\lambda}}{e^{-2\mathcal{V}}} \frac{\partial}{\partial r} (e^{-2\mathcal{V}}) + \frac{\partial}{\partial r} e^{-2\lambda} \right] \\ &= -\frac{1}{2} \kappa (\mathcal{E} + p) \left[\Gamma^2 (g_{00} - V^2 g_{11}) \right] \\ &= -\frac{1}{2} \kappa (\mathcal{E} + p) \bar{\bar{\Gamma}}, \end{aligned} \quad (4)$$

where Γ is the Lorentz factor, whereas $\bar{\bar{\Gamma}}$ corresponds to the modified Lorentz factor:

$$\bar{\bar{\Gamma}}^2 = \begin{cases} \frac{g_{00} + V^2 g_{11}}{g_{00} - V^2 g_{11}} & : \text{Generalform} \\ 1 & : \text{Hydrostatic cores embedded in curved spacetimes} \\ \frac{1 + \beta^2}{1 - \beta^2} & : \text{Flat spacetime.} \end{cases} \quad (5)$$

In fact, $\bar{\bar{\Gamma}}$ was identified in [23] to be more inductive to the state of matter embedded in dynamically varying spacetimes than the classical Lorentz factor

Γ .

We term the function $\mathcal{V}(r, t)$ as the “gravitational potential” and r is the distance from the preferred observer $\mathcal{H}_0$ located at the center of the progenitor, *i.e.* at $r = 0$.

Equation (4) is solved together with the time-dependent equations of mass and momentum transfer, which read:

$$\begin{aligned} & \left(\frac{1}{\sqrt{-g}} \frac{\partial}{\partial t} (\sqrt{-g} \mathcal{D}) + \frac{1}{\sqrt{-g}} \frac{\partial}{\partial r} \sqrt{-g} (\mathcal{D} V) \right) = 0 \\ & \frac{1}{\sqrt{-g}} \frac{\partial}{\partial t} (\sqrt{-g} \mathcal{M}^r) + \frac{1}{\sqrt{-g}} \frac{\partial}{\partial r} (\sqrt{-g} \mathcal{M}^r V) = -\frac{\partial P}{\partial r} + \frac{\mathcal{M}^t}{2} (g_{tt,r} + V^2 g_{rr,r}). \end{aligned} \quad (6)$$

$V, \mathcal{D}$ here are the transport velocity and relativistic density, *i.e.* the normal rest density times the Lorentz factor. $\sqrt{-g} (= r^2 R^3 \sin(\theta) / \sqrt{GW})$ is the determinant of the $\mathcal{H}$ -metric. $\mathcal{M}^\sigma (\doteq \mathcal{D} h u^\sigma)$ and u^σ stand for the four-momenta and four-velocities (in $\sigma = \{t, r, \theta, \phi\}$ coordinates), respectively, whereas h stands for the enthalpy. To close the set of equations, an adiabatic equation of state (EOS) is used.

The above set of equations is solved implicitly in combination with the finite-volume strategy and within the framework of the operator-splitting approach strategy (see [24] [25] for further details).

The hydrodynamical calculations are run using the following initial conditions:

- The initial mass of the progenitor is $M_p = 10^{24} M_\odot$, which is solely made of incompressible SuSu-matter. $M_\odot$ here is solar mass.
- The spacetime inside and outside the progenitor is initially flat. This is a consequence of the underlying hypothesis that SuSu-matter with zero-entropy is neither capable of interacting nor affecting the topology of the embedding spacetime (This addresses the possibilities that there might be a hidden connection between gravity and entropy [26]).
- At $t = 0$, the membrane confining the incompressible SuSu-matter in the entire progenitor is removed, and as a consequence, a hadronization front forms at the progenitor’s surface and start propagating inwards with speed V_f , thereby converting SuSu-matter into the normal matter at a rate (non-dimensional):

$$\dot{M}^{nor} = \begin{cases} 1 + t^2 - 2t & : 0 \leq t \leq \tau_{dyn} \\ 0 & : t > \tau_{dyn} \end{cases} \quad (7)$$

thereby increasing the mass of normal matter at the rate:

$$M^{nor} = \begin{cases} 3t - 3t^2 + t^3 & : 0 \leq t \leq \tau_{dyn} \\ 1 & : t > \tau_{dyn} \end{cases} \quad (8)$$

where τ_{dyn} is the dynamical time scale, which is equal to the radius of the progenitor r_p divided by V_f . In the present calculations, we set $V_f = 0.9999$ in the unit of c . As the density of SuSu-matter is $\rho = \rho_{uni}^{max} = 3\rho_0$, we obtain a pro-

genitor's radius $r_p = 8.38 \times 10^{13}$ cm and $\tau_{dyn} = 2.79 \times 10^3$ s. We use $M_p = (4\pi/3)\rho_{cr}r_0^3$ and $\dot{M}^{nor}$ in $(4\pi r_0^2) \times (\rho_{cr} V_f)$ as reference quantities for non-dimensionalizing the equations.

Based thereon, we show in **Figure 2** the time-evolution of the radial velocity V , the modified Lorentz factor $\bar{u}^t (\doteq \bar{\Gamma})$ and the gravitational redshift Z_{grav} during the hadronization and the immediately following epochs. The achieved time here is $11\tau_{dyn}$, which is equivalent to the first 7.5 hours of the big bang explosion.

The results obtained here reconfirm the following arguments:

- Both the radial velocity and the kinetic energy per mass, increase with the distance self-consistently, *i.e.* $dV/dr > 0$, and don't violate the causality condition.

In the very outer shells of the progenitor, the newly created normal matter is jettisoned into the surrounding flat spacetime with ultra-relativistic velocities, almost unaffected by the curved topology of the embedding spacetime, as the total available mass of normal matter, $m_n(r, t)$ is still small, relatively.

During hadronization, the gravitational well in the central regions of the fire-ball becomes increasingly deeper, enforcing thereby the newly created normal matter here to lose a considerable part of its kinetic energy to overcome the potential well. This may be considered as enforced balance between outward advection of momentum and deceleration through curvature:

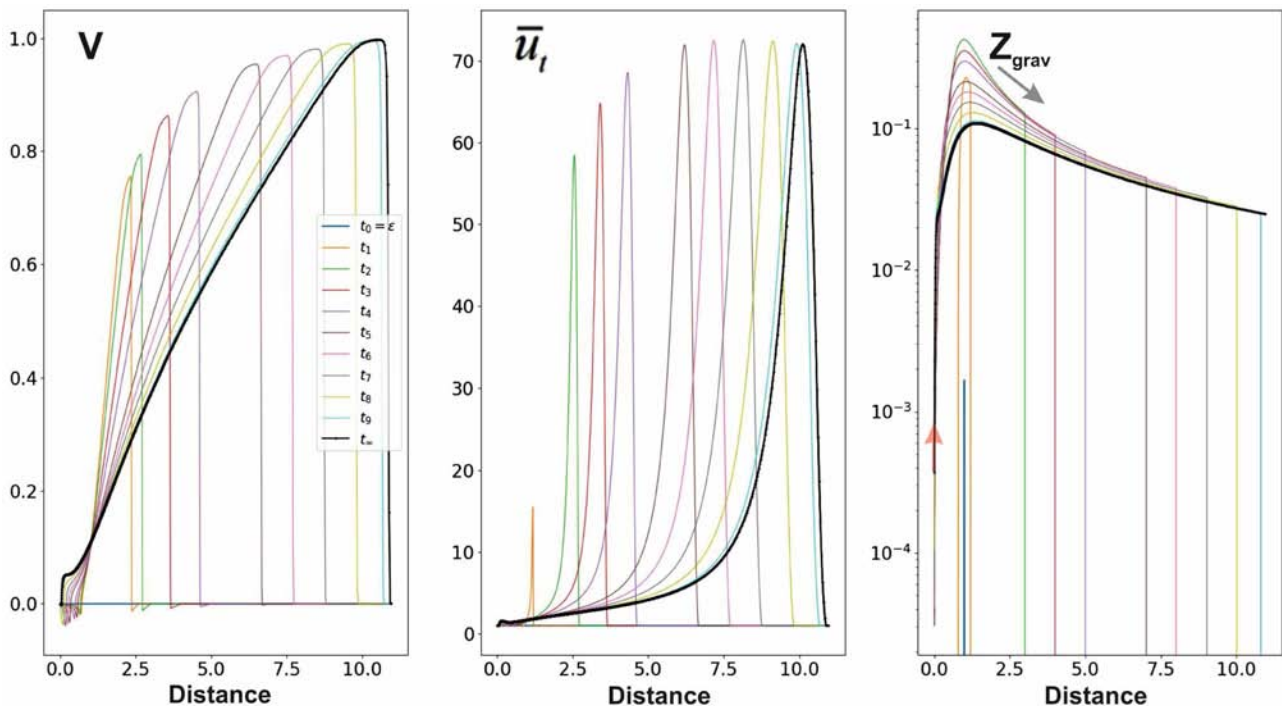


Figure 2. Different snapshots of the radial distributions of the velocity V , the modified Lorentz factor $\bar{u}^t$ and the gravitational redshift Z_{grav} , during the hadronization and the immediately following epochs. The time-sequence of the snapshots are marked with different colours, starting with blue and ending with black. The tendency of flattening of the spacetime is emphasized by two arrows in the third panel of Z_{grav} .

$$\frac{1}{\sqrt{-g}} \frac{\partial}{\partial r} (\sqrt{-g} \mathcal{M} V) \approx \frac{\mathcal{M}^t}{2} (g_{u,r} + V^2 g_{rr,r}). \quad (9)$$

when this equation is integrated over volume and simplified, we obtain the usual balance of potential energy and kinetic energy prescribed by Newtonian mechanics:

$$V^2(r) - V_f^2 \approx g_{u,r} (r - r_f). \quad (10)$$

In the early stages of the hadronization epoch, $g(t)_{u,r}$ in the outer shells is negligibly small, and therefore the propagational velocity, V and the jettisoning/detonation velocity, V_f become almost identical. Once the hadronization process has been completed, then the deep gravitational well must have reached its maximum, and therefore particles in the very central region must lose considerable fractions of their kinetic energies to climb up the potential well. This is a consequence of the spatially and temporarily varying topology of the spacetime embedding the expanding fireball, which differs fundamentally from the stationary spacetimes, such as the Schwarzschild one, where $g_{u,r}$ attains its maximum at the surface of the object.

Both plots of V and $\bar{u}^t$ show that the bulk of the kinetic energy is transported outwards, where the velocity converges to V_f , thereby giving rise to a constant outflow rate in which $\rho \sim 1/r^2$. This trend has been made possible as the gravitational redshift tends to decrease with time, as depicted with arrows in the third panel of **Figure 2**. As the spacetime embedding the fireball expands, its curvature must flatten with time as prescribed by the minimum energy theorem (see [10], and the references therein).

- In **Figure 3** we show different snapshots of V and $\bar{u}^t$ in the innermost region. Here the low and inward-oriented motion implies that the newly created normal matter is trapped in the fireball's potential well, therefore slowing its motion and delaying its escape into the ambient flat spacetime. However, the potential well is not sufficiently deep to trigger a collapse of the matter in the central regions into a BH. On cosmic scales, the matter here is expected to cool down and form the galaxies of our local universe, whose redshifts are relatively small.

The Cosmic Evolution and the Hubble Parameter

Prior to the big bang explosion, the progenitor was levitating at the background of our infinitely large and flat parent universe, which populated all types of galaxies and objects. UNIMOUN predicts the spacetime surrounding the progenitor of our big bang to have been populated, but with at least 13.8 Gyr old and inactive galaxies. When the outward-propagating normal matter of the expanding fireball hit the surrounding galaxies with the ultra-relativistic speeds, mass with enormous momentum was transferred to these galaxies, turning them active and setting them into outward-oriented motions, whose increase rates depend on different parameters, and mainly on their cross-sections, masses and distances from the BB-event [20].

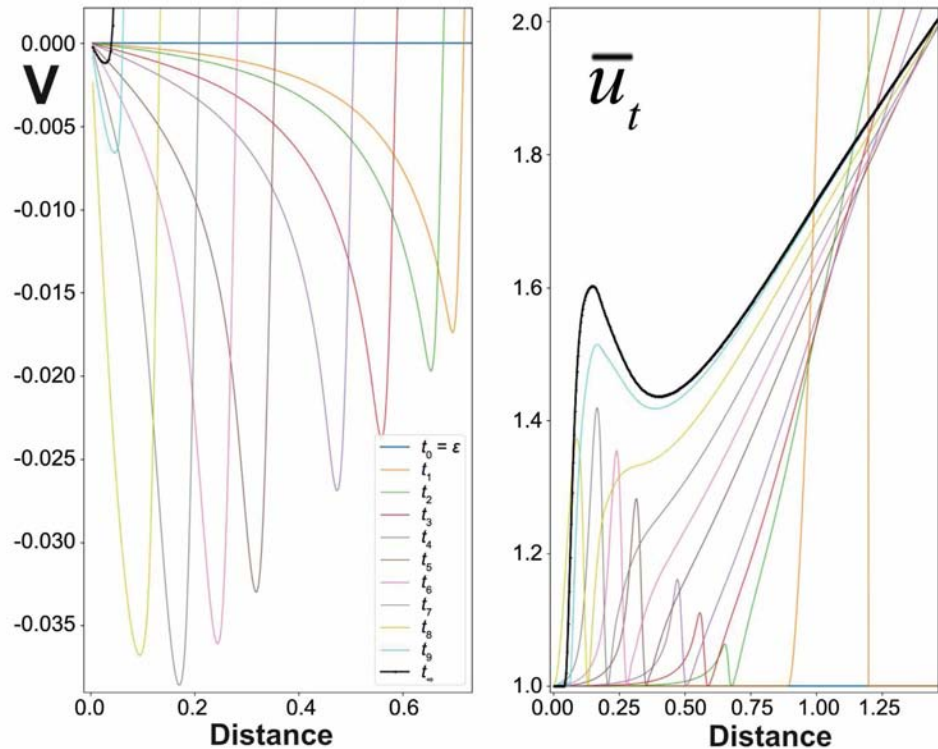


Figure 3. The hydrodynamics of the newly created normal matter in the central region during the hadronization phase and thereafter is shown through a time-sequence of snapshots of the radial velocity, V and the modified Lorentz factor $\bar{u}^t$. The small negative radial velocity, though relatively small, which is a consequence of the relatively high compactness of the progenitor, would considerably slow the receding velocities of the galaxies formed in the vicinity of the big bang. In the right panel, the continuous flattening of the spacetime embedding the fireball is sketched in accord with the minimum energy theorem.

As the observable effect of momentum transfer may need much longer time than the dynamical time scale of the progenitor, we need to integrate Equations (6) over the volume of the smallest spherical sector enclosing a prototype of an old inactive galaxy. In this case, the integration of the continuity equation yields:

$$\frac{\delta \mathcal{M}}{\delta t} + (\dot{\mathcal{M}}_o - \dot{\mathcal{M}}_{in}) = 0, \quad (11)$$

where $\delta \mathcal{M}, \dot{\mathcal{M}}_{in} (= A_{cone}^G \mathcal{D}V)$ and $\dot{\mathcal{M}}_o$ stand for the total mass absorbed by the galaxy and mass inflow in and outflow from the galaxy, respectively. A_{cone}^G is the cross-section of the spherical sector of the cone enclosing the concerned galaxy located at r_G . For $r_G \gg r_p$, the spacetime surrounding the galaxy is set to be initially flat.

For simplicity, we consider the mass ejection rate out of the galaxy to be $\dot{\mathcal{M}}_o = \alpha \dot{\mathcal{M}}_{in}$, where $\alpha < 1$. Moreover, we assume the collision between the normal matter of the expanding fireball with the motionless galaxy to be inelastic, *i.e.*,

$$\delta \mathcal{M} \beta_{in} + M_G \beta_G^{\rightarrow 0} = (\delta \mathcal{M} + M_G) \beta_f, \quad (12)$$

where $\beta_{in}, \beta_G, \beta_f$ are the inflow velocity of normal matter through the left boundary of the galaxy, the velocity of the galaxy prior collision, which is set to vanish, and the final combined velocity of the galaxy together with the captured matter, respectively. $\mathcal{M}_G$ stands for the mass of the galaxy. In this case, the final velocity of the combined object reads:

$$\frac{\beta_f}{\beta_{in}} = \frac{\delta\mathcal{M}}{M_G + \delta\mathcal{M}} = \frac{(1-\alpha)\frac{\beta}{\sqrt{1-\beta^2}}t}{M_G + (1-\alpha)\frac{\beta}{\sqrt{1-\beta^2}}t}, \quad (13)$$

where we inserted $\delta\mathcal{M}$ from the integrated continuity equation (Equation (11)).

For $M_G > 0$, the causality condition is satisfied self-consistently.

In **Figure 4** the final velocity of the bombarded galaxy is plotted versus time (in τ_{dyn} units) for different injection/ejection rates and masses. Here the following trends may be identified:

- Independent of M_G and $\dot{\mathcal{F}}$, the acceleration rate of all galaxies is almost identical.
- Massive or distant galaxies require, as expected, much longer times to accelerate their outward motions compared to low masses or close ones.

Based thereon, the fireball may have affected the observed galaxies at least in three different ways (see **Figure 5**):

- Type G_A galaxies may have formed after the BB and populated the local universe. Their relative motions may not be well-defined, though they must have slow-receding velocity components. These receding velocity components are direct consequences of the expanding spacetime embedding the expanding fireball enclosing a finite mass of normal matter. Following the minimum energy theorem, the spacetime must flatten as depicted by the arrows in Panel 3 of **Figure 2**.
- Old and inactive galaxies, G_B , that populated the spacetime surrounding the BB then at intermediate distances from the BB event. The cross-sections and distances of these galaxies should have laid in a certain regime of parameters, which allows them to gain sufficient mass and momentum and still accelerate their outward-oriented motions without deforming their internal structures completely. Meanwhile the gained mass-momentum enabled them to reach ultra-high relativistic velocities, and currently, they populate the very outer regions of the observable universe.
- Relatively remote, old and inactive galaxies, G_C , that populated the parent universe, should have been hit by the expanding fireball of the BB. Due to their low effective cross sections, the overall gained mass and momentum were relatively low. Still, they were sufficient to turn them on and accelerate their outward-oriented motions up to intermediate values.

For the given dependence of β_f on time, the corresponding Hubble parameter reads:

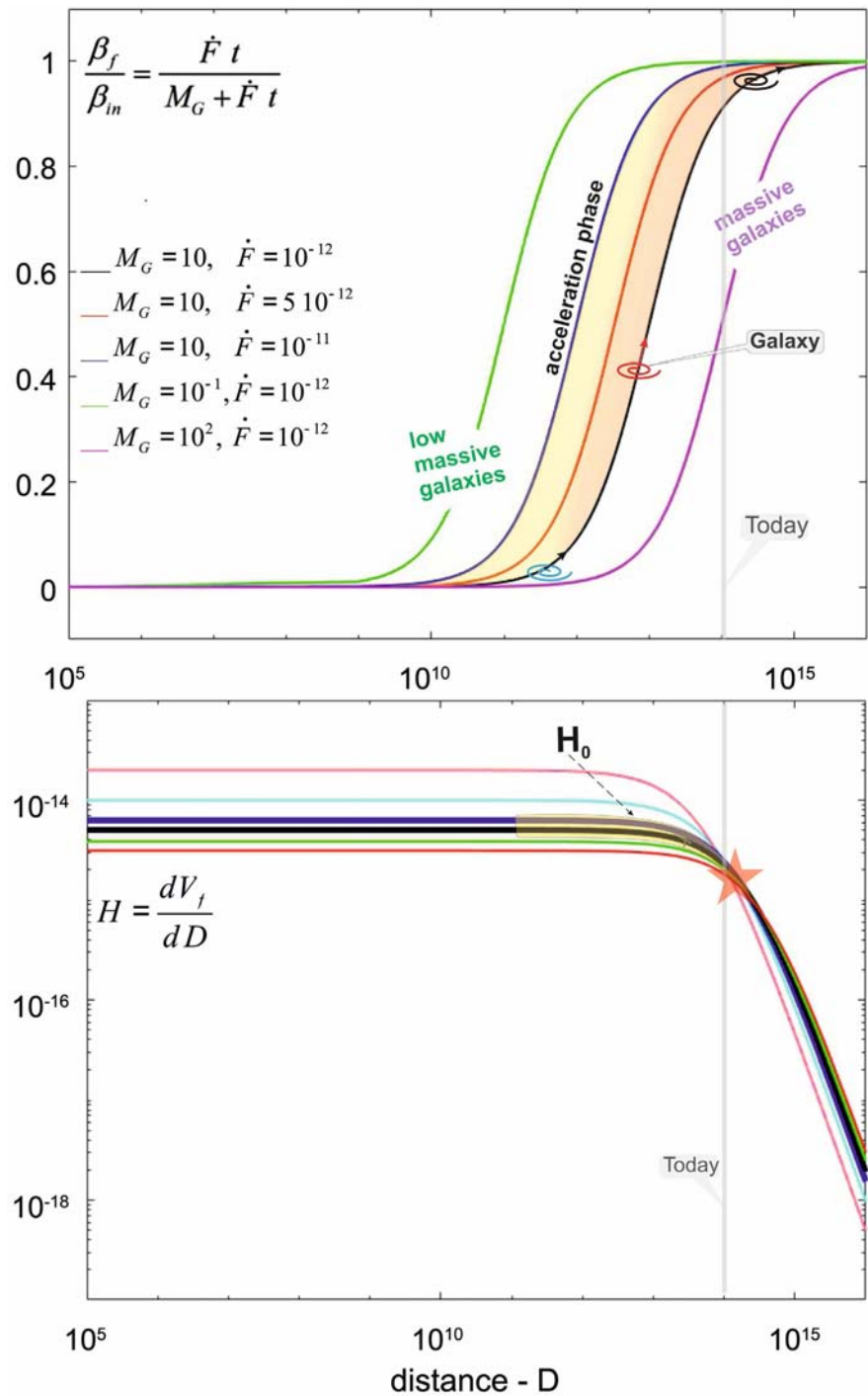


Figure 4. The time-development of the receding velocity of galaxies for different galaxy masses and incident fluxes $\dot{F}$ (top). The blue, red and black lines correspond to galaxies having the same mass $M_G = 10$ and incident fluxes $\dot{F} = 10^{-11}, 5 \times 10^{-12}, 10^{-12}$, respectively. The green and cyan lines correspond to low and massive galaxies, respectively. In the lower panel, the Hubble parameter, $H = dV/dD$ is shown for different galaxy masses and incident fluxes. In all cases, the H decreases slowly with the cosmic time, *i.e.* with distance, though at different rates. Galaxies with high effective masses display relatively lower Hubble constants, but in roughly 21 Gyr from now, the different values converge to a single value, beyond which H evolves reversibly.

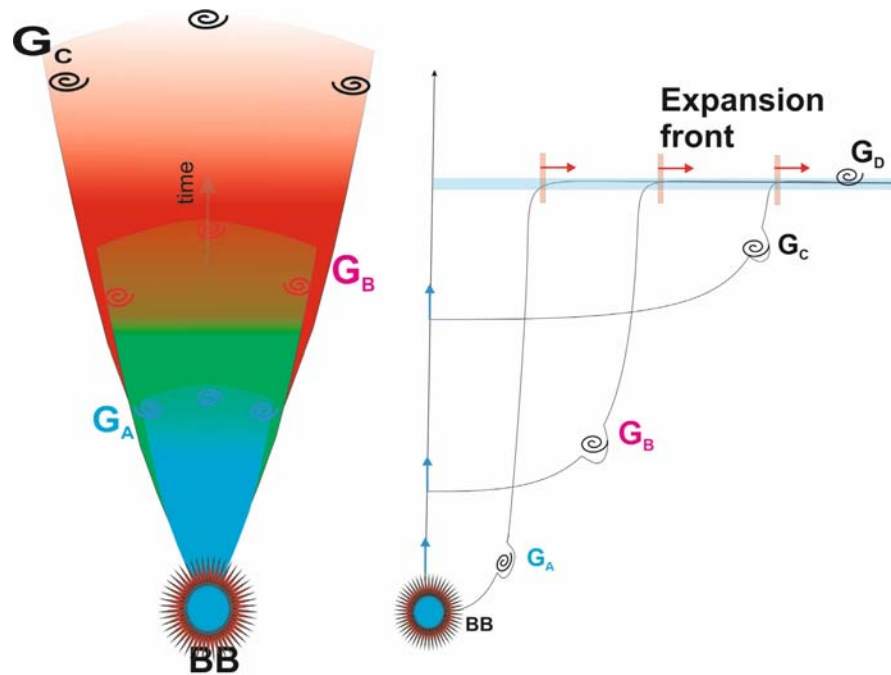


Figure 5. A schematic description of the three types of galaxies that may affect the time-evolution of the Hubble “constant.” G_A are galaxies that were formed from the newly created normal matter in the central region of the fireball. The matter here must have been anchored deeply in the gravitational potential-well of the fireball and therefore have relatively low kinetic energy. G_B are the old inactive galaxies that have been hit by the fireball, absorbed an enormous amount of mass and momentum and were forced to accelerate their radially oriented motion relatively fast and later became to high redshift galaxies. G_C are relatively distant old and inactive galaxies that have been hit by the fireball and subsequently turned into an active mode whilst accelerating their outward velocities. G_D are galaxies that populate the parent universe and are beyond the particle horizon.

$$H = \frac{dV}{dD} = \frac{d\beta_f}{dt}. \quad (14)$$

In **Figure 4**, $H(t)$ is plotted for different injection rates between $[10^{-12} \leq F \leq 10^{-11}]$. Here dV/dD is not really a constant but rather a slowly varying function of time. In the early universe, $H(t)$ had higher values than in the late universe. This suggests that when the normal matter of the expanding fireball hit the first G_B -type galaxies, it set them into an acceleration mode relatively early, and they currently should have reached higher redshifts than their distant counterparts. In fact, their close distance to the BB event makes their effective cross-sections relatively larger, and therefore the possibility of absorbing normal matter and momentum would be much higher. For causality reasons, this trend must change: extremely high redshift galaxies cannot continue increasing their velocities with distance indefinitely but eventually converge to a constant velocity, namely to the incident velocity V_f . On the other hand, those initially distant galaxies would continue increasing their velocities with distance as far as $V < V_f$, though at much longer time scales. The crossover of $H(t)$ is depicted with a star in **Figure 4**.

3. Summary and Conclusions

In this paper, two additional aspects of UNIMOUN have been addressed, namely:

- UNIMOUN accommodates nicely the observational fact that the galaxies in the local universe recede at much slower rates than their distant counterparts?
- The Hubble parameter is a slowly varying function with the cosmic time: it starts with higher values in the early universe and decreases toward small ones in the far future, which is a causally consistent trend.

Based on the time-dependent $\mathcal{H}$ -metric, both the analytical and GR-hydrodynamical calculations confirm that the newly created normal matter in the innermost region of the fireball was gravitationally bound. Once the matter has cooled down, the contraction would lead to the formation of galaxies that we observe today in the local universe. Nonetheless, the expansion of spacetime forces these galaxies to recede at relatively low rates, which is a consequence of the minimum energy theorem applied to a fireball with a finite mass.

Our study also confirms that the Hubble parameter is, in fact, a slowly varying function of cosmic time. Its value in the early times must have been higher than in the late universe, which is a self-consistent behaviour in UNIMOUN. Here the matter and momentum transferred from the fireball to G_B -type galaxies must have been sufficient to accelerate their motions up to relativistic velocities observed today as high redshift galaxies. The reason, therefore, is that although the spacetimes inside these galaxies are locally curved, the ambient parent spacetime is flat. Therefore, due to momentum conservation, any external bombardments of these levitating galaxies in the flat parent universe must set them into accelerating motions.

We note here that the “early universe” in Λ CDM-cosmologies differs fundamentally from that in UNIMOUN: While extremely high redshift galaxies in Λ CDM-cosmologies correspond to those that were formed shortly after the BB, in UNIMOUN, however, they may correspond to old and inactive galaxies that populated the ambient space of the BB, but started receding due to bombardment by the fireball to be seen today as very high redshift galaxies. In fact, the data from the CMB with/without Planck gives rise to a Hubble constant of roughly $67 \text{ km}\cdot\text{s}^{-1}\cdot\text{Mpc}^{-1}$, which is lower than the value revealed by Cepheids-SNIa measurements. This implies that H_0 would increase rather than decrease with cosmic time as predicated by UNIMOUN. Recalling the sensitivity of the CMB-Planck data to the underlying cosmological model, we conclude that UNIMOUN is a viable alternative to Λ CDM-cosmologies.

Acknowledgements

The author acknowledges the financial support of the IWR and KAUST. Max Camenzind, Mario Livio and Adi Nasser are acknowledged for their valuable comments and suggestions.

Conflicts of Interest

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

References

- [1] Bahcall, N.A. (2015) *PNAS*, **112**, 3173-3175.
<https://doi.org/10.1073/pnas.1424299112>
- [2] Di Valentino, E., Mena, O., *et al.* (2021) *Classical and Quantum Gravity*, **38**, Article ID: 153001.
- [3] Dianotti, M.G., De Simone, B., *et al.* (2021) *The Astrophysical Journal*, **912**, 150.
<https://doi.org/10.3847/1538-4357/abeb73>
- [4] Riess, A.G., *et al.* (2019) *The Astrophysical Journal*, **876**, 85.
<https://doi.org/10.3847/1538-4357/ab1422>
- [5] Aghanim, N., *et al.* (2020) *A&A*, **641**, A5-A6.
- [6] Carroll, S.M. (2001) *LRR*, **4**, 1.
- [7] Perlmutter, S. *et al.* (1999) *The Astrophysical Journal*, **517**, 565.
- [8] Sato, K. (1981) *Monthly Notices of the Royal Astronomical Society*, **195**, 467-479.
<https://doi.org/10.1093/mnras/195.3.467>
- [9] Steinhardt, P.J. (2011) *Scientific American*, **304**, 18.
<https://doi.org/10.1038/scientificamerican0311-18b>
- [10] Hujeirat, A.A. (2021) *Journal of Modern Physics*, **12**, 7.
- [11] Hujeirat, A.A. and Samtaney, R. (2019) *Journal of Modern Physics*, **10**, 1696-1712.
<https://doi.org/10.4236/jmp.2019.1014111>
- [12] Hujeirat, A. and Samtaney, R. (2020) *Journal of Modern Physics*, **11**, 1779-1784.
<https://doi.org/10.4236/jmp.2020.1111110>
- [13] Eskola, K.J. (2019) *Nature Physics*, **15**, 1111-1112.
<https://doi.org/10.1038/s41567-019-0643-0>
- [14] Espinoza, C.M., Lyne, A.G., Stappers, B.W. and Kramer, C. (2011) *Monthly Notices of the Royal Astronomical Society*, **414**, 1679-1704.
<https://doi.org/10.1111/j.1365-2966.2011.18503.x>
- [15] Roy, J., Yashwant Gupta, Y. and Lewandowski, W. (2012) *Monthly Notices of the Royal Astronomical Society*, **424**, 2213-2221.
<https://doi.org/10.1111/j.1365-2966.2012.21380.x>
- [16] Abbott, *et al.* (2017) *The Astrophysical Journal*, **848**, L12.
- [17] Hujeirat, A.A. (2018) *Journal of Modern Physics*, **9**, 4.
- [18] Camenzind, M. (2007) *Compact Objects in Astrophysics*. Springer, Heidelberg.
- [19] Hujeirat, A.A. and Samtaney, R. (2020) *Journal of Modern Physics*, **11**, 395-406.
<https://doi.org/10.4236/jmp.2020.113025>
- [20] Hujeirat, A.A. (2022) *Journal of Modern Physics*, **13**, 1581-1597.
<https://doi.org/10.4236/jmp.2022.1312096>
- [21] Brousso, R. (2002) *Reviews of Modern Physics*, **74**, 825-874.
<https://doi.org/10.1103/RevModPhys.74.825>
- [22] Trimble, V. (1987) *Annual Review of Astronomy and Astrophysics*, **25**, 425-472.
<https://doi.org/10.1146/annurev.aa.25.090187.002233>

- [23] Hujeirat, A.A. (2022) *Journal of Modern Physics*, **13**, 1474-1498.
<https://doi.org/10.4236/jmp.2022.1311091>
- [24] Hujeirat, A.A. (2005) *Computer Physics Communications*, **168**, 1-24.
<https://doi.org/10.1016/j.cpc.2005.01.013>
- [25] Fischer, M.S. and Hujeirat, A.A. (2020) Time-Implicit Schemes in Fluid Dynamics?— Their Advantage in the Regime of Ultra-Relativistic Shock Fronts.
<https://ui.adsabs.harvard.edu/abs/2020arXiv200412310F/abstract>
- [26] Verlinde, E.P. (2010) *Journal of High Energy Physics*, **1104**, 29.

Electron Shape Calculated for the Dual-Charge Dual-Mass Model

Arlen Young

Independent Researcher, Palo Alto, CA, USA

Email: arlen_young@yahoo.com

How to cite this paper: Young, A. (2023) Electron Shape Calculated for the Dual-Charge Dual-Mass Model. *Journal of Modern Physics*, 14, 198-207.
<https://doi.org/10.4236/jmp.2023.143014>

Received: January 5, 2023

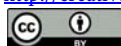
Accepted: February 5, 2023

Published: February 8, 2023

Copyright © 2023 by author(s) and Scientific Research Publishing Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

A model for the internal structure of the electron using classical physics equations has been previously published by the author. The model employs both positive and negative charges and positive and negative masses. The internal attributes of the electron structure were calculated for both ring and spherical shapes. Further examination of the model reveals an instability for the ring shape. The spherical shape appears to be stable, but relies on tensile or compressive forces of the electron material for stability. The model is modified in this document to eliminate the dependency on material forces. Uniform stability is provided solely by balancing electrical and centrifugal forces. This stability is achieved by slightly elongating the sphere along the spin axis to create a prolate ellipsoid. The semi-major axis of the ellipsoid is the spin axis of the electron, and is calculated to be 1.20% longer than the semi-minor axis, which is the radius of the equator. Although the shape deviates slightly from a perfect sphere, the electric dipole moment is zero. In the author's previously published document, the attributes of the internal components of the electron, such as charge and mass, were calculated and expressed as ratios to the classically measured values for the composite electron. It is interesting to note that all of these ratios are nearly the same as the inverse of the Fine Structure Constant, with differences of less than 15%. The electron model assumed that the outer surface charge was fixed and uniform. By allowing the charge to be mobile and the shape to have a particular ellipticity, it is shown that the calculated charge and mass ratios for the model can be exactly equal to the Fine Structure Constant and the Constant plus one. The electron radius predicted by the model is 15% greater than the Classical Electron Radius.

Keywords

Electron Shape, Classical Electron Model, Dual-Charge Dual-Mass Model, Electron Radius, Negative Mass, Electron Mass Inconsistency, Electron Charge Inconsistency, Fine Structure Constant

1. Introduction

Reference [1] addresses a great inconsistency between the measured electron spin magnetic dipole moment and the moment calculated from spinning the electron charge. It also addresses a similar inconsistency between the spin angular momentum derived from quantum theory and the classical momentum calculated from spinning the electron mass. The model presented in [1] is referred to herein as the dual-charge dual-mass (DCDM) model. The internal attributes for the model were calculated for both ring and spherical shapes. The ring shape appears to have an instability along the spin axis, while the spherical shape depends on the electron material tensile or compressibility properties for stability.

The ring shape has an outer shell of positively-charged negative mass and a central core of negatively-charged positive mass. Both lie on the equatorial plane. If a perturbation were to displace the central core along the spin axis, an unbalanced attractive electrical force would be created between the positive and negative masses. Since negative mass moves in a direction opposite to an applied force, the outer shell is repelled from the central core, creating an even stronger attractive force between the two. As a result, the outer shell and central core would fly apart.

The spherical shape appears to be stable for perturbations in all directions. It does, however, depend on the tensile or compressibility properties of the negative mass electron material. The intent of this document is to remove this dependency by slightly altering the spherical shape. Stability along the spin axis will be provided by only the internal electrical force between the outer shell and central core. There will be no tensile or compressive forces anywhere on the outer shell.

Table 1 contains electron constants that will be used in this document. Ratios of outer shell and central core charges and masses to the measured charge and mass of the composite electron are presented in **Table 2** below. For comparison,

Table 1. Table of electron constants.

constant	symbol	value [cgs] [1]
charge	q	-4.8032×10^{-10}
outer shell charge for DCDM sphere	q^+	7.507×10^{-8}
central core charge for DCDM sphere	q^-	-7.555×10^{-8}
classical electron radius	R	2.82×10^{-13}
radius of DCDM sphere	R_q	3.711×10^{-13}
mass	m	9.1094×10^{-28}
spin angular momentum	S	9.1329×10^{-28}
magnetic dipole moment	M	$-9.284764 \times 10^{-21}$
fine structure constant inverse	$1/\alpha$	137.0360 [2]
speed of light	c	$2.99792458 \times 10^{10}$

Table 2. Table of internal/external attribute ratios [1].

attribute	ring	sphere
$\frac{\text{charge radius}}{\text{classical electron radius}}$	1.07	1.32
$\frac{\text{outer shell mass}}{\text{electron mass}} = \frac{\text{outer shell charge}}{\text{electron charge}}$	-127.6	-156.3
$\frac{\text{central core mass}}{\text{electron mass}} = \frac{\text{central core charge}}{\text{electron charge}}$	128.6	157.3
$\frac{1}{\text{fine structure constant}}$ [3]	137.0	137.0

the Fine Structure Constant is also included in the table. It can be seen that for a spherical shape, the charge and mass ratio values are close to the Fine Structure Constant inverse value, with differences of less than 15%. An approximate calculation for an elliptical shape shows that the attribute ratios could actually be exactly equal to the inverse of the Fine Structure Constant and the inverse of the Constant plus one.

Except where otherwise noted, all constants and equations in this article are expressed in cgs units.

2. Stretch the Spherical Shape to an Ellipsoid

2.1. Introduction

In the equatorial plane of the DCDM electron model, the centrifugal force is balanced by the attractive electrical force between the outer shell and central core and the repulsive electrical force on the outer shell. Along the spin axis, the same electrical forces exit, but there is no counterbalancing centrifugal force. By stretching the sphere slightly along the spin axis, two things happen:

- The radius of curvature of the surface at the pole is increased, thereby increasing the repulsive electrical force on the outer shell at that point;
- The distance between the central core and the outer shell at the pole is increased, thereby reducing the attractive force between the two.

An ellipsoid, or ellipse of revolution, seems like a reasonable approximation for the shape required by the model.

2.2. Strategy

The parameters for an ellipsoid are to be calculated for which the electrical forces on the electron surface at the spin axis pole are balanced. Thus, there will be no force on the surface at the pole. The two forces to be balanced are

- The attractive force on the outer shell by the central core, and
- The repulsive force created by the outer shell's charge.

The repulsive electric field at the ellipsoid pole is calculated by slicing the ellipsoid into many circle rings, all having axes along the spin axis. The charge on the surface of each ring is first calculated. Then, the electric field along the spin

axis at the pole is calculated for each charged ring. Finally, the fields for all rings are added together, and from this sum, the ellipsoid parameters are calculated.

2.3. Ellipsoid Ring Charge

The equation for an ellipse is:

$$\left(\frac{x}{A}\right)^2 + \left(\frac{y}{B}\right)^2 = 1 \quad \text{or} \quad y = B\sqrt{1 - \left(\frac{x}{A}\right)^2}$$

where x = major semi-axis which aligns with the electron spin axis,

y = minor semi-axis which lies in the electron equatorial plane,

A is the intersection of the ellipse (pole) with the spin axis, and

B is the intersection of the ellipse with the equatorial plane.

The ellipsoid is formed by revolving the ellipse about its major axis. To calculate the difference between A and B required to provide a balance of electrical forces at the pole, the ellipsoid is divided into many circular rings, all having axes along the spin axis. The radius of each ring is y , and the location of the ring from the electron center is x . To calculate the charge on each ring, the width Δc of each ring is first calculated.

$$(\Delta c)^2 = (\Delta x)^2 + (\Delta y)^2$$

where Δx and Δy are the projections of Δc on the x and y axes, respectively.

$$y - \Delta y = B\sqrt{1 - \left(\frac{x + \Delta x}{A}\right)^2}$$

$$\frac{y\Delta y}{B^2} = 1 - \left(\frac{x}{A}\right)^2 - \sqrt{\left[1 - \left(\frac{x}{A}\right)^2\right]\left[1 - \left(\frac{x + \Delta x}{A}\right)^2\right]}$$

$$y\Delta c = \left[(y\Delta x)^2 + (y\Delta y)^2\right]^{1/2} = \left[B^2(\Delta x)^2\left(1 - \left(\frac{x}{A}\right)^2\right) + (y\Delta y)^2\right]^{1/2}$$

$\Delta q \equiv$ ring charge

$\sigma \equiv$ charge surface density

$$\frac{\Delta q}{A^2} = \frac{2\pi\sigma y\Delta c}{A^2}$$

For the following equation expansions,

$$B = A + \Delta A$$

$$\left(\frac{B}{A}\right)^2 = \left(1 + \frac{\Delta A}{A}\right)^2 = 1 + 2\frac{\Delta A}{A} + \left(\frac{\Delta A}{A}\right)^2 = 1 + 2\frac{\Delta A}{A}$$

It is assumed that $\frac{\Delta A}{A} \ll 1$, so terms containing higher orders of $\frac{\Delta A}{A}$ are deleted. The final evaluation of $\frac{\Delta A}{A}$ will validate this assumption. The expanded

equation for $\frac{\Delta q}{A^2}$ is

$$\frac{\Delta q}{A^2} = 2\pi\sigma\sqrt{k_1}\left[1 + (k_2 + 1)\frac{\Delta A}{A}\right]$$

$$\text{where } k_1 = \left[1 - \left(\frac{x}{A}\right)^2\right]\left(\frac{\Delta x}{A}\right)^2 + \left(\frac{y\Delta y}{B^2}\right)^2 \text{ and } k_2 = \frac{\left(\frac{y\Delta y}{B^2}\right)^2}{k_1}.$$

2.4. Distance from A Ring Charge to Pole A

The distance d between the charge on a ring and a polar point A is

$$d = \left[(A-x)^2 + y^2\right]^{1/2} = A^2\left(1 - \frac{x}{A}\right)^2 + B^2\left[1 - \left(\frac{x}{A}\right)^2\right].$$

$$\frac{d}{A} = \sqrt{2}\left[1 - \frac{x}{A} + \left[1 - \left(\frac{x}{A}\right)^2\right]\frac{\Delta A}{A}\right]^{1/2}$$

$$\left(1 - \frac{x}{A}\right)\left(\frac{d}{A}\right)^{-3} = 2^{-3/2}\left(1 - \frac{x}{A}\right)^{-1/2}\left[1 - \frac{3}{2}\frac{1 - \left(\frac{x}{A}\right)^2}{1 - \frac{x}{A}}\frac{\Delta A}{A}\right]$$

2.5. Electric Field at Pole A from A Single Ring

The electric field along the spin axis at a pole A due to a single ring is

$$\frac{\Delta q}{d^2}\left(\frac{A-x}{d}\right) = \frac{\Delta q}{A^2}\left(1 - \frac{x}{A}\right)\left(\frac{d}{A}\right)^{-3} = \frac{\pi\sigma}{\sqrt{2}}k_3\left[1 + k_4\frac{\Delta A}{A}\right]$$

$$\text{where } k_3 = \sqrt{k_1}\left(1 - \frac{x}{A}\right)^{-1/2} \text{ and } k_4 = 1 + k_2 - \frac{3}{2}\frac{1 - \left(\frac{x}{A}\right)^2}{1 - \frac{x}{A}}.$$

2.6. Electric Field at Pole A from All of the Rings

The following electric fields along the spin axis at pole A are defined:

E_e^+ $\equiv$ repulsive field created by the ellipsoid surface charge q^+
 $\equiv$ the sum of the electric fields due to all of the ellipsoid rings. $\Delta A \neq 0$.

E_s^+ $\equiv$ repulsive field created by a spherical shell having a radius A and a surface charge q^+

$$= E_e^+ \text{ for } \Delta A = 0 = \frac{q^+}{A^2}$$

E_s^- $\equiv$ attractive field created by a charge q at the center of the sphere of radius A

$$= \frac{|q^-|}{A^2}$$

For the DCDM electron model with a spherical shape, the attractive and repulsive forces on the outer shell at the pole A are unbalanced. The ratio of the two fields at A is

$$\frac{E_s^+}{E_s^-} = \frac{\frac{q^+}{A^2}}{\frac{|q^-|}{A^2}} = \frac{q^+}{|q^-|} \quad \text{and} \quad E_s^- = \frac{|q^-|}{q^+} E_s^+$$

The sphere is stretched along the spin axis x to an ellipsoidal shape to balance the attractive and repulsive electrical fields, so that

$$\frac{E_e^+}{E_s^-} = 1 \quad \text{and} \quad E_s^- = E_e^+ = \frac{|q^-|}{q^+} E_s^+$$

The following equation was solved numerically for ΔA with the ellipsoid sliced into 200 rings:

$$\frac{E_e^+}{E_s^+} = \frac{|q^-|}{q^+} = \frac{\sum_{x=-A}^{x=A} E_e^+ (\Delta A \neq 0)}{\sum_{x=-A}^{x=A} E_e^+ (\Delta A = 0)} = \frac{\sum_{x=-A}^{x=A} k_3 \left(1 + k_4 \frac{\Delta}{A}\right)}{\sum_{x=-A}^{x=A} k_3}$$

2.7. Ellipticity of the Electron Surface

The solution to the above equation for $\frac{\Delta A}{A}$ is:

$$\frac{\Delta A}{A} = -\frac{\sum_{x=-A}^{x=A} k_3 \left(\frac{|q^-|}{q^+} - 1\right)}{\sum_{x=-A}^{x=A} k_4} = -0.0119$$

Note that $\frac{\Delta A}{A}$ is negative, so the ellipsoid type is prolate, *i.e.*, the distance from the center to the intersection of the electron spin axis with its surface (pole A) is 1.20% longer than the distance from the center to the intersection of the equatorial plane with the surface (R_q).

3. Outer Shell Charge Distribution

3.1. Introduction

The mysterious Fine Structure Constant [3] appears in many instances in physics, for example [4] [5] [6], although it is not completely understood why. It is a unitless constant, and is the ratio of other constants having the same unit. It has been suggested [7] that it is one of the most, and maybe the most, fundamental constant in physics.

Table 2 above lists the ratios of attributes of the electron model's internal components, as calculated according the dual-charge dual-mass model. The table also includes the value of the inverse of the Fine Structure Constant. There is a striking similarity between the attribute ratios and the Constant. In the worst case, the attribute ratio differs from the Constant by less than 15%. Possibly, the attribute ratio should actually equal the Constant, or at least have a value very much closer to it.

Attribute ratios are listed in the table for both ring and spherical electron model shapes. The attribute ratios for a ring shape are less than the Constant, while those for the spherical shape are greater. It has been assumed in the electron model that the charge density on the outer shell surface is uniform. But maybe it isn't. The fact that the Constant value lies between the ring and sphere attribute ratio values suggests that the charge density might actually be greater near the equator than elsewhere. By changing the surface charge profile from uniform to one of increasing charge density toward the equator, it should be possible to achieve attribute ratios that are within $\pm 0.4\%$ of the Constant value.

3.2. Surface Charge Distribution

For a perfectly spherical outer shell charge, the central core charge exerts a radial force on each charge increment. There is no tangential component to the force. The outer shell charge may be mobile or fixed. It makes no difference. If mobile, the mutually repulsive forces between the charge increments will cause them to spread out uniformly across the shell surface. However, if the outer shell is not perfectly spherical and in fact is a prolate ellipsoid, there will be a tangential force exerted by the central core on each charge increment, which tends to pull it toward the equator. (The exception is at exactly the equator and the poles.) This force is further increased by the lesser distance between the outer shell increment and the central core for increments closer to the equator than farther away. Therefore, for a prolate ellipsoidal shape and a mobile surface charge, the surface charge density will increase from the pole to the equator. The greater the ellipticity, the greater the concentration of charge near the equator.

In reality, the ellipsoid's charge density near the equator would smoothly taper down toward the poles. A rough approximation to a realistic charge profile is an overlay of a ring charge on a sphere at the equator. The relation between the ring charge value and the sphere's charge value is calculated in the following. Of course, the values calculated for this composite are approximate, but provide a general idea of how the electron's internal attribute ratios could be brought closer to the fine structure constant.

3.3. Calculation of Approximate Charge Distribution

A ring-shaped outer shell is superimposed on the spherical outer shell at the equator. The two shells have the same radius r . The charge on the ring is kq^+ , and $(k-1)q^+$ on the sphere. The scaling factor k is calculated as follows:

The magnetic moments for the ring and sphere are

$$M(\text{ring}) = -\frac{kq^+}{2}r, \text{ and } M(\text{sphere}) = -\frac{(k-1)q^+}{3}r. \quad [1]$$

The composite magnetic moment is

$$M = M(\text{ring}) + M(\text{sphere}) = -\frac{(k+2)q^+}{6}r,$$

where the radius r is

$$r = -\left(\frac{m^-}{m}\right) \frac{q^4}{mc^3 S} \quad [1]$$

Assume that

$$-\frac{q^+}{q} = -\frac{m^-}{m} = \frac{1}{\text{fine structure constant}} = 137$$

and therefore

$$\frac{q^-}{q} = \frac{m^+}{m} = \frac{1}{\text{fine structure constant}} + 1 = 137 + 1.$$

By merging and rearranging the above equations,

$$k = \frac{6Mmc^3 S}{137^2 q^5} - 2 = 0.60$$

$$\text{ring charge} = 0.60q^+$$

$$\text{sphere charge} = 0.40q^+$$

3.4. Radius

The radius for a ring-shaped outer shell in the dual-charge dual-mass model can be calculated from

$$R(\text{ring}) = \sqrt{\frac{2M}{mS} \left(\frac{q}{c}\right)^3} \quad [1]$$

The radius for the ring and sphere charge composite can be calculated by replacing “ $2M$ ” with $\frac{6}{k+2}M$ in the above equation.

$$R(\text{ring} + \text{sphere}) = 3.255 \times 10^{-13} = 1.15R$$

Therefore, by allowing the outer shell charge to be mobile on an ellipsoidal surface and assuming the outer shell charge and mass ratios to be equal to the inverse of the Fine Structure Constant, the resulting electron radius is only 15% greater than the Classical Electron Radius.

4. Summary

Reference [1] describes a model for the internal structure of the electron. Attributes for two possible shapes, a ring and a sphere, were calculated. The purpose of this document is to explain a stability problem for the ring shape and to modify the spherical shape slightly such that its stability is not dependent upon the tensile or compression properties of the electron material. The calculations presented in this document show that this dependency can be eliminated for a shape close to that of a prolate ellipsoid having a length 1.20% greater along its spin axis than its width in the equatorial plane. The net force at every point on the outer shell surface will be zero.

Table 3. Table of electron model attributes.

attribute	outer shell charge	
	fixed, uniform	mobile
$\frac{\text{outer shell mass}}{\text{electron mass}} = \frac{\text{outer shell charge}}{\text{electron charge}}$	-156.3	-137.0
$\frac{\text{central core mass}}{\text{electron mass}} = \frac{\text{central core charge}}{\text{electron charge}}$	157.3	137.0 + 1.0
$\frac{\text{central core mass}}{\text{outer shell mass}} = \frac{\text{central core charge}}{\text{outer shell charge}}$	$-\left(1 + \frac{1}{156.3}\right)$	$-\left(1 + \frac{1}{137.0}\right)$
$\frac{1}{\text{fine structure constant}}$	137.0	137.0
$\frac{\text{charge radius}}{\text{classical electron radius}}$	1.32	1.15
ellipticity	1.012	1.035

Ratios of the charge and mass attributes for the electron model were calculated in [1] assuming a fixed and uniform surface charge, and are summarized in **Table 3**. It is interesting to note that all of these ratios are nearly equal to the inverse of the Fine Structure Constant, with differences of less than 15%. A model of the surface charge has been calculated that shows if the charge is mobile and more concentrated near the equator than elsewhere, the charge and mass attribute ratios can be exactly equal to the inverse of the Fine Structure Constant (137.0) and the inverse plus one (138.0). These results are summarized in **Table 3**. An explanation is presented which describes how the ellipsoidal shape of the outer core and charge mobility could result in such a charge concentration.

The ellipticity in this table is defined as the distance along the spin axis from the center to a pole divided by the radius at the equator.

Conflicts of Interest

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

References

- [1] Young, A. (2022) *Journal of Modern Physics*, **13**, 1287-1294.
<https://doi.org/10.4236/jmp.2022.139077>
- [2] NIST (2018) CODATA Value: Fine-Structure Constant. The NIST Reference on Constants, Units, and Uncertainty. 20 May 2019.
<https://physics.nist.gov/cgi-bin/cuu/Value?alph>
- [3] Feynman, R.P. (1985) QED: The Strange Theory of Light and Matter. Princeton University Press, Princeton, p. 129.
- [4] Riazuddin, F. (2012) A Modern Introduction to Particle Physics. 3rd Edition, World Scientific, Singapore, p. 4.
- [5] Bedford, D. and Krumm, P. (2004) *American Journal of Physics*, **72**, 969-970.

<https://doi.org/10.1119/1.1646135>

- [6] Poelz, G. (6 October 2021) An Electron Model with Synchrotron Radiation. arXiv: 1206.0620v36 [physics.class-ph].
- [7] Hawking, S. (1988) A Brief History of Time. Bantam, New York, p. 7, 125.

Average Luminous Mass of Early Galaxies at $z = 10 - 20$ Predicted as $\sim 10^9$ Solar Masses

T. R. Mongan

Sausalito, CA, USA

Email: tmongan@gmail.com

How to cite this paper: Mongan, T.R. (2023) Average Luminous Mass of Early Galaxies at $z = 10 - 20$ Predicted as $\sim 10^9$ Solar Masses. *Journal of Modern Physics*, 14, 208-211.

<https://doi.org/10.4236/jmp.2023.143015>

Received: January 7, 2023

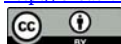
Accepted: February 5, 2023

Published: February 8, 2023

Copyright © 2023 by author(s) and Scientific Research Publishing Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

This paper predicts the average luminous mass of galaxies that will be detected by JWST space telescope at redshift $z \approx 10 - 20$. The prediction, derived in the paper, is based on holographic analysis, developed from quantum mechanics, general relativity, thermodynamics, and Shannon information theory. Consistent with early JWST data, $\sim 10^9$ solar masses is the predicted average luminous mass of early galaxies at $z \approx 10 - 20$ that will be detected by JWST.

Keywords

Galaxies: High-Redshift

1. Introduction

Luminous galaxy candidates at $z > 10$ revealed by JWST indicate “future deep JWST observations may identify relatively bright galaxies to much earlier epochs than might have been expected” [1]. As explained below, holographic analysis based on quantum mechanics, general relativity, thermodynamics, and Shannon information theory indicates galaxies with mass $\sim 10^9$ solar masses should be expected at $z = 10 - 20$. The analysis does not assume or require a theory specifically relating arrangements of bits of information on holographic screens to distribution (as opposed to amount) of mass within such screens.

Behroozi *et al.* [2] listed some early predictions for what JWST will see, and then used a supercomputer simulation to predict JWST will find galaxies $z > 10$ with masses considerably below those revealed by early JWST data. In contrast, the prediction from holographic analysis below, $\sim 10^9$ solar masses for average luminous mass of galaxies at $z \approx 10 - 20$, is consistent with early JWST results.

2. Holographic Analysis

Our post-inflationary universe, dominated by vacuum energy, has cosmological constant $\Lambda = 1.088 \times 10^{-56} \text{ cm}^{-2}$ [3] and event horizon radius

$R_H = \sqrt{\frac{3}{\Lambda}} = 1.661 \times 10^{28} \text{ cm}$. Holographic analysis [4] finds only a finite number

$N = \frac{\pi}{\ln(2)} \left(\frac{R_H}{l_p} \right)^2 = 4.741 \times 10^{122}$ of bits of information on the event horizon are available to describe our universe, where Planck length

$$l_p = \sqrt{\frac{\hbar G}{c^3}} = 1.616 \times 10^{-33} \text{ cm}.$$

Friedmann's general relativistic equation $H_0^2 = \frac{8\pi G}{3} \rho_{crit} + \frac{\Lambda c^2}{3}$ for a flat Euclidean universe with critical density and cosmological constant requires

$\Omega_\Lambda \equiv \frac{\Lambda c^2}{3H_0^2}$. PDG 2022 [3] lists present day Hubble expansion rate

$H_0 = 67.4 \text{ km}/(\text{sec} \cdot \text{Mpc})$, critical density $\rho_{crit} = \frac{3H_0^2}{8\pi G} = 8.53 \times 10^{-30} \text{ g}/\text{cm}^3$, dark

energy density parameter $\Omega_\Lambda = 0.685$, Hubble length $c/H_0 = 1.37 \times 10^{28} \text{ cm}$, and

matter density $\rho_m = (1 - \Omega_\Lambda) \rho_{crit}$. PDG parameters result in $\frac{\Lambda c^2}{3\Omega_\Lambda H_0^2} = 0.997$,

so our universe is indistinguishable from flat Euclidean space to three significant figures. Mass within the event horizon

$M_H = \frac{4}{3} \pi (1 - \Omega_\Lambda) \rho_{crit} R_H^3 = \frac{(1 - \Omega_\Lambda) c^2}{2G} \sqrt{\frac{3}{\Lambda}} = 5.14 \times 10^{55} \text{ g}$ is constant in time,

and $M_H = \left[\frac{(1 - \Omega_\Lambda) c^2}{2G} \left(\frac{H_0}{c} \right)^2 \sqrt{\frac{3}{\Lambda}} \right] R_H^2 = (0.187 \text{ g}/\text{cm}^2) R_H^2$. The constant mass

per bit of information $m_{bit} = M_H / N = 1.08 \times 10^{-67} \text{ g}$.

In a fundamental sense, information specifies location of matter in space, and holographic analysis indicates 4.741×10^{122} bits of information on the spherical holographic screen (SHS) of the event horizon are associated with matter within our observable universe. Holographic analysis then indicates information and associated mass M within isolated gravitationally bound systems relates to radii R of spherical holographic screens around system centers of mass by

$$M = (0.187 \text{ g}/\text{cm}^2) R^2.$$

Cosmic microwave background (CMB) radiation density at redshift z is

$\rho_r(z) = (1+z)^4 \rho_r(0)$, where mass equivalent of today's radiation energy density $\rho_r(0) = 4.59 \times 10^{-34} \text{ g}/\text{cm}^3$. Matter density $\rho(z)$ is much greater than radiation

density and Jeans mass [5] $M_J = \frac{\pi}{48\rho_m^2} \left[\frac{2c}{3} \sqrt{\frac{\pi\rho_r(0)}{G}} \right]^3 = 2.30 \times 10^{50} \text{ g}$, the upper

limit on mass of gravitationally bound systems stable against gravitational collapse, is independent of z . Large scale structures at $z > 10$ are gravitationally bound systems of individual stars with masses between Jeans mass M_J and

minimum stellar mass $m_{*min}(z)$ at redshift z .

3. Minimum Stellar Mass at Redshift z

Minimum stellar mass $m_{*min}(z)$ is estimated by setting escape velocity of protons at SHS radius R_{*min} for minimum stellar mass equal average velocity of protons in equilibrium with CMB radiation outside the SHS for $m_{*min}(z)$. Protons in equilibrium with CMB outside the SHS for stellar systems with mass $< m_{*min}(z)$ can transfer energy to those systems until they reach $m_{*min}(z)$. Escape velocity v for protons of mass m_p gravitationally bound at radius R from system center of mass M is calculated from $\frac{1}{2}m_p v^2 = \frac{GMm_p}{R}$. Escape velocity of protons on the SHS for minimum mass stars with mass M at redshift z is velocity of protons in thermal equilibrium with CMB, so $\frac{3}{2}kT(z) = \frac{GMm_p}{R}$,

where CMB temperature $T(z) = (1+z)2.725$ K and Boltzmann constant $k = 1.38 \times 10^{-16} (\text{g} \cdot \text{cm}^2 / \text{sec}^2) / \text{K}$. With radius $R = \sqrt{M / (0.187 \text{ g/cm}^2)}$ for

structures of mass M , $m_{*min}(z) = \frac{1}{0.187} \left(\frac{1.5k(1+z)2.725}{Gm_p} \right)^2 \text{ g}$, so

$m_{*min}(10) = 8.26M_\odot$, $m_{*min}(12) = 11.5M_\odot$, and $m_{*min}(20) = 30.1M_\odot$, with solar mass $M_\odot = 2 \times 10^{33} \text{ g}$. If outgoing protons at the SHS are in thermal equilibrium with outgoing photon flow from minimum mass stars, stars must have mass $> m_{*min}(z)$ to appear against the CMB background. Maximum star mass $6 \times 10^{35} \text{ g} \approx 300M_\odot$ [6] coincided with minimum star mass at $z \approx 65$, consistent with indications stars first formed at $z \approx 65$ [7]. At $z = 0$, minimum star mass $\approx 1.4 \times 10^{32} \text{ g} = 0.07M_\odot$, consistent with hydrogen burning mass threshold separating brown dwarfs from lowest mass stars [8].

4. Average Galactic Mass at Redshift z

With no good reason to suspect information describing gravitationally bound systems is not uniformly distributed between mass bins, the number of gravitationally bound systems of mass m at redshift z can be taken as $n(m) = K/m$ with constant K . Then total mass of observable gravitationally bound systems at redshift z is $M_H \approx \int_{m_{*min}(z)}^{M_J} m \left(\frac{K}{m} \right) dm = K(M_J - m_{*min}(z))$ and

$K \approx M_H / M_J = 2.24 \times 10^5$. The number of observable gravitationally bound structures in Jeans mass M_J at redshift z is

$N(z) = \int_{m_{*min}(z)}^{M_J} \frac{K}{m} dm = K \ln(2.30 \times 10^{50} \text{ g} / m_{*min}(z))$. Estimated average total mass

of observable gravitationally bound structures at redshift z ,

$M_{avg}(z) = M_J / N(z)$, at $z = 10 - 20$ is $\sim 1.4 \times 10^{10} M_\odot$, between estimated Milky Way mass $\sim 10^{45} \text{ g} \sim 10^{12} M_\odot$ and dwarf galaxy masses $\sim 10^{40} \text{ g} \sim 10^7 M_\odot$.

PDG 2022 [3] lists baryon density fraction of the universe $\Omega_b = 0.0493$, so stellar mass is $\Omega_b / (1 - \Omega_\Lambda) = 0.0720$ times total galactic mass. Then stellar mass

$\sim 10^9 M_{\odot}$ of galaxy candidates at redshift $z \approx 10 - 12$ identified by Naidu *et al.* in JWST data is consistent with “early appearance of UV-luminous galaxies with stellar masses as high as $\approx 10^9 M_{\odot}$ already at few 100 Myr after the Big Bang” [1].

5. Conclusion

Holographic analysis predicts the average galaxy detected by JWST space telescope at redshift $z \approx 10 - 12$ will have luminous mass of about 10^9 solar masses, consistent with early JWST data.

Conflicts of Interest

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

References

- [1] Naidu, R., Oesch, P., van Dokkum, P., *et al.* (2022) *The Astrophysical Journal Letters*, **940**, L14. <https://doi.org/10.3847/2041-8213/ac9b22>
- [2] Behroozi, P., Conroy, C., Wechsler, R., *et al.* (2020) *Monthly Notices of the Royal Astronomical Society*, **499**, 5702-5718. <https://doi.org/10.1093/mnras/staa3164>
- [3] Workman, R., Burkert, V., Crede, V., *et al.* (2022) *Progress of Theoretical and Experimental Physics*, **2022**, 083C01.
- [4] Bousso, R. (2002) *Reviews of Modern Physics*, **74**, 825-874. <https://doi.org/10.1103/RevModPhys.74.825>
- [5] Mongan, T. (2013) *Journal of Modern Physics*, **4**, 50-54. <https://doi.org/10.4236/jmp.2013.47A1006>
- [6] Crowther, P., Schnuur, O., Hirschi, R., *et al.* (2010) *Monthly Notices of the Royal Astronomical Society*, **408**, 731-751.
- [7] Naoz, S., Noter, S. and Barkana, R. (2006) *Monthly Notices of the Royal Astronomical Society: Letters*, **373**, L98-L102. <https://doi.org/10.1111/j.1745-3933.2006.00251.x>
- [8] Mireles, I., Shporer, A., Grievess, N., *et al.* (2020) *AJ*, **160**, 133.

An Outline of the Grand Unified Theory of Gauge Fields

Bi Qiao

Wuhan University of Technology, Wuhan, China

Email: biqiao@gmail.com

How to cite this paper: Qiao, B. (2023) An Outline of the Grand Unified Theory of Gauge Fields. *Journal of Modern Physics*, 14, 212-236.

<https://doi.org/10.4236/jmp.2023.143016>

Received: December 20, 2022

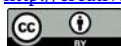
Accepted: February 10, 2023

Published: February 13, 2023

Copyright © 2023 by author(s) and Scientific Research Publishing Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

This paper attempts to propose a grand unified guiding principle of gauge fields from the mathematical and physical picture of fiber bundles: it is believed that our universe may have more fundamental interactions than the four fundamental interactions, and the gauge fields of these fundamental interactions are just a unified gauge potential on the fiber bundle manifold or the components connected to the bottom manifold, that is, our universe; these components can meet the transformation of gauge potential, and even can be transformed from a fundamental interaction gauge potential to another fundamental interaction gauge potential, and can be summarized into a unified equation, namely the expression of the generalized gauge equation, corresponding to the gauge transformation invariance; so gauge transformation invariance is a necessary condition to unify field theory, but quantization of field is not a necessary condition; the four (or more) fundamental interaction fields of the universe are unified into a universal gauge field defined by the connection of the principal fiber bundle on the cosmic base manifold.

Keywords

Gauge Field, Principal Fiber Bundle, Gauge Transformation Invariance, Grand Unified Theory of Physics

1. Introduction

In 1954, Yang Zhenning and Mills extended the $U(1)$ group local gauge symmetry to the $SU(2)$ group, which represents the local gauge symmetry of isotope spin in the non-Abelian group, and established the Yang-Mills gauge field theory [1]. Following the idea of gauge field theory, there is a gauge symmetry group $U(1) \times SU(2)$, which has achieved great success in unifying electromagnetic interaction and weak interaction. Such a theory has been proved to be con-

sistent with renormalization in the sense of spontaneous symmetry breaking Higgs mechanism [2] [3] [4] [5]. Moreover, the idea of gauge field theory provides a successful unified description of the dynamics (QCD) of gauge symmetry group $SU(3)$ [6] [7] of electric weak interaction and strong interaction characterized by quantum color, so as to construct a quantum gauge field theory dominated by symmetry, that is, gauge group $U(1) \times SU(2) \times SU(3)$ with spontaneous symmetry breaking, the standard model. This is a milestone achievement of gauge field theory in unifying electromagnetic, weak and strong fundamental interactions.

This led many scholars after Einstein to try to expand the gauge “quantum” field theory to the category of gravity, hoping to establish a grand unified theory of four fundamental interactions, including gravity and electromagnetic forces [8] [9] [10]. But until now, we may believe that the gravitational field is a gauge field, but the quantization theory of the gravitational field has always been inconsistent with the microscopic quantum field theory [11] [12] [13], this has also become an exciting point for the creation of the famous superstring theory [14] [15], which still inspires us to think about a question today: can we say for sure that gravity can be quantized?

The second question is whether there are more than four fundamental interactions in nature? It seems that no principle can limit the fundamental interactions in nature to four kinds, namely, gravitational, electromagnetic, and weak and strong interactions. Dark matter and dark energy put forward interpretations of this question from the perspective of astrophysics or cosmic scale [16] [17]. Are dark matter and dark energy the real existence or the representation of some unknown fundamental interactions? In fact, so far, many normative unified field theories have attempted to construct some direct product forms of very specific structural groups [18] [19] [20] [21]. These direct product forms often correspond to some specific basic particle fields, thus explaining the existing basic particles and astrophysical theories and experimental phenomena, which is like establishing the world’s unity on some specific basic particles. If they also can be decomposed, or can it decay? Or are there not only four fundamental interactions in the world (for example, dark energy does not exist, but a cosmic expansion caused by fundamental interactions on a cosmic scale)?

And what is the “over-distance” effect of quantum entanglement? The so-called force field interaction between two objects that we know is just the exchange of virtual particles between two objects, which seems to be invalid here, and the interaction seems to be a whole that occurs on two entangled objects at the same time [22] [23].

On the other hand, as the mathematical basis of the gauge field theory, the theory of principal fiber bundles has been greatly developed [24]. In a sense, the gauge theory of physics is just the principal fiber bundle (principal bundle) theory in mathematics [25]. The question is how to deeply understand or develop the relevant concepts or meanings of physics implied by the principal bundle theory? Thanks Professor Liang Canbin for teaching and building bridges in this

area over the years [26]. It is based on these “plays” and “bridges” that the author can consider the above-mentioned issues about the grand unification of physics in this paper, and try to propose an outline program for the grand unification of gauge fields from the mathematical physics picture of principal fiber bundles.

2. Basic Points of the Outline

Why is it so difficult to unify gravity and electromagnetic force among the four fundamental interactions? Why must the gauge transformation remain unchanged? Is there a unified outline program of cosmic gauge field? How to understand the “teleportation” between quantum entangled states? What is the interaction between entanglements? After a long time of meditation, the author of this paper proposes a “unified outline program for the gauge field of space and time in the universe”, the main points of which are as follows:

1) The whole universe structure can be simply described as the structure of a region ordinary principal bundle and associated bundle, which can be called “the principal associated bundles picture of the universe”: the universe is the bottom manifold M ; the structure group of the principal bundle is defined as a Lie group G (also a manifold), which reflects the laws or rules of the universe; then the fiber bundle of principal bundle, bottom manifold and group G form a direct product of manifold, which can be expressed as $P = G \times M$. It is recorded as $P(M, G)$. The fiber bundle associated to principal bundle (Associated bundle) is composed of manifold $Q(M, F)$ and fiber bundles determined by the coordination of P, M, G, F and related maps. The structure of the principal bundle and associated bundle can be shown by the following “principal associated bundles picture of the universe” (Figure 1).

Here, the precise definitions and mathematical and physical expressions of the mappings $\tau, \hat{\tau}, T_U, \hat{T}_U, \pi, \hat{\pi}$, etc., will be given one after another in the following sections.

2) The cross section of principal bundle $\sigma: M \rightarrow P$ represents a choice of gauge (embodied as an internal frame of gauge field); the transformation between two different cross sections of the principal bundle gives a gauge transformation, and vice versa. The cross section of associated bundle $\hat{\sigma}: M \rightarrow Q$ represents a physical gauge field (or elementary particle field) on the base manifold.

$$\begin{array}{ccc}
 & G \times F & \\
 \tau \downarrow & & \downarrow \hat{\tau} \\
 P \times G \xleftarrow{T_U} P & & Q \xrightarrow{\hat{T}_U} P \times F \\
 \pi \downarrow & & \downarrow \hat{\pi} \\
 & M &
 \end{array}$$

Figure 1. The picture of the principal associated bundles of the universe. M represents our universe, and above M stands the high-dimensional space-time (heaven) of the laws or rules governing the universe.

3) The gauge potential corresponds to the connection of the principal bundle, and the gauge field strength corresponds to the curvature on the principal bundle. So it can be said that the connection is the gauge potential and the curvature is the gauge field strength.

4) The gauge potential on the bottom manifold M is a 1-form field function of the Lie algebra taking G in a space-time region. Different gauge potentials satisfy the formula of gauge potential transformation, but from P , these different gauge potentials are just the gauge potentials of the principal bundle (connection field $\tilde{\omega}$) projection component on the base manifold under the cross section transformation, total gauge potential $\tilde{\omega}$ is invariant, and the transformation between cross sections is the gauge transformation; this general gauge potential $\tilde{\omega}$ corresponds to a cosmic space-time gauge field, and the four fundamental interactions in reality only correspond to the projection components of this cosmic space-time gauge field under different gauge choices (internal frame choices); this is the meaning of gauge transformation invariance of gauge field.

5) There may be many components of connection $\tilde{\omega}$, so there are not only four fundamental interactions in the universe, but also five, six, and so on. These corresponding fundamental interactions are spatio-temporal regional field functions of different gauge choices (internal frame choices), and may also be the superposition or combination of four (or other fundamental) interactions.

6) The so-called gauge transformation may be a transformation from one fundamental interaction (such as electromagnetic force) to another fundamental interaction (such as gravity) in the picture of the principal associated bundles of the universe; the so-called gauge invariance means that the gauge field of space and time in the universe is invariant under the gauge transformation, and gravity, electromagnetic force, or strong or weak force are all the expressions of its components. The gauge transformation is only the transformation between these components, and the gauge invariance is that the total space-time gauge field does not change with the gauge transformation (showing the Lagrangian invariance of the physical field).

7) Quantization is not necessarily a necessary condition for the unification of the gauge field, but the invariance of gauge transformation is the necessary condition for the unification of the gauge field!

8) The four fundamental fields of the universe (gravity, electromagnetic force, weak force, and strong force) are unified in one cosmic space-time gauge potential $\tilde{\omega}$, which corresponds to a cosmic space-time gauge field, and the mutual transformation between the four fundamental gauge fields can be described by a generalized gauge potential Equation (referred to as GG equation):

$$\sigma_V^* \tilde{\omega}(Y) = \mathcal{A}d_{g_{UV}(x)^{-1}} \sigma_U^* \tilde{\omega}(Y) + L_{g_{UV}(x)^*}^{-1} g_{UV^*}(Y), \forall x \in U \cap V, Y \in T_x M \quad (\text{GGE})$$

here, σ_V^* and σ_U^* is the pull back mapping of the principal bundle cross section to the bottom manifold region V or U respectively, showing that all the different gauge potentials are just the components of the cosmic space-time unified

potential in different regions of the bottom manifold; the physical and mathematical meanings of other symbols will be introduced later.

Let's use mathematical physics to construct a strict framework for the above arguments; the main references are [25] (Chapter 6, Appendix 3), [26] (Appendix G and I in Volume II and III) however, the author of this paper has made some generalization in the process of derivation and proof.

3. Relevant Mathematical Physics Framework

3.1. Concepts of Principal Fiber Bundle

The principal fiber bundle $P(M, G)$ consists of the bundle manifold P , the base manifold M , and the Lie group G of the structure group, and satisfies the following three conditions:

1) G has a free right action R on P , $R: P \times G \rightarrow P$; here free means: if $g \neq e$, then there is $pg \neq p, \forall g \in G, p \in P$. Since the dimension of P is greater than G , so this is an embedding map. It makes fibers also Lie groups (both manifolds and groups), and is isomorphic to G .

2) There is a C^∞ projection on to mapping $\pi: P \rightarrow M$, and satisfies: $\pi^{-1}[\pi(p)] = \{pg | g \in G\}, \forall p \in P$. In order to ensure the existence of π^{-1} , $[\]$ represents a singleton subset.

3) For each $x \in M$, the existence of open neighborhood $U \subset M$ and differential homeomorphism (local trivialization mapping) $T_U: \pi^{-1}[U] \rightarrow U \times G$ can be expressed as: $T_U(p) = (\pi(p), S_U(p)), \forall p \in \pi^{-1}[U]$, where mapping S_U is defined as $\pi^{-1}[U] \rightarrow G$, and meets (i.e., the "core" requirement of local trivialization T_U : corresponding to $p \rightarrow pg$ on fiber or orbit): $S_U(pg) = S_U(p)g, \forall g \in G$. Here $S_U: \pi^{-1}[U] \rightarrow G$ is not a differential homeomorphism, because $\pi^{-1}[U]$ manifold is larger than G , but $S_U: \pi^{-1}[x] \rightarrow G$ is a differential homeomorphism. If p is a special point, it is marked as $\check{p}_U(x), S_U(\check{p}_U) = e$, and $R_{\check{p}_U}: G \rightarrow \pi^{-1}[x]$ is also a differential homeomorphism, so $R_p = S_U^{-1}$.

Here $\pi^{-1}[x], x \in U \subset M$ can be called the fiber above the point $x \in M$; $\pi^{-1}[\pi(p)] = \{pg | g \in G\}$ is equal to the group G right acting R to the orbit passing through point p , that is, the fiber is equal to the orbit, $\{pg | g \in G\} \subset P$, where the fiber bundle forms a bundle manifold; the principal fiber bundle $P(M, G)$ can be abbreviated as P .

3.2. Concept of Associated Bundle

Let $P(M, G)$ be the principal bundle. If manifold F is selected, make G have a left action χ on F (no free requirement), $\chi: G \times F \rightarrow F$, i.e.

$\chi_g(f) = gf, \forall g \in G, f \in F$. Then the free right action R of group G on P and the left action χ of group G on F jointly induce the free right action ξ of G on $P \times F$, $\xi: (P \times F) \times G \rightarrow P \times F$, which is defined as

$\xi_g(p, f) := (pg, g^{-1}f) \in P \times F$, here $\forall g \in G, p \in P, f \in F$. If $\tau: P \times F \rightarrow P$ represents the natural projection mapping, that is, $\tau(p, f) := p, \forall p \in P, f \in F$,

each track of ξ on $P \times F$ is regarded as an element, the set of all elements is defined as Q , any element $q \in Q$ represents a orbit on $P \times F$, and Q is a manifold, which we call the associated bundle. Here q can be defined as

$q \equiv \{pg, g^{-1}f \mid g \in G\} \equiv p \cdot f$, which is called the orbit passing (p, f) point and is formed by $\xi_g(p, f)$. Note that $p \cdot f = pg \cdot g^{-1}f = p \cdot gg^{-1}f = p' \cdot f'$.

3.3. Relations of Natural Projection Mapping of Principal Associated Bundles

1) $\hat{\tau} : P \times F \rightarrow Q$, is defined as $\hat{\tau}(p, f) := p \cdot f \in Q$, i.e. $\forall p \in P$, $\hat{\tau}_p : F \rightarrow Q$; from this, we can define a topology for Q , let $\phi \subset Q$ is open if and only if $\hat{\tau}^{-1}[\phi] \subset P \times F$ is open, so Q is topological space, $\hat{\tau}$ is a continuous mapping. Not only that, but also Q can be proved to be a manifold.

2) $\hat{\pi} : Q \rightarrow M$, defined as $\hat{\pi}(q) := \pi(p) \in M$, $\forall q = p \cdot f \in Q$. Hence more precisely, $\hat{\tau}_p : F \rightarrow \hat{\pi}^{-1}[x]$, $x \equiv \pi(p)$ while $R_p : G \rightarrow \pi^{-1}[x]$, $x \equiv \pi(p)$, and $\hat{\tau}_p, R_p$ is a differential homeomorphic mapping. In other words, $\hat{\tau}_p$ or R_p brings the manifold structure of F or G respectively to the fiber $\hat{\pi}^{-1}[x]$ of the companion bundle Q or the fiber $\pi^{-1}[x]$ of the principal bundle P .

3) $\tau : P \times F \rightarrow P$, $\tau(p, f) := p$, $\forall p \in P$, $f \in F$.

4) $\pi : P \rightarrow M$, and meet: $\pi^{-1}[\pi(p)] = \{pg \mid g \in G\}$, $\forall p \in P$.

Here, the relevant definitions in (3), (4) have been given by 3.1 and 3.2.

If for every $x \in M$ there exists an open neighborhood $U \subset M$, its inverse image $\pi^{-1}[U]$ is diffeomorphic to the product manifold $U \times G$, i.e. $\pi^{-1}[U] = U \times G$, then the corresponding T_U is said to be locally trivial, and the corresponding principal bundle is locally trivial; if $U = M$, that is, $\pi^{-1}[U] = P = M \times G$, is said to be integrally trivial. It can be said that any principal bundle is locally trivial. Therefore, the local trivialization can be extended to the principal associated bundles picture of the universe (see **Figure 2**).

Where, note that $T_U : \pi^{-1}[U] \rightarrow U \times G$, $T_U(p) := (\pi(p), S_U(p))$, therefore, local trivial $\hat{T}_U : \pi^{-1}[U] \rightarrow U \times F$, $\hat{T}_U(p) := (\hat{\pi}(q), \check{f}_U)$; here, $S_U(\check{p}_U) = e$, $q = \check{p}_U \cdot f = \check{p}_U \cdot \check{f}_U$. So Q is an associated bundle of the principal bundle P , and F is called the typical fiber of Q . If there is $U \cap V \neq \emptyset$, then one has $\check{p}_V = \check{p}_U g_{UV}(x) \Rightarrow g_{UV}(x) \check{f}_U = \check{f}_V$. In addition, the local cross section of the associated bundle can be defined as $\hat{\sigma} : U \rightarrow Q$, namely satisfying $\pi(\hat{\sigma}(x)) = x$.

$$\begin{array}{ccc}
 & \pi^{-1}[U] \times F & \\
 \tau \downarrow & & \downarrow \hat{\tau} \\
 U \times G \xleftarrow{T_U} \pi^{-1}[U] & & \hat{\pi}^{-1}[U] \xrightarrow{\hat{T}_U} U \times F \\
 \pi \downarrow & & \downarrow \hat{\pi} \\
 U & = & U
 \end{array}$$

Figure 2. The more specific structure of the principal associated bundles of the universe, $\pi^{-1}[U]$ represents the principal fiber bundle on U ; U refers to the overall trivial or locally trivial region of M .

3.4. Frame Bundle

M is a set as n -dimensional manifold, $P \equiv \{x, \{e_\mu\} | x \in M\}$, $\{e_\mu\}$ is a $T_x M$ base, abbreviated as $e_\mu T_x$ stands the tangent space for $x \in M$. Then P can be proved to be a $n + n^2$ dimensional manifold. Now select $GL(n)$ as the structure group G , then a frame bundle can be constructed by the following three steps:

1) Define the right action R of matrix group $GL(n)$ on P , $R: P \times GL(n) \rightarrow P$ is: $R_g(x, e_\mu g_\mu^\nu)$, here g_μ^ν represents a matrix element of g .

2) Define projection mapping $\pi: P \rightarrow M$, namely $\pi(x, e_\mu) := x, \forall (x, e_\mu) \in P$.

3) Define local trivial $T_U: \pi^{-1}[U] \rightarrow U \times G$, namely $T_U(x, e_\mu) := (x, h)$, here

$h \equiv S_U(x, e_\mu) \in G$, $\left. \frac{\partial}{\partial x^\nu} \right|_x h_\mu^\nu = e_\mu$, and $S_U(pg) = S_U(p)g, \forall g \in G$. So T_U is differential homeomorphic.

The principal bundle $P(M, GL(n))$ constructed by the above three steps is called the frame bundle, which is recorded as FM .

3.5. Tangent Bundle

On the basis of FM , take the manifold $F = \mathbb{R}^n$, then F is a vector space, and $f \in F$ can be expressed as a column matrix composed of n real numbers, that is, $(f^1, \dots, f^n)$; so we can define the left action $\chi: G \times F \rightarrow F$ as $(\chi_g(f))^\mu := g_\nu^\mu f^\nu, \forall g \in GL(n), f \in F$; then a ξ is determined by the right and left actions, $\xi: (P \times F) \times G \rightarrow P \times F$, $\xi_g: P \times F \rightarrow P \times F$, specifically $\xi_g(p, f) = (pg, g^{-1}f) \Rightarrow \xi_g(x, e_\mu; f^\rho) = (x, e_\nu g_\mu^\nu; (g^{-1})^\rho_\sigma f^\sigma)$. Here $(x, e_\mu; f^\rho) \in P \times F$ can produce $v \equiv e_\mu f^\mu \in T_x M$, and on the same orbit $v = e_\mu f^\mu = v' = e'_\mu f'^\mu$; that is to say, each $q \in \hat{\pi}^{-1}[x]$ point (represents an orbit) 1-1 corresponds to a vector v in $T_x M$, and all different v in $T_x M$ correspond to different q above it to form a tangent bundle $\hat{\pi}^{-1}[x]$, that is, $\hat{\pi}^{-1}[x] \xrightarrow{1-1} T_x M$; so the tangent bundle $Q = P \times F$ is the associated bundle of FM . Furthermore, Q can be regarded as the tangent bundle TM of M , so any cross section $\hat{\sigma}: U \rightarrow Q$ (because 1-1 corresponds to the vector of the tangent space on U) is a vector field on $U \subset M$. Since it is a vector field, at least it is preliminarily explained that the section $\hat{\sigma}: U \rightarrow Q$ is related to coordinates, different section corresponds to different coordinate, and there is a relationship of coordinate transformation between sections.

On the basis of FM if the manifold $F = (\mathbb{R}^n)^* = T_{\mathbb{R}^n}^*(0, 1)$, $f = (f_1, \dots, f_n) \in F$, $(\chi_g(f))_\mu := (g^{-1})_\mu^\nu f_\nu$, then given any point $(x, e_\mu; f_\rho) \in P \times F$, one can produce: $\beta \equiv e^\mu f_\mu \in T_x^* M$ (the dual space of $T_x M$), and there is $\beta = \beta'$ on the same orbital; all the different β in $T_x^* M$ correspond to the different q above it to form a cotangent bundle $\hat{\pi}^{-1}[x]$, that is, $\hat{\pi}^{-1}[x] \xrightarrow{1-1} T_x^* M$; so the cotangent bundle $Q = P \times F$ is also an associated bundle of FM . Any of its cross sections $\hat{\sigma}: U \rightarrow Q$ is a covector field on $U \subset M$. Note that in general, $\chi: G \times F \rightarrow F$ is a Lie transformation group: $\hat{G} \equiv \{\chi_g: F \rightarrow F | g \in G\}$.

In addition, one of the simplest structures of the principal associated bundle of the universe constructed by us may be the frame bundle plus the associated tangent bundle. However, in order to make this structure of the principal associated bundles of the universe able to accommodate the most universal gauge field of the universe and the corresponding unknown fundamental interactions, we still use the general abstract structure of the principal associated bundles to replace the concrete model.

4. Principal Bundle and Gauge Field

4.1. Gauge Selection and Cross Section

Definition: Let $P(M, G)$ be the principal bundle, U be the open subset of M , C^∞ mapping $\sigma: U \rightarrow P$ is called a local cross section, if $\pi(\sigma(x)) = x, \forall x \in U$. Here if $U = M$, then $\sigma: M \rightarrow P$, which is called the overall section. Next, in the case of local cross section, we further explore the physical meaning of the local cross section: let $\mathbb{R}^4$ be an open subset of the bottom manifold M , and G is the structural group to construct a non-trivial principal bundle $\mathbb{R}^4 \times G$, where the free right-hand action of G on P is $R: (\mathbb{R}^4 \times G) \times G \rightarrow \mathbb{R}^4 \times G$, namely $\forall g_1 \in G$, define $R_{g_1}: \mathbb{R}^4 \times G \rightarrow \mathbb{R}^4 \times G$ as: $R_{g_1}(x, g_2) := (x, g_2 g_1)$, $\forall (x, g_1) \in \mathbb{R}^4 \times G$. Let $\sigma: \mathbb{R}^4 \rightarrow P$ and $\sigma': \mathbb{R}^4 \rightarrow P$ be the local sections of P respectively, then $\forall x \in \mathbb{R}^4$ has a unique group element field $g: \mathbb{R}^4 \rightarrow G$ such that:

$\forall g(x) \in G, x \in \mathbb{R}^4, \sigma'(x) = \sigma(x)g(x)^{-1}$. Therefore, there exists a representation group element such that $U(x) \equiv \rho(g(x)) \in \hat{G}$, $\rho: G \rightarrow \hat{G}$, thus creating a local gauge transformation: $\phi'(x) = U(x)\phi(x) \equiv \rho(g(x))\phi(x), \forall \phi(x) \in V$, where V is the representation space of $\hat{G}$, and $\hat{G}$ is a representation of G . At this time, $\phi(x)$ is actually a column matrix, and $\rho(g(x))$ is a square matrix. Then define the associated bundle: choose $F = V$, define the left action $\chi: G \times F \rightarrow F$ as: $\forall g_1 \in G, \chi_{g_1}: F \rightarrow F, \chi_{g_1}(f_1) := \rho(g_1)(f_1), \forall f_1 \in F$, then there is an associated bundle $\phi(x) \equiv q = p \cdot f = \sigma(x) \cdot f(x) \in \hat{\pi}^{-1}[x] \subset Q$, where

$f: M \rightarrow F = V, \forall f(x) \in F = V$. So $\phi(x)$ is determined by the cross sections σ and f . In addition, $g(x)$ can generate: 1) $\sigma'(x) = \sigma(x)g(x)^{-1}$, 2)

$f'(x) = \chi_{g(x)} f(x) = \rho(g(x)) f(x) = g(x) f(x)$ (i.e. gauge transformation),

$\Phi'(x) = \sigma'(x) \cdot f'(x) = \sigma(x)g(x)^{-1} \cdot g(x)f(x)$

which is equivalent to

$$= \sigma(x) \cdot g(x)^{-1} g(x) f(x) = \sigma(x) \cdot f(x) = \Phi(x) \in \hat{\pi}^{-1}[x].$$

Furthermore, we can generally say that the so-called local (global) gauge transformation for the principal bundle is actually the transformation cross section $\sigma(x) \rightarrow \sigma'(x)$. This is equivalent to the transformation of the frame and the transformation of the components (of a physical field) under the internal frame field: $f(x) = \phi(x) \rightarrow f'(x) = \phi'(x)$, but the overall physical field (internal vector $\Phi(x)$) is unchanged, i.e. $\Phi'(x) = \Phi(x)$. The so-called gauge selection is to select different cross sections, while a cross section on the associated bundle $\hat{\sigma}$ is just a physical field $\Phi(x)$! In a word, the change of the cross section on the principal bundle is the change of the internal frame. If the internal

frame is changed, it is equivalent to the transformation of a gauge. Therefore, to select a cross section of the principal bundle is to select a gauge. Note that the construction and related proofs here are relatively general. Even if $\mathbb{R}^4$ is changed to a submanifold of general M , the conclusion is still correct, and there is no restriction on the type of gauge potential or the type of fundamental interaction.

4.2. Invariance of Global Gauge Transformation of Physical Field

Assume that the physical fields ϕ_1 and ϕ_2 are two independent particle fields in Minkowski space-time $(\mathbb{R}^4, \eta_{ab})$, have the same mass m , and spin 0, then according to the spirit of quantum field theory, they must respectively obey Klein-Gordon equation:

$$\partial^a \partial_a \phi_j - m^2 \phi_j = 0, j = 1, 2 \quad (1)$$

The total Lagrangian density of both is:

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2 = -\frac{1}{2} \left[(\partial^a \phi_1) \partial_a \phi_1 + m^2 \phi_1^2 + (\partial^a \phi_2) \partial_a \phi_2 + m^2 \phi_2^2 \right] \quad (2)$$

Introducing a complex scalar field

$$\begin{cases} \phi = \frac{1}{\sqrt{2}} (\phi_1 + i\phi_2) \\ \bar{\phi} = \frac{1}{\sqrt{2}} (\phi_1 - i\phi_2) \end{cases} \quad (3)$$

Then the Klein-Gordon Equation (1) becomes:

$$\begin{cases} \partial^a \partial_a \phi - m^2 \phi = 0 \\ \partial^a \partial_a \bar{\phi} - m^2 \bar{\phi} = 0 \end{cases} \quad (4)$$

Then the total Lagrangian density becomes:

$$\mathcal{L} = - \left[(\partial^a \bar{\phi}) \partial_a \phi + m^2 \phi \bar{\phi} \right] \quad (5)$$

Then an overall gauge transformation can be introduced as follows:

$$\begin{cases} \phi' = e^{-iq\theta} \phi \\ \bar{\phi}' = e^{iq\theta} \bar{\phi} \end{cases} \quad (6)$$

where θ is a real number, q is an integer, it is actually a group $U(1) = \left\{ e^{-iq\theta} \middle| \theta \in \mathbb{R} \right\}$, more generally: $U(1) \rightarrow \hat{G}$, $e^{-iq\theta} \rightarrow \text{diag}(e^{-iq_1\theta}, \dots, e^{-iq_N\theta})$; hence one has

$$\begin{cases} \partial_a \phi' = e^{-iq\theta} \partial_a \phi \\ \partial_a \bar{\phi}' = e^{iq\theta} \partial_a \bar{\phi} \end{cases} \quad (7)$$

therefore, the total Lagrangian density (5) remains unchanged, namely

$$\mathcal{L} = - \left[(\partial^a \bar{\phi}') \partial_a \phi' + m^2 \phi' \bar{\phi}' \right] = - \left[(\partial^a \bar{\phi}) \partial_a \phi + m^2 \phi \bar{\phi} \right] \quad (8)$$

It can be seen that the total Lagrangian density is invariant under the field transformation of the gauge transformation (6), which reflects the symmetry

under the field transformation and is different from the space-time symmetry. Therefore, we call this invariance the internal symmetry of the field, which, like the space-time symmetry, can lead to the conservation law and can be expressed by the Noether theorem [27].

4.3. Local Gauge Invariance for Non-Abelian Fields

In gauge transformation (6) θ does not change with time and space points, which is called the overall gauge transformation. The converse is called local gauge transformation. If the multiplication of the corresponding Lie group G is commutative, it is called an Abelian field, otherwise it is called a non-Abelian field. This non-Abelian field is also called Yang-Mills field, which was proposed by Yang-Mills in 1954 by extending G from $U(1)$ to $SU(2)$ [1]. From then on, it was the first time to open a river that the gauge field theory unified the fundamental physical interaction. Following the basic spirit of Yang-Mills field, we discuss the local gauge invariance of a generalized non-Abelian field:

First, Equation (8) is generalized to the local gauge transformation:

$$\begin{cases} \phi'(x) = U(\theta(x))\phi(x) = e^{-iL\cdot\theta(x)}\phi(x) \\ \bar{\phi}'(x) = \bar{\phi}(x)U(\theta(x))^{-1} = \bar{\phi}(x)e^{iL\cdot\theta(x)} \end{cases} \quad (9)$$

Then, we make the following assumptions:

1) The structure group and its representation group of the system are denoted as G and $\rho: G \rightarrow \hat{G}$, $\dim G \equiv R$, the physical field concerned is a multi-component complex particle field $\phi(x) \in V$ (representing space), which can be called the generalized Yang-Mills field, can be accompanied by R potentials (gauge potentials). These gauge potentials (also known as Yang-Mills potentials) are R complementary potential fields $\{A_\mu^r | r=1, \dots, R\}$ that can be generalized in general manifolds (including space-time manifolds of general relativity). Their coordinate components are marked as $A_\mu^r(x)$.

2) The total Lagrangian density of the system can be expressed as:

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_{YM} \quad (10)$$

Here $\mathcal{L}_{YM}$ is the Lagrangian density of the generalized Yang-Mills field, $\mathcal{L}_1$ can be expressed as:

$$\mathcal{L}_1 = \mathcal{L}_0(\phi(x), D_\mu\phi(x); \bar{\phi}(x), D_\mu\bar{\phi}(x)) \quad (11)$$

Here $D_\mu\phi(x)$ and $D_\mu\bar{\phi}(x)$ are respectively defined as:

$$\begin{cases} D_\mu\phi(x) = \partial_\mu\phi(x) - ikL \cdot A_\mu(x)\phi(x) \\ D_\mu\bar{\phi}(x) = \partial_\mu\bar{\phi}(x) + ik\bar{\phi}(x)L \cdot A_\mu(x) \end{cases} \quad (12)$$

which can also be defined as a derivative covariant operator, where k is defined as a coupling constant (and if $G = U(1)$, then $k = e \leftrightarrow$ Electromagnetic gauge field; if $G = SO(N)$, then $k = -1 \leftrightarrow$ Gravitational gauge field); and $L \cdot A_\mu(x)$ is defined as:

$$L \cdot A_\mu(x) = L_1 A_\mu^1(x) + \dots + L_R A_\mu^R(x) = L_r A_\mu^r(x) \quad (13)$$

Then under the local gauge transformation (9), formula (8) can be expressed as:

$$\begin{cases} D'_\mu \phi'(x) = U(\theta(x)) D_\mu \phi(x) \\ D'_\mu \bar{\phi}'(x) = [D_\mu \bar{\phi}(x)] U(\theta(x))^{-1} \end{cases} \quad (14)$$

and make $\mathcal{L}_1$ unchanged. To ensure these, we require the transformation form of $A'_\mu(x)$ is as follows:

$$\begin{aligned} (a) \quad D'_\mu \phi'(x) &= \partial_\mu \phi'(x) - ik\mathbf{L} \cdot \mathbf{A}'_\mu(x) \phi'(x) \\ &= U(\theta(x)) \partial_\mu \phi(x) + [\partial_\mu U(\theta(x))] \phi(x) - ik\mathbf{L} \cdot \mathbf{A}'_\mu(x) U(\theta(x)) \phi(x) \end{aligned} \quad (15)$$

$$\begin{aligned} (b) \quad U(\theta(x)) D_\mu \phi(x) &= U(\theta(x)) [\partial_\mu \phi(x) - ik\mathbf{L} \cdot \mathbf{A}_\mu(x) \phi(x)] \\ &= U(\theta(x)) \partial_\mu \phi(x) - ikU(\theta(x)) \mathbf{L} \cdot \mathbf{A}_\mu(x) \phi(x) \end{aligned} \quad (16)$$

Obviously, considering formula (14), and comparing formula (16) with formula (15), to make (a) = (b), we only need to have:

$$\partial_\mu U(\theta(x)) - ik\mathbf{L} \cdot \mathbf{A}'_\mu(x) U(\theta(x)) = -ikU(\theta(x)) \mathbf{L} \cdot \mathbf{A}_\mu(x) \phi(x), \quad (17)$$

after sorting out, it is found that for group $U(1)$, there is:

$$A'^1_\mu(x) = A^1_\mu(x) - e^{-1} \partial_\mu \theta(x) \quad (18)$$

For general case, defining $-iL_r \equiv \rho_* e_r \in \hat{\mathcal{G}}$, here $\hat{\mathcal{G}}$ is the representation of a Lie algebra of G , or Lie algebra of $\hat{G}$; e_r is a basis of Lie algebra of $\hat{G}$; ρ_* is a forward mapping of ρ , then $\hat{A}_\mu(x) \equiv -i\mathbf{L} \cdot \mathbf{A}_\mu(x) = -iL_r A^r_\mu(x) \in \hat{\mathcal{G}}$ can be defined. Therefore, Equation (17) becomes the following equation, which can be also proved that its right side belongs to $\hat{\mathcal{G}}$.

$$\hat{A}'_\mu(x) = U(\theta(x)) \hat{A}_\mu(x) U(\theta(x))^{-1} - k^{-1} \partial_\mu U(\theta(x)) U(\theta(x))^{-1} \quad (19)$$

what needs to be emphasized here is that formula (19) is essentially a form of “generalized gauge transformation”, that is, the general transformation formula between two connections of the bottom manifold in the intersection domain: the generalized gauge equation, if we also look at the gauge potential as it is a connection. We shall return to this issue later.

Now let's introduce the generalized Yang-Mills field strength: because R gauge potentials $A^r_\mu(x)$ have been introduced, so there should be R gauge field strengths $F^r_{\mu\nu}(r=1, \dots, R)$, which can be expressed as:

$$F^r_{\mu\nu}(x) = \partial_\mu A^r_\nu - \partial_\nu A^r_\mu + k \sum_{s,t=1}^R C^r_{st} A^s_\mu(x) A^t_\nu(x), (r=1, \dots, R) \quad (20)$$

where C^r_{st} represents the structural constant of the Lie algebra $\hat{\mathcal{G}}$ of G under the basis $\{e_r\}$. So it can be defined:

$$\mathcal{L}_{\text{YM}} := -\frac{1}{16\pi} \sum_{r=1}^R F^r_{\mu\nu} F^{r\mu\nu} \quad (21)$$

here, the metric $g^{\mu\alpha}$, $g^{\nu\beta}$ can be used to lift subscripts of $F^r_{\alpha\beta}$:

The following task is to prove that $\mathcal{L}_{\text{YM}}$ is also invariant under transformation (19). For this, we first introduce the simplified notation: $\hat{F}^r_{\mu\nu}(x) \equiv -iL_r F^r_{\mu\nu}(x) \in \hat{\mathcal{G}}$, similar Equation (20) can be changed to

$$\hat{F}_{\mu\nu}(x) = \partial_\mu \hat{A}_\nu(x) - \partial_\nu \hat{A}_\mu(x) + k[\hat{A}_\mu(x), \hat{A}_\nu(x)] \quad (22)$$

where $[\hat{A}_\mu(x), \hat{A}_\nu(x)]$ is the Lie brackets of the Lie algebra elements $\hat{A}_\mu(x)$ and $\hat{A}_\nu(x)$. From this, it can be proved that $\mathcal{L}_{YM}$ is invariant under the transformation (19), that is, $\mathcal{L}'_{YM}(x) = \mathcal{L}_{YM}(x)$. The proof is as follows:

Firstly we prove that $16\pi\lambda_\rho\mathcal{L}_{YM} = \text{tr}(F_{\mu\nu}F^{\mu\nu})$. Let $\{e_r\}$ be the orthonormal basis, and let $F_{\mu\nu}(x) = e_r F_{\mu\nu}^r(x) \in \hat{\mathcal{G}}$, then we can introduce ρ , so that $\rho_* F_{\mu\nu}(x) = \hat{F}_{\mu\nu}(x) \in \hat{\mathcal{G}}$, then we have

$$\begin{aligned} \text{tr}(F_{\mu\nu}F^{\mu\nu}) &= \text{tr}[(\rho_* F_{\mu\nu})(\rho_* F^{\mu\nu})] = \text{tr}\left[\left(\rho_* \sum_{r=1}^R F_{\mu\nu}^r e_r\right)\left(\rho_* \sum_{s=1}^R F^{\mu\nu s} e_s\right)\right] \\ &= \sum_{r,s=1}^R F_{\mu\nu}^r F^{\mu\nu s} \text{tr}[(\rho_* e_r)(\rho_* e_s)] = \sum_{r,s=1}^R F_{\mu\nu}^r F^{\mu\nu s} \tilde{K}(e_r, e_s) \quad (23) \\ &= \lambda_\rho \sum_{r,s=1}^R F_{\mu\nu}^r F^{\mu\nu s} K(e_r, e_s) = -\lambda_\rho \sum_{r,s=1}^R F_{\mu\nu}^r F^{\mu\nu s} \delta_{rs} = 16\pi\lambda_\rho \mathcal{L}_{YM} \end{aligned}$$

where $\tilde{K}$ is the generalized Cartan metric satisfying $\tilde{K} = \lambda_\rho K$, and $K(e_r, e_s) = -\delta_{rs}$. The last step uses formula (21).

Secondly, we can prove that $\text{tr}(\hat{F}'_{\mu\nu} \hat{F}'^{\mu\nu}) = \text{tr}(\hat{F}_{\mu\nu} \hat{F}^{\mu\nu})$, namely

$$\text{tr}(\hat{F}'_{\mu\nu} \hat{F}'^{\mu\nu}) = \text{tr}(U \hat{F}_{\mu\nu} U^{-1} U \hat{F}^{\mu\nu} U^{-1}) = \text{tr}(U^{-1} U \hat{F}_{\mu\nu} \hat{F}^{\mu\nu}) = \text{tr}(\hat{F}_{\mu\nu} \hat{F}^{\mu\nu}) \quad (24)$$

where the last step used the rotation of the matrix under the tr operator. Therefore, from formula (24), $\mathcal{L}'_{YM}(x) = \mathcal{L}_{YM}(x)$.

q.e.d.

4.4. Gauge Potential and Connection

The above discussion (including the generalization of Yang-Mills potential) shows that in a very general principal associated bundles structure, *i.e.* in the equation $\phi'(x) = U(\theta(x))\phi(x)$, one can choose a definition,

$U(\theta(x)) \equiv \rho(g(x)) \in \hat{\mathcal{G}}$, to construct a local gauge transformation:

$\phi'(x) = U(x)\phi(x) \equiv \rho(g(x))\phi(x)$, $\forall \phi(x) \in V$. Here V is the representation space of $\hat{\mathcal{G}}$. So here $\phi(x)$ is a column matrix, $\rho(g(x))$ is a square matrix.

Reselect $F = V$, $\forall f(x) \in F = V$, define $\Phi(x) \equiv \sigma(x) \cdot f(x) \in \hat{\pi}^{-1}[x] \subset Q$,

where $f(x) \in F$ (a typical fiber), one can introduce

$\Phi'(x) \equiv \sigma'(x) \cdot f'(x) \in \hat{\pi}^{-1}[x] = \Phi(x)$. In addition, for the principal bundle FM and associated bundle TM , there is

$q = \sigma(x) \cdot f^\mu = (x, e_\mu) f^\mu = e_\mu f^\mu = e'_\mu f'^\mu \equiv V$ which is called space-time vector, while $\Phi(x)$ can be called internal vector; $\sigma(x)$ is called the internal frame, and $f(x)$ is called the component of the internal vector expanded with the internal frame. But if $\sigma(x)g(x)^{-1} = \sigma'(x)$, $f(x) = f'(x)$, then: $\Phi(x) = \Phi'(x)$.

In the above discussion, we also see that in order to ensure the invariance of the total Lagrangian density $\mathcal{L}$ under the local gauge transformation, that is, equivalent invariance for the cross section transformation to the principal bundle P , the gauge potential $A_\mu(x)$ must be introduced, and make it under gauge transform (*i.e.* the section transform $\sigma'(x) = \sigma(x)g(x)^{-1}$, where $g(x)$ is a group element field) to change to $A'_\mu(x)$, then $A_\mu(x) \rightarrow A'_\mu(x)$ (that is, the transformation of gauge potential satisfying Equation (19)) also corresponds to

an absolute invariance, that is, the invariance of the connection on the principal bundle $\tilde{\omega}$. In other words, the connection on the principal bundle is the gauge potential, which is invariant under the transformation of the gauge potential. For this reason, we firstly give three definitions of connection on the principal bundle:

1) A connection on the principal bundle $P(M, G)$ is to specify a horizontal subspace $H_p \subset T_p P$ for each point $p \in P$, satisfying:

- a) $T_p P = V_p \oplus H_p$, $\forall p \in P$,
 - b) $R_g^* [H_p] = H_{pg}$, $\forall p \in P, g \in G$,
 - c) H_p gives a C^∞ n -dimensional distribution on P .
- Here $V_p := \{X \in T_p P \mid \pi_*(X) = 0\}$ is the vertical subspace.

2) A connection on the principal bundle $P(M, G)$ is a first-order differential form (abbreviation: 1 form) field $\tilde{\omega}$ of a Lie algebra $\mathcal{G}$ value of C^∞ on P , and satisfies:

- a) $\tilde{\omega}_p(A_p^*) = A, \forall A \in \mathcal{G}, p \in P$,
- b) $\tilde{\omega}_{pg}(R_g^* X) = \text{Ad}_{g^{-1}} \tilde{\omega}_p(X), \forall p \in P, g \in G, X \in T_p P$,

Here A_p^* is a vertical vector field generated by A on $p \in P$; while $\text{Ad}_{g^{-1}}$ represents an automorphism called adjoint isomorphism constructed by element g^{-1} through an isomorphic mapping $I_{g^{-1}} := g^{-1}hg, \forall h \in G$ to induce a push forward mapping (tangent mapping) as $I_{g^{-1}*} := \text{Ad}_{g^{-1}}$ at $I_{g^{-1}}(e) = e$ point, that is, $\text{Ad}_{g^{-1}} : \mathcal{G} \rightarrow \mathcal{G}$ is a linear transformation on Lie algebra $\mathcal{G}$.

3) A connection on the principal bundle $P(M, G)$ is to the local trivial $T_U : \pi^{-1}[U] \rightarrow U \times G$ specifies a 1-form field ω_U of the $\mathcal{G}$ value of C^∞ on U , that is a connection on $U \subset M$. At this time, if $T_V : \pi^{-1}[V] \rightarrow V \times G$ is another local triviality, that is, $U \cap V \neq \emptyset$, and the conversion function from T_U to T_V is g_{UV} , then there is

$$\omega_V(Y) = \text{Ad}_{g_{UV}(x)^{-1}} \omega_U(Y) + L_{g_{UV}(x)*}^{-1} g_{UV*}(Y) = B + A, \forall x \in U \cap V, Y \in T_x M \quad (25)$$

where $L_{g_{UV}(x)}^{-1}$ is the inverse map of the left translation $L_{g_{UV}(x)}$ generated by $g_{UV}(x) \in G$, $L_{g_{UV}(x)*}^{-1} \equiv (L_{g_{UV}(x)}^{-1})_*$.

Furthermore, it can be shown that the three definitions are equivalent to each other. But the emphases are different: definition (1) means that the connection is the horizontal subspace of the point-tangent space on the fiber of the principal bundle; definition (2) means that the connection is the 1-form field with Lie algebra value on the fiber point of the principal bundle; definition (3) means that the connection is the 1-form field of a point on the fiber-bottom manifold of the principal bundle, that is, $\omega_U \equiv \sigma_U^* \tilde{\omega} \in \Lambda_U(1, \mathcal{G})$, where the lower $*$ indicates the push forward mapping, and the upper $*$ indicates the pull back mapping. The connection coordination formula (25) in the third definition of connection is a form of generalized gauge potential transformation, where the local trivial T_U corresponds to the local cross section $\sigma_U := U \rightarrow P$, and there is a pull back mapping

$$\begin{cases} \omega_U := \sigma_U^* \tilde{\omega} \\ \omega_V := \sigma_V^* \tilde{\omega} \end{cases} \quad (26)$$

Proof of the Equation (25): Firstly

$$\omega_V|_x(Y) = \left(\sigma_V^* \tilde{\omega}|_{\sigma_V^*(x)} \right)(Y) = \tilde{\omega}_{\sigma_V(x)}(\sigma_{V*}Y) \quad (27)$$

but (the last step, Leibniz's law, can be proved in the work [25])

$$\begin{aligned} \sigma_{V*}Y &= \sigma_{V*} \frac{d}{dt} \Big|_0 \eta(t) = \frac{d}{dt} \Big|_0 \sigma_V(\eta(t)) = \frac{d}{dt} \Big|_0 [\sigma_U(\eta(t)) g_{UV}(\eta(t))] \\ &= \frac{d}{dt} \Big|_0 [\sigma_U(\eta(t)) g_{UV}(x)] + \frac{d}{dt} \Big|_0 [\sigma_U(x) g_{UV}(\eta(t))] \end{aligned} \quad (28)$$

$$\frac{d}{dt} \Big|_0 [\sigma_U(\eta(t)) g_{UV}(x)] = \frac{d}{dt} \Big|_0 R_{g_{UV}(x)} \sigma_U(\eta(t)) = R_{g_{UV}(x)*} \sigma_{U*}(Y) \quad (29)$$

$$\begin{aligned} \sigma_U(x) g_{UV}(\eta(t)) &= \sigma_U(x) g_{UV}(\eta(t)) = \sigma_V(x) g_{UV}(x)^{-1} g_{UV}(\eta(t)) \\ &= \sigma_V(x) \left[L_{g_{UV}(x)^{-1}} g_{UV}(\eta(t)) \right] = R_{\sigma_V(x)} \left[L_{g_{UV}(x)^{-1}} g_{UV}(\eta(t)) \right] \end{aligned} \quad (30)$$

$$\frac{d}{dt} \Big|_0 [\sigma_U(x) g_{UV}(\eta(t))] = R_{\sigma_V(x)*} \left[L_{g_{UV}(x)*}^{-1} g_{UV*}(Y) \right] = R_{\sigma_V(x)*} A = A_{\sigma_V(x)}^* \quad (31)$$

Therefore

$$\sigma_{V*}Y = R_{g_{UV}(x)*} \sigma_{U*}(Y) + A_{\sigma_V(x)}^* \quad (32)$$

$$\begin{aligned} \omega_V|_x(Y) &= \tilde{\omega}|_{\sigma_V(x)}(\sigma_{V*}Y) = \tilde{\omega}|_{\sigma_V(x)} \left(R_{g_{UV}(x)*} \sigma_{U*}(Y) \right) + \tilde{\omega}|_{\sigma_V(x)} \left(A_{\sigma_V(x)}^* \right) \\ &= \tilde{\omega}|_{\sigma_V(x)} \left(R_{g_{UV}(x)*} \sigma_{U*}(Y) \right) + A = \tilde{\omega}_{pg} \left(R_{g*}X \right) + A \\ &= Ad_{g^{-1}} \tilde{\omega}_p(X) + A = B + A \end{aligned} \quad (33)$$

where

$$\begin{aligned} Ad_{g^{-1}} \tilde{\omega}_p(X) &= Ad_{g_{UV}(x)^{-1}} \tilde{\omega}|_{\sigma_U(x)}(\sigma_{U*}Y) = Ad_{g_{UV}(x)^{-1}} \left(\sigma_U^* \tilde{\omega}|_{\sigma_U(x)} \right)(Y) \\ &= Ad_{g_{UV}(x)^{-1}} \omega_U|_x(Y) = B \end{aligned} \quad (34)$$

q.e.d.

If the structure group G is a $N \times N$ matrix group, $V \equiv \{N \times N \text{ matrix}\}$, then V is a vector space, and $G \subset V$, $\mathcal{G} \subset V$, then Equation (25) can be expressed in a simplified form:

$$\omega_V = g_{UV}^{-1} \omega_U g_{UV} + g_{UV}^{-1} dg_{UV} \quad (35)$$

For example, if we take the general gauge potential on the bottom manifold (that is, not limited to the electromagnetic gauge potential, it may also include the gravitational gauge potential, etc.) as: $A_\mu^r(x) \rightarrow A_\mu^r(x)$ (1 form field of real or complex value), then there is: $e_r A_\mu^r \in \Lambda_U(1, \mathcal{G})$, where e_r is the basis in Lie algebra $\mathcal{G}$. In addition, $\omega = \sigma^* \tilde{\omega} \rightarrow \omega' = \sigma'^* \tilde{\omega}$, $\forall \Lambda_U(1, \mathcal{G})$, where $\Lambda_U(1, \mathcal{G})$ is a set of 1 form fields of Lie algebra $\mathcal{G}$ on U . So we can define:

$\omega_a \equiv ke_r A_\mu^r \in \Lambda_M(1, \mathcal{G})$, or $\omega \equiv ke_r A^r \equiv kA \in \Lambda_M(1, \mathcal{G})$, note that here M is a

suitable general bottom manifold (that is, our “universe” manifold satisfying the local trivial condition, which can be much more complicated than 4 dimensional real or complex manifolds and is the same as M in the structure diagram of the principal associated bundles of the universe). Now we want to prove that the ω defined in this way satisfies the transformation relation (25) of the connection definition (3) above, $\omega_V(Y) \rightarrow \omega_U(Y)$.

Proof: Let $\omega_\mu \equiv ke_r A_\mu^r \in \Lambda_M(0, \mathcal{G})$, then $\omega_\mu(x) \equiv ke_r A_\mu^r(x) \in \mathcal{G}$, $\omega = \omega_\mu dx^\mu$, now let again $\sigma_V(x) = \sigma_U(x) g_{UV}(x) = \sigma'(x) = \sigma(x) g^{-1}(x)$, then Equation (25) becomes

$$\omega'(Y) = \mathcal{A}d_{g(x)} \omega(Y) + L_{g(x)*} g_*^{-1}(Y), \forall x \in M, Y \in T_x M \quad (36)$$

If one taking $Y = \frac{\partial}{\partial x^\mu} \Big|_{x_0}$, then one gets $\omega(Y) = \omega_\mu(x_0) dx^\mu \Big|_{x_0} = \omega_\mu(x_0)$,

so proving Equation (36) only needs to prove

$$\omega'_\mu(x_0) = \mathcal{A}d_{g(x_0)} \omega_\mu(x_0) + L_{g(x_0)*} g_*^{-1} \left(\frac{\partial}{\partial x^\mu} \Big|_{x_0} \right) \quad (37)$$

Then, by considering $\omega_a \equiv ke_r A_a^r \equiv kA_a^r$, and using the formula:

$$\begin{aligned} \hat{\omega}_\mu(x) &\equiv k\hat{A}_\mu^r(x) \equiv k(-iL_r A_\mu^r(x)) = k(\rho_*(e_r) A_\mu^r(x)) \\ &= \rho_*(k(e_r A_\mu^r(x))) = \rho_*(\omega_\mu(x)) \in \hat{\mathcal{G}}, \end{aligned}$$

one can change Equation (19) as

$$\begin{aligned} \hat{\omega}'_\mu(x) &= \rho(g(x)) \hat{\omega}_\mu(x) \rho(g(x))^{-1} - [\partial_\mu \rho(g(x))] \rho(g(x))^{-1} \\ &= \rho(g(x)) \hat{\omega}_\mu(x) \rho(g(x))^{-1} + \rho(g(x)) [\partial_\mu \rho(g(x))^{-1}] \\ &= \rho(g(x)) \hat{\omega}_\mu(x) \rho(g(x))^{-1} + \rho(g(x)) \left[\frac{\partial}{\partial x^\mu} \Big|_x \rho(g(x))^{-1} \right] \end{aligned} \quad (38)$$

Then, by defining $\frac{\partial}{\partial x^\mu} \Big|_x \rho(g(x))^{-1} = \frac{\partial}{\partial t} \Big|_0 \rho(g(\eta(t)))^{-1}$, the second item of the above formula can be proved to be ([26], Volume III p. 308)

$$\begin{aligned} &\rho(g(x)) \left[\frac{\partial}{\partial x^\mu} \Big|_x \rho(g(x))^{-1} \right] \\ &= \frac{\partial}{\partial t} \Big|_0 \rho(g(x)) \rho(g(\eta(t)))^{-1} = \frac{\partial}{\partial t} \Big|_0 L_{\rho(g(x))} \rho(g(\eta(t)))^{-1} \\ &= \frac{\partial}{\partial t} \Big|_0 L_{\rho(g(x))} \rho(g^{-1}(\eta(t))) = L_{\rho(g(x))*} \rho_* g_*^{-1} \frac{\partial}{\partial t} \Big|_0 \eta(t) \\ &= L_{\rho(g(x))*} \rho_* g_*^{-1} \frac{\partial}{\partial x^\mu} \Big|_x = \rho_* \left[L_{g(x)*} g_*^{-1} \frac{\partial}{\partial x^\mu} \Big|_x \right] \end{aligned} \quad (39)$$

Then the first item of Equation (38) is:

$$\begin{aligned} \rho(g(x)) \hat{\omega}_\mu(x) \rho(g(x))^{-1} &= g_1 B g_1^{-1} = \mathcal{A}d_{g_1} B = \mathcal{A}d_{\rho(g(x))} \hat{\omega}_\mu(x) \\ &= \mathcal{A}d_{\rho(g(x))} \rho_*(\omega_\mu(x)) = \rho_* [\mathcal{A}d_{\rho(g(x))} \omega_\mu(x)] \end{aligned} \quad (40)$$

therefore

$$\hat{\omega}'_{\mu}(x) = \rho_* \left[\text{Ad}_{\rho(g(x))} \omega_{\mu}(x) + L_{g(x)*} \mathcal{G}^{-1} \frac{\partial}{\partial x^{\mu}} \Big|_x \right] \quad (41)$$

So from $\hat{\omega}'_{\mu}(x) = \rho_*(\omega'_{\mu}(x))$, Equation (36) can be obtained.

q.e.d.

The above proof shows that the connection on the principal bundle is invariable, but the connection on the base manifold is variable and varies with the selection of cross sections. The transformation relationship is the same as that of the gauge potential. What is invariable is the connection on the principal bundle $\tilde{\omega}$, therefore $\tilde{\omega}$ is the (quite generalized) gauge potential! The author should emphasize that Equation (25) is actually a generalized gauge potential transformation, and $\omega'(Y)$ or $\omega(Y)$ is just components of $\tilde{\omega}$, which include not only the electromagnetic gauge potential and other three basic interaction gauge potentials, but also the gravitational gauge potential; the transformation relationship between these components is exactly the gauge potential transformation relationship (25), but no matter how the gauge transformation, $\tilde{\omega}$ is the same. This has implicitly pointed out that there is a unified universal gauge field corresponding to the invariable gauge potential in the universe!

Not only that, but it can be seen from formula (20) that the gauge potential can define the gauge field strength. At this time, if there are physically:

$A_{\mu}^r \rightarrow A_{\mu}^{r'}, F_{\mu\nu}^r \rightarrow F_{\mu\nu}^{r'}$, so what transformation does it correspond to in mathematics? What is the constant quantity?

5. Gauge Field and Space-Time Curvature

5.1. Gauge Field Strength and Curvature

The answer to the above question is the curvature transformation on the bottom manifold, that is, $\Omega \rightarrow \Omega'$, and the constant “original quantity” is the curvature on the principal bundle, denoted as $\hat{\Omega}$. Now first define the curvature on the principal bundle:

Let K be a manifold and $\mathcal{G}$ be a Lie algebra, then $\Lambda_K(i, \mathcal{G})$ is the set of smooth i -form fields of all Lie algebras $\mathcal{G}$ valued on K , namely vector space. Set up $\varphi \in \Lambda_K(i, \mathcal{G})$, $\psi \in \Lambda_K(j, \mathcal{G})$, $[\varphi, \psi] \in \Lambda_K(i+j, \mathcal{G})$, then there is

$$[\varphi, \psi](X_1, \dots, X_{i+j}) := \frac{1}{i!j!} \sum_{\pi} \delta_{\pi} \left[\varphi(X_{\pi(1)}, \dots, X_{\pi(i)}), \psi(X_{\pi(i+1)}, \dots, X_{\pi(i+j)}) \right] \quad (42)$$

Here π represents the arrangement, δ_{π} is defined as the even permutation is 1, and the odd permutation is -1, then $[\varphi, \psi](X_1, \dots, X_{i+j}) \in \Lambda_K(o, \mathcal{G})$ becomes a Lie bracket when $(X_1, \dots, X_{i+j})$ is fixed. If $i=2, j=1$, then

$$\begin{aligned} & [\varphi, \psi](X_1, X_2, X_3) \\ &= \frac{1}{2} \left([\varphi(X_1, X_2), \psi(X_3)] + [\varphi(X_3, X_1), \psi(X_2)] + [\varphi(X_2, X_3), \psi(X_1)] \right. \\ & \quad \left. - [\varphi(X_2, X_1), \psi(X_3)] - [\varphi(X_1, X_3), \psi(X_2)] - [\varphi(X_3, X_2), \psi(X_1)] \right) \end{aligned} \quad (43)$$

In addition, if $\omega \in \Lambda_K(1, \mathcal{G})$, then one has

$$[\omega, \omega](X_1, X_2) = [\omega(X_1), \omega(X_2)] - [\omega(X_2), \omega(X_1)] = 2[\omega(X_1), \omega(X_2)] \quad (44)$$

Usually, $\Lambda_K(i+j, \mathcal{G})$ is called the graded Lie algebra, if the Lie bracket operations are also defined in $\Lambda_K(i+j, \mathcal{G})$.

The covariant exterior differential is defined as follows: let the principal bundle with connection $(P, \tilde{\omega})$ have $\varphi \in \Lambda_p(i, \mathcal{G})$, define:

1) $\varphi^H \in \Lambda_p(i, \mathcal{G})$ is the horizontal component of φ ,

$$\varphi^H(X_1, \dots, X_i) := \varphi(X_1^H, \dots, X_i^H) \quad (45)$$

where $X_1, \dots, X_i$ is an arbitrary vector field on P , respectively.

2) The covariant external differential is defined as:

$$D\varphi := (d\varphi)^H \in \Lambda_p(i+1, \mathcal{G}) \quad (46)$$

3) The curvature $\tilde{\Omega}$ of a connection $\tilde{\omega} \in \Lambda_p(1, \mathcal{G})$ is defined as (note that connections on the principal bundle can determine the curvature.)

$$\tilde{\Omega} := D\tilde{\omega} \equiv (d\tilde{\omega})^H \in \Lambda_p(2, \mathcal{G}) \quad (47)$$

Then it can be proved that:

$$\tilde{\Omega} = d\tilde{\omega} + \frac{1}{2}[\tilde{\omega}, \tilde{\omega}] \quad (48)$$

Proof:

Science

$$\begin{aligned} \tilde{\Omega}_p(X, Y) &= D\tilde{\omega}_p(X, Y) = (d\tilde{\omega})^H|_p(X, Y) = d\tilde{\omega}|_p(X^H, Y^H) \\ &= d\tilde{\omega}_p(X, Y) + \frac{1}{2}[\tilde{\omega}, \tilde{\omega}]_p(X, Y) \\ &= d\tilde{\omega}_p(X, Y) + [\tilde{\omega}_p(X), \tilde{\omega}_p(Y)] \end{aligned} \quad (49)$$

therefore

$$d\tilde{\omega}|_p(X^H, Y^H) = d\tilde{\omega}_p(X, Y) + [\tilde{\omega}_p(X), \tilde{\omega}_p(Y)] \quad (50)$$

Then there are three cases:

a) $X, Y \in H_p \Rightarrow \tilde{\omega}_p(X) = 0, \tilde{\omega}_p(Y) = 0$, Equation (49) holds;

b) $X, Y \in V_p \Rightarrow X = A_p^*, Y = B_p^*, 0 = d\tilde{\omega}_p(X, Y) + [\tilde{\omega}_p(X), \tilde{\omega}_p(Y)]$,

hence

$$d\tilde{\omega}|_p(A_p^*, B_p^*) = e_r d\tilde{\omega}^r|_p(A_p^*, B_p^*) = -[\tilde{\omega}_p(A_p^*), \tilde{\omega}_p(B_p^*)] \quad (51)$$

The formula is used here

$$\begin{aligned} d\tilde{\omega}(A^*, B^*) &= e_r d\tilde{\omega}^r(A^*, B^*) \\ &= e_r A^*(\tilde{\omega}^r(B^*)) - e_r B^*(\tilde{\omega}^r(A^*)) - e_r \tilde{\omega}^r([A^*, B^*]) \\ &= e_r A^*(\tilde{\omega}^r(B^*)) - e_r B^*(\tilde{\omega}^r(A^*)) - \tilde{\omega}([A^*, B^*]) \\ &= A^*(e_r \tilde{\omega}^r(B^*)) - B^*(e_r \tilde{\omega}^r(A^*)) - \tilde{\omega}([A^*, B^*]) \end{aligned}$$

$$\begin{aligned}
&= A^* (\tilde{\omega}(B^*)) - B^* (\tilde{\omega}(A^*)) - \tilde{\omega}([A^*, B^*]) \\
&= d\tilde{\omega}(B^*)(A^*) - d\tilde{\omega}(A^*)(B^*) - \tilde{\omega}([A^*, B^*]) \\
&= -\tilde{\omega}([A^*, B^*]) = -\tilde{\omega}([A, B]^*) = -[A, B] \\
&= -[\tilde{\omega}_p(A_p^*), \tilde{\omega}_p(B_p^*)]
\end{aligned} \tag{52}$$

and used definitions and equations:

$$\begin{aligned}
e_r A^* (\tilde{\omega}^r(B^*)) &:= e_r d\tilde{\omega}^r(B^*)(A^*) = de_r \tilde{\omega}^r(B^*)(A^*) \\
&= d\tilde{\omega}(B^*)(A^*) = A^* (\tilde{\omega}(B^*))
\end{aligned} \tag{53}$$

Also $\tilde{\omega}(B^*)$, $\tilde{\omega}(A^*)$ is a constant scalar field with a value of $\mathcal{G}$,
 $\therefore d\tilde{\omega}(B^*) = 0$, $d\tilde{\omega}(A^*) = 0$.

c) $X \in V_p, Y \in H_p \Rightarrow X = A_p^*, Y = \tilde{Z}_p$ ($\tilde{Z}_p$ that is, the value of a horizontal lift field at p), then one gets

$$\begin{aligned}
d\tilde{\omega}|_p(0, \cdot) &= 0 = d\tilde{\omega}|_p(A_p^*, \tilde{Z}_p) + [\tilde{\omega}_p(A_p^*), \tilde{\omega}_p(\tilde{Z}_p)] \\
&= d\tilde{\omega}|_p(A_p^*, \tilde{Z}_p) + [A, 0] = d\tilde{\omega}|_p(A_p^*, \tilde{Z}_p) \\
&= A_p^* (\tilde{\omega}_p(\tilde{Z}_p)) - \tilde{Z}_p (\tilde{\omega}_p(A_p^*)) - 0 \\
&= 0 - \tilde{Z}_p (\tilde{\omega}_p(A_p^*)) - 0 = 0
\end{aligned} \tag{54}$$

A theorem has been used here: $[A^*, \tilde{Z}] = 0$, and a vector field acting on a constant scalar field is 0, $\tilde{Z}(A) = 0$, so there is

$$\begin{aligned}
d\tilde{\omega}(A^*, \tilde{Z}) &= A^* (\tilde{\omega}(\tilde{Z})) - \tilde{Z} (\tilde{\omega}(A^*)) - \tilde{\omega}([A^*, \tilde{Z}]) \\
&= A^* (\tilde{\omega}(\tilde{Z})) - \tilde{Z} (\tilde{\omega}(A^*)) \\
&= 0 - \tilde{Z} (\tilde{\omega}(A^*)) = -\tilde{Z}(A) = 0
\end{aligned} \tag{55}$$

q.e.d.

Furthermore, it can be proved that:

$$D\tilde{\Omega} = 0 \tag{56}$$

In fact, by definition (47), one gets,

$$D\tilde{\Omega} = (d\tilde{\Omega})^H \tag{57}$$

that is

$$\begin{aligned}
D\tilde{\Omega}(X, Y, Z) &= (d\tilde{\Omega})^H(X, Y, Z) \\
&= (d\tilde{\Omega})(X^H, Y^H, Z^H) = \frac{1}{2} d[\tilde{\omega}, \tilde{\omega}](X^H, Y^H, Z^H) \\
&= \frac{1}{2} \{[d\tilde{\omega}, \tilde{\omega}] - [\tilde{\omega}, d\tilde{\omega}]\}(X^H, Y^H, Z^H) = [d\tilde{\omega}, \tilde{\omega}](X^H, Y^H, Z^H) \\
&= \frac{1}{2} \{[d\tilde{\omega}(X^H, Y^H), \tilde{\omega}(Z^H)] + [d\tilde{\omega}(Z^H, X^H), \tilde{\omega}(Y^H)] \\
&\quad + [d\tilde{\omega}(Y^H, Z^H), \tilde{\omega}(X^H)] - [d\tilde{\omega}(X^H, Z^H), \tilde{\omega}(Y^H)]\}
\end{aligned}$$

$$\begin{aligned}
& -[d\tilde{\omega}(Z^H, Y^H), \tilde{\omega}(X^H)] - [d\tilde{\omega}(Y^H, X^H), \tilde{\omega}(Z^H)] \} \\
& = \frac{1}{2} \{ [d\tilde{\omega}(X^H, Y^H), 0] + [d\tilde{\omega}(Z^H, X^H), 0] + [d\tilde{\omega}(Y^H, Z^H), 0] \\
& \quad - [d\tilde{\omega}(X^H, Z^H), 0] - [d\tilde{\omega}(Z^H, Y^H), 0] - [d\tilde{\omega}(Y^H, X^H), 0] \} \\
& = 0
\end{aligned} \tag{58}$$

where

$$d\tilde{\Omega} = 0 + \frac{1}{2} d[\tilde{\omega}, \tilde{\omega}] = \frac{1}{2} ([d\tilde{\omega}, \tilde{\omega}] - [\tilde{\omega}, d\tilde{\omega}]) = [d\tilde{\omega}, \tilde{\omega}] \tag{59}$$

q.e.d.

And there is also a theorem that holds:

$$d\tilde{\Omega} = [\tilde{\Omega}, \tilde{\omega}] \tag{60}$$

Proof:

$$[\tilde{\Omega}, \tilde{\omega}] = [d\tilde{\omega}, \tilde{\omega}] + \frac{1}{2} [[\tilde{\omega}, \tilde{\omega}], \tilde{\omega}] = [d\tilde{\omega}, \tilde{\omega}] + 0 = d\tilde{\Omega} \tag{61}$$

Again by the previous definition:

$$\omega_U := \sigma_U^* \tilde{\omega}, \omega \equiv ke_r A_\mu^r dx^\mu = kA, A \equiv e_r A_\mu^r dx^\mu$$

one can introduce:

$$\Omega_U := \sigma_U^* \tilde{\Omega}, F \equiv \frac{1}{2} e_r F_{\mu\nu}^r dx^\mu \wedge dx^\nu \tag{62}$$

From this we can prove the theorem: Let $F \equiv \frac{1}{2} e_r F_{\mu\nu}^r dx^\mu \wedge dx^\nu$, then $\Omega_U = kF$, where k is a scale coefficient, and F can be defined as the gauge field strength.

Proof:

$$\begin{aligned}
\Omega &= d\omega + \frac{1}{2} [\omega, \omega] = d(ke_r A_\mu^r dx^\mu) + \frac{1}{2} [ke_s A_\mu^s dx^\mu, ke_t A_\nu^t dx^\nu] \\
&= ke_r dA_\mu^r \wedge dx^\mu + \frac{1}{2} k^2 [e_s, e_t] A_\mu^s A_\nu^t dx^\mu \wedge dx^\nu \\
&= \frac{1}{2} ke_r (\partial_\mu A_\nu^r - \partial_\nu A_\mu^r) dx^\mu \wedge dx^\nu + \frac{1}{2} k^2 C_{st}^r e_r A_\mu^s A_\nu^t dx^\mu \wedge dx^\nu \\
&= \frac{1}{2} ke_r (\partial_\mu A_\nu^r - \partial_\nu A_\mu^r + k C_{st}^r A_\mu^s A_\nu^t) dx^\mu \wedge dx^\nu \\
&= \frac{1}{2} ke_r F_{\mu\nu}^r dx^\mu \wedge dx^\nu = kF
\end{aligned} \tag{63}$$

here, $[e_s, e_t] = C_{st}^r e_r$, as well

$$\begin{aligned}
dA_\mu^r \wedge dx^\mu &= \frac{\partial A_\mu^r}{\partial x^\nu} dx^\nu \wedge dx^\mu = (\partial_\nu A_\mu^r) dx^\nu \wedge dx^\mu \\
&= (\partial_{[\mu} A_{\nu]}^r) dx^\mu \wedge dx^\nu = \frac{1}{2} (\partial_\mu A_\nu^r - \partial_\nu A_\mu^r) dx^\mu \wedge dx^\nu
\end{aligned} \tag{64}$$

The above proof shows that the defined $\tilde{\Omega}$ is the gauge field strength, and the defined $\tilde{\omega}$ is the gauge potential, which is constant. However, Ω and ω change with the cross section, and the corresponding transformation also cor-

responds to the transformation of the gauge potential. These transformations are caused by the transformation of the cross section (that is, the gauge transformation), but the “grand quantities” connection $\tilde{\omega}$ and curvature $\tilde{\Omega}$ in $(P, \tilde{\omega})$ and $(P, \tilde{\Omega})$ are constant, and the change are only their components $\omega \rightarrow \omega', \Omega \rightarrow \Omega'$, respectively.

5.2. Transformation Relation of Curvature

Under the cross section transformation σ , the transformation relationship of $\omega \rightarrow \omega'$ on the bottom manifold (that is, our world) is (25); what is the transformation relationship of $\Omega \rightarrow \Omega'$? To do this, we need to prove the following theorems:

$$(\text{Theorem A}) \quad \Omega_U = d\omega_U + \frac{1}{2}[\omega_U, \omega_U] \quad (65)$$

Proof:

$$\begin{aligned} \Omega_U &= \sigma_U^* \tilde{\Omega} = \sigma_U^* \left(d\tilde{\omega} + \frac{1}{2}[\tilde{\omega}, \tilde{\omega}] \right) \\ &= d(\sigma_U^* \tilde{\omega}) + \frac{1}{2}[\sigma_U^* \tilde{\omega}, \sigma_U^* \tilde{\omega}] \\ &= d\omega_U + \frac{1}{2}[\omega_U, \omega_U] \end{aligned} \quad (66)$$

(Theorem B) Equation (63) corresponds to Equation (20), namely

$$\left\{ \begin{array}{l} \Omega = d\omega + \frac{1}{2}[\omega, \omega] \\ \Downarrow \\ F_{\mu\nu}^r(x) = \partial_\mu A_\nu^r - \partial_\nu A_\mu^r + k \sum_{s,t=1}^R C_{st}^r A_\mu^s A_\nu^t(x), (r=1, \dots, R) \end{array} \right. \quad (67)$$

Proof: the same as Equation (63),

$$\begin{aligned} \Omega &= k e_r d(A_\mu^r dx^\mu) + \frac{1}{2}[e_s, e_t] k^2 A_\mu^s A_\nu^t dx^\mu \wedge dx^\nu \\ &= \frac{1}{2} k e_r (\partial_\mu A_\nu^r - \partial_\nu A_\mu^r + C_{st}^r e_s e_t k^2 A_\mu^s A_\nu^t) dx^\mu \wedge dx^\nu \\ &= \frac{1}{2} k e_r F_{\mu\nu}^r dx^\mu \wedge dx^\nu \\ \therefore F_{\mu\nu}^r &= \partial_\mu A_\nu^r - \partial_\nu A_\mu^r + C_{st}^r e_s e_t k^2 A_\mu^s A_\nu^t \end{aligned}$$

Here Equations (62) and (63) are considered.

(Theorem C) Let $g_{UV}: U \cap V \rightarrow G$ be the local trivial transformation function from T_U to T_V , then on $U \cap V$ of the bottom manifold one gets

$$\Omega_V = \mathcal{A} d_{g_{UV}^{-1}} \Omega_U \quad (68)$$

Proof:

$\forall x \in U \cap V, X, Y \in T_x M$, one has

$$\Omega_V(X, Y) = (\sigma_V^* \tilde{\Omega})(X, Y) = \tilde{\Omega}(\sigma_V^* X, \sigma_V^* Y) \quad (69)$$

Again considering Equations (31) and (32), one gets

$$\sigma_V^* X = R_{g_{UV}(x)}^* \sigma_U^* X + \left[L_{g_{UV}(x)}^{-1} g_{UV}(X) \right]_{\sigma_V(x)}^* = X_1 + X_2 \quad (70)$$

$$\sigma_V^* Y = R_{g_{UV}(x)}^* \sigma_U^* Y + \left[L_{g_{UV}(x)}^{-1} g_{UV}(Y) \right]_{\sigma_V(x)}^* = Y_1 + Y_2 \quad (71)$$

Hence Equation (69) changes

$$\begin{aligned} \Omega_V(X, Y) &= \tilde{\Omega}(X_1 + X_2, Y_1 + Y_2) \\ &= \tilde{\Omega}(X_1, Y_1) + \tilde{\Omega}(X_1, Y_2) + \tilde{\Omega}(X_2, Y_1) + \tilde{\Omega}(X_2, Y_2) \end{aligned} \quad (72)$$

The action objects of the last three items on the right side of the above formula involve the vertical vector X_2 or Y_2 , so they are all 0, so it can be obtained

$$\begin{aligned} \Omega_V(X, Y) &= \tilde{\Omega}(X_1, Y_1) = \tilde{\Omega}(R_{g_{UV}(x)}^* \sigma_U^* X, R_{g_{UV}(x)}^* \sigma_U^* Y) \\ &= (R_{g_{UV}(x)}^* \tilde{\Omega})(\sigma_U^* X, \sigma_U^* Y) = (Ad_{g_{UV}^{-1}} \tilde{\Omega})(\sigma_U^* X, \sigma_U^* Y) \\ &= Ad_{g_{UV}^{-1}}(\sigma_U^* \tilde{\Omega})(X, Y) = Ad_{g_{UV}^{-1}} \Omega_U(X, Y) \end{aligned} \quad (73)$$

here, the theorem is used: $R_g^* \tilde{\Omega} = Ad_{g^{-1}} \tilde{\Omega}, \forall g \in G$.

q.e.d.

(Theorem D) If the structure group is a matrix group, formula (68) can be further expressed as:

$$\Omega_V = g_{UV}^{-1} \Omega_U g_{UV} \quad (74)$$

Therefore, under the cross section transformation σ , the transformation relationship of $\omega \rightarrow \omega'$ on the bottom manifold (that is, the world we assume) is the Equation (25), and the corresponding transformation relationship of $\Omega \rightarrow \Omega'$ is the Equation (68)!

5.3. Gravitational Gauge Field

Physics has roughly unified the three fundamental interactions of electromagnetism, weak and strong, but the fundamental interaction of gravitation has not been unified, which makes us pay special attention to the problem of the gravitational gauge field. 1) Is gravity a gauge field? The answer is: suppose it is. Many authors have proposed gravitational gauge theory in this regard, such as Duan Yishi's general theory of relativity and gravitational gauge theory and Yue-Liang Wu's unified theory of gauge field [25] [28]; 2) the general gauge equation proved above (25) and the corresponding gauge potential in Equation (68) also include the gravitational gauge potential? The answer is yes! In fact, in Chapter 6 and Appendix 3 of Duan Yishi's book, the author has proved relatively well that in the principal fiber bundle where the $SO(n)$ group is the structural group, the connection is the gravitational gauge potential, and the curvature is the gravitational gauge field strength. And there is a very classic saying, "The gauge theory of physics is just the principal fiber bundle (principal bundle) theory in mathematics" [25]. Considering these circumstantial evidences, the author of this paper confirms that Equations (25) and (68) are some forms of GGE

(generalized gauge equations) that are generally valid for the gauge potential of fundamental interaction; the connection of space-time is the gauge potential, and the curvature of space-time is the gauge field strength; no matter for gravitation or for electromagnetic force and weak and strong interaction forces, Equations (25) and (68) are true for the four fundamental interactions.

So under the condition that the structure group is large enough, the generalized gauge Equation (GGE) can express the transformation of the gauge potential on the bottom manifold (our universe) from one fundamental interaction gauge potential to another fundamental interaction gauge potential, such as the transformation from electromagnetic gauge potential to gravitational gauge potential, while the universal unified gauge potential on the principal bundle (similar to the “heaven” of high-dimensional space-time) is unchanged, these local gauge potentials on the base manifold are just the projection components of the choice of the unified gauge potential under different gauge (or the cross section of principal bundle). However, this transformation is not only the transformation within the same gauge field as understood in the traditional view, but may be the transformation between different gauge fields, or even the transformation between gauge potentials with cross fundamental interactions. We can call this transformation from gauge potentials with one fundamental interaction to gauge potentials with another fundamental interaction as “cross fundamental gauge potential transformation”. This is exactly the meaning of the great unity of physics revealed by Equation (25) and the corresponding Equation (68); it is also the outline significance of the grand unification of the cosmic gauge field revealed by the structural picture of the principal associated bundles of the universe.

5.4. Connection of Entangled States

The above “grand unified outline” can first be used to explain the “paradox” of quantum entangled states. What is quantum entanglement first? The rough answer of quantum theory is that if the wave functions of two particles cannot be written as the scalar product of two wave functions, then the two particles are entangled. Fan Hongyi believed in the book [29] that the entangled state is the common eigenstate of the commutator $P_1 \pm P_2$ and $Q_1 \mp Q_2$, i.e.

$[P_1 \pm P_2, Q_1 \mp Q_2]$, here the momentum operator of two particle is respectively P_1 or P_2 and the position operator of two particle is respectively as Q_1 or Q_2 . Hence for $[P_1 + P_2, Q_1 - Q_2]$, it can be expressed as

$$\begin{cases} (Q_1 - Q_2)|\eta\rangle = \sqrt{2}\eta_1|\eta\rangle \\ (P_1 + P_2)|\eta\rangle = \sqrt{2}\eta_2|\eta\rangle \end{cases} \quad (75)$$

Here, the form of the entangled state $|\eta\rangle$ is

$$|\eta\rangle = \exp\left(-\frac{1}{2}|\eta|^2 + \eta a_1^\dagger - \eta^* a_2^\dagger + a_1^\dagger a_2^\dagger\right)|00\rangle \quad (76)$$

Among them, $\eta = \eta_1 + \eta_2$, $Q_j = \frac{a_j + a_j^\dagger}{\sqrt{2}}$, $P_j = \frac{a_j - a_j^\dagger}{\sqrt{2}i}$; and $|\eta\rangle$ satisfies the completeness relation and is also orthogonal.

One of the strangeness of the entangled state is hidden in the Schmidt decomposition in its coordinate representation:

$$|\eta\rangle = e^{-i\eta_1\eta_2} \int_{-\infty}^{+\infty} dq |q\rangle_1 \otimes |q - \sqrt{2}\eta_1\rangle_2 e^{i\eta_2 q} \quad (77)$$

The above formula shows that when particle 1 is in its coordinate eigenstate $|q\rangle_1$, particle 2 is simultaneously in its coordinate eigenstate $|q - \sqrt{2}\eta_1\rangle_2$, and these two particles have a distance, and the distance may even be great. Why does the state change of particle 1 cause the state of particle 2 to change correspondingly “simultaneously”? From the general gauge Equation (25), if all the interactions in the universe are regarded as the interaction of the gauge field and a kind of space-time regional force, then the various strange performances of the quantum entanglement state are easy to understand, so our answer is as follows:

1) The so-called quantum entanglement means that the space-time of two particles establishes a connection, thus a gauge potential appears, and a gauge field acts on the two particles.

2) This connection is a special connection, the regional gauge potential, which makes the sum of the two particles' momentum and the difference of their coordinates limit the common eigenstates of the two particles in this region, and as a whole, it is affected by the regional gauge field at the same time. This is different from the common sense of connection or gauge potential. The difference lies in the “simultaneous regional integration”.

3) Any operation to change the state of one of the particles is equivalent to changing the structure of the gauge potential in this area at the same time, which must affect the role of all gauge fields in this area at the same time, and must also affect the corresponding state of another particle.

In a word, quantum entanglement is to establish the gauge field of the whole region between two particles, so that the whole of the two particles are affected by the gauge field at the same time. Any change in the state of one particle must change the state of the other particle at the same time. Here, the gauge effect is simultaneous for the whole region, without the meaning of time and transmission speed. In fact, the quantum entangled state may be the common eigenstate of the operator of the relative distance between the two quanta and the total momentum operator of the two quanta, which indicates that the relative distance between the two quanta can change in the space time region where the entanglement occurs, and their total momentum can also change, but the region of the space time regional force that causes the entanglement will also change, still covering the entire region where the entangled state is located. Therefore, any “operation or measurement” behavior that changes a single quantum state will “simultaneously” cause the gauge field force of the whole space-time region enveloped to act on another entangled quantum to change its state. There is no

superluminal force transmission, which is the concept of assumption that the role of this special global gauge field force is still considered as a local part.

6. Conclusions and Outlook

A) In this paper, a general unified program of physics is proposed, and the unified formula is the generalized gauge equation GGE or Equation (25). All interactions in the world are unified in the gauge potential (gauge field) in the universe picture of the principal associated bundles, and the fundamental interaction of our universe on the bottom manifold is just the representation of the component of the gauge potential of this principal bundle, which follows the transformation of the generalized gauge equation from one component to another, or even the transformation of one fundamental interaction into another.

B) The invariance of the gauge transformation or the satisfaction of the generalized gauge Equation (EEG) is a necessary condition for the universal unified field theory, but the quantization of the field is not a necessary condition for the universal unified field theory.

C) Outlook: 1) The fundamental interactions may not be limited to four types, and the discovery of new fundamental interactions is possible. 2) The fundamental interaction may be transformed into each other, and the basic equation of transformation is just GGE. 3) It is found that the structure group can express more gauge field components; so it is an important task for physics in the future to simplify and solve EEG so that one can concretely express the transformation relationship between any two gauge field components, especially the transformation relationship between electromagnetic force and gravity (which is extremely important for solving human aerospace dynamics).

Acknowledgements

The author would like to express his high respect to Professor Canbin Liang!

Conflicts of Interest

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

References

- [1] Yang, C.N. and Mills, R.L. (1954) *Physical Review*, **96**, 191-195.
<https://doi.org/10.1103/PhysRev.96.191>
- [2] 't Hooft, G. and Veltman, M. (1972) *Nuclear Physics B*, **44**, 189.
[https://doi.org/10.1016/0550-3213\(72\)90279-9](https://doi.org/10.1016/0550-3213(72)90279-9)
- [3] Higgs, P. (1964) *Physical Review Letters*, **13**, 508.
<https://doi.org/10.1103/PhysRevLett.13.508>
- [4] Englert, F. and Brout, R. (1964) *Physical Review Letters*, **13**, 321.
<https://doi.org/10.1103/PhysRevLett.13.321>
- [5] Aquino, F.W. and Wong, B.M. (2018) *The Journal of Physical Chemistry Letters*, **9**, 6456-6462. <https://doi.org/10.1021/acs.jpclett.8b02786>

- [6] Gross, D.J. and Wilczek, F. (1973) *Physical Review Letters*, **30**, 1343.
<https://doi.org/10.1103/PhysRevLett.30.1343>
- [7] Politzer, H.D. (1973) *Physical Review Letters*, **30**, 1346.
<https://doi.org/10.1103/PhysRevLett.30.1346>
- [8] Georgi, H. and Glashow, S.L. (1974) *Physical Review Letters*, **32**, 438.
<https://doi.org/10.1103/PhysRevLett.32.438>
- [9] Georgi, H. (1975) *Particles and Fields 1974*. Amer. Inst. of Physics, New York.
- [10] Fritzsche, H. and Minkowski, P. (1975) *Annals of Physics*, **93**, 193.
[https://doi.org/10.1016/0003-4916\(75\)90211-0](https://doi.org/10.1016/0003-4916(75)90211-0)
- [11] Hehl, F.W., von der Heyde, P., Kerlick, G.D. and Nester, J.M. (1976) *Reviews of Modern Physics*, **48**, 393. <https://doi.org/10.1103/RevModPhys.48.393>
- [12] Ivanenko, D. and Sardanashvily, G. (1983) *Physics Reports*, **94**, 1.
[https://doi.org/10.1016/0370-1573\(83\)90046-7](https://doi.org/10.1016/0370-1573(83)90046-7)
- [13] Hehl, F., McCrea, J., Mielke, E. and Neeman, Y. (1995) *Physics Reports*, **258**, 1.
- [14] Polchinski, J. (2005) *String Theory*. Vol. 1, 2, Cambridge University Press, Cambridge.
- [15] Kaku, M. (1999) *Introduction to Superstrings and M-Theory*. Springer, New York.
<https://doi.org/10.1007/978-1-4612-0543-2>
- [16] Farnes, J.S. (2018) *Astronomy and Astrophysics*, **620**, A92.
<https://doi.org/10.1051/0004-6361/201832898>
- [17] Olive, K.A. (2014) *Review of Particle Physics, Chinese Physics C*, **38**, Article ID: 090001. <https://doi.org/10.1088/1674-1137/38/9/090001>
- [18] Chou, K.C. and Wu, Y.L. (1998) *Science China A*, **41**, 324.
<https://doi.org/10.1007/BF02879052>
- [19] Percacci, R. (1984) *Physics Letters B*, **144**, 37.
[https://doi.org/10.1016/0370-2693\(84\)90171-0](https://doi.org/10.1016/0370-2693(84)90171-0)
- [20] Percacci, R. (1991) *Nuclear Physics B*, **353**, 271.
[https://doi.org/10.1016/0550-3213\(91\)90510-5](https://doi.org/10.1016/0550-3213(91)90510-5)
- [21] Nesti, F. and Percacci, R. (2010) *Physical Review D*, **81**, Article ID: 025010.
<https://doi.org/10.1103/PhysRevD.81.025010>
- [22] Nielsen, M.A. and Chuang, I.L. (2000) *Quantum Computation and Quantum Information*. Cambridge University, Cambridge.
- [23] Ren, J.G., Xu, P., Yong, H.-L., *et al.* (2017) *Nature*, **549**, 70-73.
<https://doi.org/10.1038/nature23675>
- [24] Chen, X.S. and Chen, W.H. (1983) *Handout on Differential Geometry*. Peking University Press, Beijing.
- [25] Duan, Y.S. (2020) *General Relativity and Gravity Gauge Theory*. Science Press, Beijing.
- [26] Lian, C.B. and Zhou, B. (2019) *Introduction to Differential Geometry and General Relativity*. Second Edition, Vol. I, II, III, Science Press, Beijing.
- [27] Peskin, M.E. and Schroeder, D.V. (2018) *An Introduction to Quantum Field Theory*. CRC Press, Boca Raton. <https://doi.org/10.1201/9780429503559>
- [28] Wu, Y.-L. (2018) *The European Physical Journal C*, **78**, 28.
<https://doi.org/10.1111/tme.12559>
- [29] Fan, H.Y. (2012) *Supplement to Quantum Mechanics Textbooks for Postgraduates*. China University of Science and Technology Press, Hefei.

Disjointed Equivalence of Gravitational and Inertial Mass

Dirk J. Pons 

Department of Mechanical Engineering, University of Canterbury, Christchurch, New Zealand
Email: dirk.pons@canterbury.ac.nz

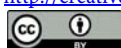
How to cite this paper: Pons, D.J. (2023) Disjointed Equivalence of Gravitational and Inertial Mass. *Journal of Modern Physics*, 14, 237-270.
<https://doi.org/10.4236/jmp.2023.143017>

Received: November 23, 2022

Accepted: February 10, 2023

Published: February 13, 2023

Copyright © 2023 by author(s) and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

Problem—Contemporary physics offers no underlying reason for the equivalence of inertial and gravitational mass. **Approach**—The equivalence is examined from the new physics provided by the cordus theory, being a non-local hidden-variable (NLHV) theory. Mathematical formalisms are derived for masses and observers in different fabric densities. **Findings**—A disjointed equivalence is predicted, whereby inertial and gravitational masses are equivalent in any one situation, but a different equivalence holds when the fabric densities change. Consequently this theory predicts that the gravitational constant G varies with fabric density, and hence would be different across the universe and across time. Not only is the gravitational constant non-constant, but the formulation of gravitation changes with fabric density. Specifically, the theory predicts gravity is stronger at genesis (and the end of the universe) such that orbit velocity $v_B \propto \sqrt{r_B}$ (where r_B is orbit radius), compared to weaker gravitation at middle life epochs with $v_B \propto \frac{1}{\sqrt{r_B}}$. The

current Earth location and epoch correspond to the latter case, *i.e.* Newtonian gravitation is recovered. The findings disfavour the existence of both dark energy and dark matter, and instead attribute these effects to differences in the fabric density. **Originality**—The work makes the contribution of deriving a mass equivalence relationship that includes fabric density, identifying a disjointed mass equivalence, and showing that the gravitation formulation itself changes with relative fabric densities.

Keywords

Identity of Mass, Gravitation, Inertia, General Relativity, Quantum Mechanics

1. Introduction

A deep issue for foundational and cosmological physics is the lack of an under-

lying principle for the equivalence of inertial and gravitational mass. The equivalence states that mass as determined from inertial behaviour (m_I), *i.e.* the resistance of a body to acceleration, is the same as determined from gravitational interaction with another body (m_G), where:

$$m_I = \frac{F_I}{a} \quad (1)$$

and

$$m_G = \frac{F_G r^2}{GM} \quad (2)$$

with force F , acceleration a , separation r between masses m and M , and gravitational constant G .

This equivalence is problematic since it seems an unnecessary coincidence. It is an assumption of Newtonian gravitation and general relativity (GR), but is not explained by them, nor by quantum mechanics (QM).

The present paper examines the equivalence from the non-local hidden-variable (NLHV) perspective, using the particle sub-structure proposed by the cordus theory [1]. We show that under this new physics the equivalence holds for each local gravitational frame, but is not universally identical in all situations. The key situational variable is identified as the fabric density. This is a vacuum attribute (described below) that has previously been identified as contributing to relativistic time dilation [2].

2. Background

Philosophically, there have been different ways to approach the equivalence problem. Newtonian and Einsteinian theories accepted the equivalence as real, and sought—unsuccessfully—the foundational principles thereof. Important points to consider in the debate are the multiple ways in which mass has physical effects, the origins of mass at a particle level (which primarily relates to the quantum theory interpretations), and the origin of gravitation (which relates to general relativity).

2.1. Multiple Attributes of Mass

Mass is the coupling for multiple physical effects that otherwise might be independent [3]:

1) Two masses attract each other (gravitational mass). The gravitation force or interaction has an unusual set of properties compared to the other forces: a) it only acts on particles with mass or energy; b) it always attracts, never repels (at least for matter-matter interactions, with antimatter it may be different [4]); c) it has infinite range; and d) it cannot be redirected or shielded. Mass is the fundamental strength variable for gravitation.

2) Resistance to acceleration (Newtonian or inertial mass). The greater the acceleration a or mass m of a body, the greater the force F required to change its speed v :

$$F = ma \quad (3)$$

or more generally

$$F = \frac{d(mv)}{dt} \quad (4)$$

3) Relativistic mass: as the speed v of a body of rest mass m_o approaches that of light c , so the effective mass m tends to infinity, or at least the resistance to acceleration does:

$$m = m_o \left(1 - \frac{v^2}{c^2} \right)^{0.5} \quad (5)$$

This effect applies even if there is no acceleration.

4) From the perspective of relativity, momentum p is a separate property to mass and the full energy-momentum formula is:

$$E = \left((pc)^2 + (m_o c^2)^2 \right)^{0.5} \quad (6)$$

5) Mass originates with particles, e.g. protons and neutrons (among others including the photon). There is no deeper explanation for why particles have the mass they do.

2.2. Origins of Mass

What determines the mass property of a particle? It is unclear whether mass is an intrinsic property of the particle, e.g. Copenhagen interpretation, or based on some other attributes. Neither QM nor GR directly address the question of the origin of mass. They also lack a common treatment of gravitation, having independent approaches based on particles and space-time. Much existing work has focused on proposing candidate mechanisms for the origins of mass for specific fundamental particles, with a particular focus on the see-saw mechanism for neutrino species e.g. [5] [6] [7] [8] [9], and quarks e.g. [10] [11] [12] [13], however none provide a comprehensive explanation. Mass arises, per QM, by interaction of the particle with a Higgs field [14] [15], the electroweak symmetry of which is broken at suitably low temperatures [16]. The particle that mediates this field is the Higgs boson [11] [17] [18] [19] [20]. The CERN findings of 2012-4 produced a signature at 125GeV consistent with such a boson [21] [22] [23]. However the mechanism for production of the Higgs boson is still unclear [24] [25]. There has been an expectation that new physics could arise from this direction [26], but this has been elusive. This is problematic because the mass generation mechanism is incompletely specified [27] [28], the mechanisms for stabilisation of the Higgs mass are unknown [23] [29], and there are discrepancies in the sizes of the fundamental interactions – the hierarchy problem [30] [31]. Possibly a deeper physics is needed [32], but if so it is unclear in what direction that lies.

A further problem is how mass arises as the aggregation of such particles. Related questions are the hierarchical structure of mass in the assembly of particles

(e.g. quarks into nucleons) [33] [34] [35], and the effect of binding energy or mass excess. Quantum chromodynamics predicts that the nucleon mass is caused by the breaking of chiral symmetry, but the process itself is unknown [36].

There are several other speculative lines of enquiry into the origins of mass, though none provide a compelling solution. These include attempts to identify the mass generation mechanisms and causes of electroweak symmetry breaking [27] [37] [38] [39], theories beyond the standard model [40], supersymmetry [30] [41] [42] [43] [44], extended or multi-Higgs frameworks [45] [46] [47] [48] [49], gravitons [50], axions [51], inflatons & dark energy [52] [53], connection of Higgs and dark matter sectors [28] [54] [55], analogy of Higgs with fluid resistance [56], technicolor particles [57], baryogenesis [54] [58], dynamics of mass-energy equivalence [59] [60], and whether there may be repulsion (negative mass) interactions between matter and antimatter [61]. Theories of mass also have implications in other areas such as variable speed of light theories [62] [63] and other gauge theories [64].

However the basic problem remains—the origins of mass are unknown for both inertial and gravitational mass. Additionally, it is unknown what attributes or structures of the particle represent the momentum vs. those represent the mass. Quantum theory is built on the premise that fundamental particles have no physical substructure and hence such questions can not be comprehended ontologically from within that theory. It is possible that QM may not have the requisite architecture to address these deeper questions, and hence there is value in exploring alternative paradigms.

2.3. Origins of Gravitation

Newtonian gravitation and then Einstein's general relativity (GR) represent gravitation as a continuous field effect, the latter based on the curvature of space-time. Hence space-time becomes an integrated structure. A foundational premise of GR is the *equivalence principle*: that gravitational and inertial effects are indistinguishable. This is related to the *relativity principle* that all inertial frames of reference are indistinguishable—no frame, neither its orientation nor velocity, is preferred to another. Hence general relativity requires, rather than proves, the equivalence of gravitational and inertial frames of reference [65]. This was Einstein's key insight and the principle of departure from Newtonian gravitation. One view might be that there has been no empirical refutation of the mass equivalence principle, hence the premise is worthy of elevation to a principle. Certainly there are not many theories that offer plausible alternatives. However it is still an assumption, and difficult to verify. Furthermore, there are conceptual difficulties with GR: the inability to describe singularities in space-time; and the lack of a particle interpretation that is consistent with quantum mechanics.

Possibly gravitation might in the future be recast as a particle-based force us-

ing virtual bosons. However a robust integration of gravitation with the other forces has not yet been achieved this way. Even so, this approach seems to require the abolition of the equivalence. GR has empirical shortcomings too, in the inability to explain anomalous rotation curves of galaxies [66] [67]. The theoretical consensus is to attribute this phenomenon to dark matter. Many different kinds have been proposed, but none have been empirically observed [68] despite considerable efforts and cost to do so.

A contrary and minority view is that anomalous galaxy rotation curves already provide an empirical confutation of mass equivalence, and that gravitation takes an alternative form. There is no absence of alternative theories. A candidate is modified Newtonian dynamics (MOND) [69] [70], which also requires the equivalence to be broken. MOND provides good quantitative fit to observations, but lacks an underpinning theory of causation. Another active area of theoretical exploration is torsional gravity [71]. There are several variants of this including teleparallel and metric-affine gauge formulations [68]. While these approaches have promise and can explain features of cosmology [68] [72], there is as yet no complete solution nor reconciliation with particle physics. Furthermore, they lack a deeper ontological explanation for the effect.

An alternative approach sees the equivalence as purely coincidental, per the anthropic principle or the multiverse [73]. That has its own difficulties as it relies on untestable metaphysics. Also, it does not explain the principles which set the presumably different ratios in the many universes.

2.4. Contributions from Hidden Variable Theories

The hidden-variable (HV) theories take a conceptual and logical approach to theory-development, and sometimes provide the vision and leadership for mathematical and empirical methods to follow afterwards. These theories tend to be premised on the assumption that particle phenomena that objectively exist are brought about by the existence of physical mechanisms at a deeper level of structure, and this implies that particles have sub-structure. This is a long-standing idea [74] that corresponds to the philosophical premise of physical realism [75]. Einstein and others expected HV theory to provide a deeper mechanics beneath QM [74], however this has been difficult to achieve.

There are two classes of HV theories: *local* and *non-local*. Locality in this context refers to the relationship between a particle and its environment: that a point object is only affected by fields at that location, not by remote values; and transmission of an effect occurs by at most the speed of light [76]. In contrast *non-locality* allows remote fields to affect the particle, and for these effects to be superluminal [77]. It is unlikely that *local* theories are viable, as they are incongruent with observable entanglement phenomena. The Bell inequalities [78] provide a mathematical formulation of this logic. However the validity of *non-local* theories is indeterminate. Some at least are excluded by the Bell-type inequalities, but not all [77]. There is a view in some parts of the literature, exem-

plified in the Colbeck and Renner (C&R) paper [79], that all types of hidden variable theories are excluded, *i.e.* that particles are zero dimensional (0-D) points without any substructure. Hence also that the attributes of the particle (mass, charge, etc.) are merely intrinsic rather than based on physical structures. However the conclusions of the C&R paper have been criticised as logically unsound [77]. Stripped of the mathematics, the logical structure of the C&R paper is a starting assumption that particles are 0-D points, that locality exists, and that quantum theory is correct. Then by deductive reasoning it was inferred that particles are 0-D points such that no substructure can exist. However that is an unsound conclusion because the output statements were already self-evident in the starting premises—the argument was self-confirming.

Even so, successful non-local hidden-variable (NLHV) theories are scarce. There have been many ideas [80] but few successful ones. Historically the most important is the de Broglie – Bohm pilot-wave theory [81] [82]. However this has not developed into a comprehensive new theory of physics, and does not extend to gravitation or the equivalence question.

A more recent NLHV theory is the cordus theory [83] which proposes that particles have a specific two-ended structure, hence *cordus*. This is not inconsistent with string theory, and the number of parameters required to define a cordus particle is the same as some variants of string theory, despite coming at the problem from completely different directions. The theory makes specific predictions about the structures at the sub-particle level (reactive ends, fibril, discrete forces), and proposes mechanisms whereby these manifest the physical behaviour of the particle as a whole [4]. The theory has been applied to explain a wide variety of phenomena [83]. Of specific relevance to the equivalence question are its explanation and mathematical representation for particle motion [84], the relativistic Doppler & time dilation [2], the identity of matter [58], and gravitation with unification of the interactions [4] [85]. It explains the gravitation field as a torsionally handed emission of discrete forces from particles [2] [4] [84] [85]. This has some parallels with torsional gravity theory, specifically how the gradient of the field arises [68]. In torsional gravity theories [68] [72], the torsional effect is attributed to spin [86] but these theories lack a deeper ontological explanation for the effect. Nor does spin have any physical representation in quantum particle theory. In contrast the cordus theory provides a physical explanation for spin [83] and torsion in field emissions [4]. Theories that involve both torsion and curvature are underpinned by absolute parallelism geometry [87], with coefficients that take discrete values reminiscent of quantum behaviour [88] [89], where the torsion term represents the interaction between the particle and the gravitational field [90]. The idea that the spin of a particle interacts in discrete steps with gravitational field was proposed on mathematical grounds by [89], and this is consistent with the cordus theory that also proposes a stepped type of motion (“gait”) [4]. The cordus theory is also compatible with the quantum hypothesis of the graviton, in that the torsional packet of discrete

forces is analogous to a quantised variable corresponding to the graviton. The theory includes the concept of fabric (described below) which is compatible (but also extends) the general relativity concept of space-time, and the quantum idea of quantum foam. Consequently it is interesting to examine what it implies for mass-equivalence.

In summary, the identity of mass is an important topic at foundational and cosmological levels. None of the existing theories satisfactorily explain why there should be an equivalence of gravitational and inertial mass, or even how mass arises.

3. Method

3.1. Objective

The purpose of this work was to explore the mass equivalence problem from the perspective of the new physics provided by the cordus theory, which is a type of non-local hidden-variable theory [1]. This is worth attempting for the potential to provide different insights to the problem.

3.2. Approach

A conceptual approach was taken, based on logical extension of the existing cordus theory. There was no a-priori expectation of whether equivalence did or did not hold. Instead we sought to identify the mechanisms that cause mass in this theory, and then infer the implications for equivalence.

This development used *inductive* reasoning. Unlike *deductive* reasoning which proceeds from explicit premises to decisive conclusions, the inductive approach starts with a limited set of premises and generalises to wider implications. One of the quality tests for this inductive methodology is whether it produces new principles that are congruent with principles arising from advancement of the theory in other directions. The anticipatory seeking of this congruence means that candidate new principles have to be checked for consistency with other aspects of the theory. In practice this means that the current results were reviewed, during development, by comparison with findings from all the other published work on the cordus theory [1] [2] [4] [58] [77] [83] [84] [85] [91]-[107]. No logical discontinuities were found. Hence the current findings are cogent with the wider cordus theory. Results are mathematical formalised, but the underlying approach is not primarily a mathematical one.

The precursor to this work was the cordus theory for the Lorentz transformation [2]. This derives the relativistic Doppler, Lorentz formulation, and time dilation from first principles from a particle perspective, which is novel. An unexpected term appeared in the formulation, in the form of the fabric density. Fabric density [93] [96] [102] [105] is a concept within the cordus theory that corresponds approximately to vacuum properties in electromagnetism.

For logical self-consistency within the cordus theory it is necessary to take the premise that mass is a contingent attribute. There is a basis to it determined by

particle identity, but the final expressed value is contingent on other variables. From this we reasoned that fabric density (to be explained) is one of those contingent variables. In the next step, we considered the implications of such an identity for mass. We concluded that a body ought to behave in internally consistent ways in both inertial and gravitational cases, because the underlying mechanisms of mass are related to a common cause in the discrete forces. This provided a rationale for believing the equivalence should hold, at least in any one fabric situation. A single fabric situation refers to all the masses and observers being immersed in the same fabric.

We then generalised to relative fabric situations where the masses and observers are in different fabric densities, and determined the corresponding mathematical formalisms. Mass interactions at cosmological scale, e.g. galaxy rotation curves, are in this category. We find that the fabric density changes the nature of the equivalence in a specific way. We conclude by exploring the implications of this.

4. Results

4.1. The Identity of Mass

4.1.1. Particle Emissions

In the cordus theory a fundamental particle comprises two reactive ends connected by a fibril, with the reactive ends energising in turn and emitting in three orthogonal directions when they do. Individual emissions comprise a sinusoidal varying pulse [84], hence these emissions may be approximated as “discrete forces”, “flux tubes”, or “continuous fields” depending on the context. The particle sub-structures for the electron and proton are shown in **Figure 1**. The emissions from a particle radially outwards and are diluted over a front comprising an expanding spherical area [85]. Hence the $1/r^2$ dependency of strength of field with separation distance r , which is a characteristic feature of Newtonian mechanics. The theory proposes that electrostatic force is carried by the direct linear action of these discrete forces, the magnetic by the bending of the flux tube, and the gravitational by the handedness of the three emissions, hence a unification is achieved [85].

4.1.2. Identity of Mass

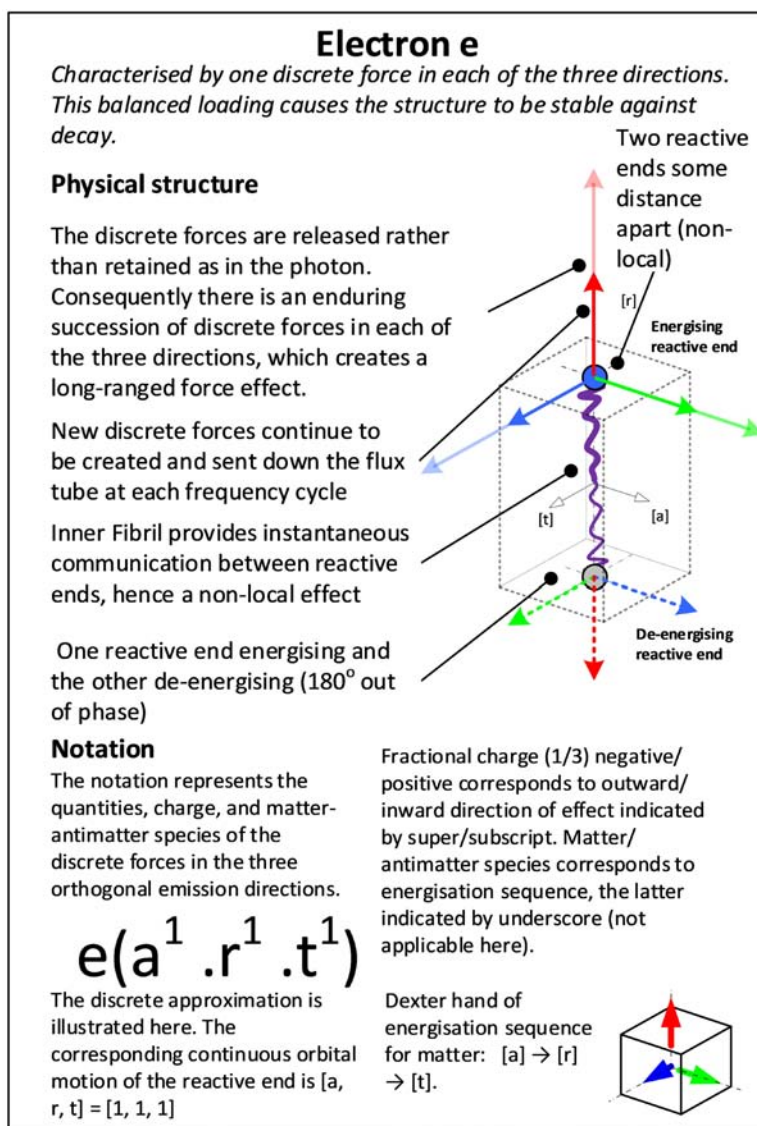
At the macroscopic level, mass appears to be a fixed attribute determined strictly by the volume and density of the body, and ultimately by the number of protons, neutrons, and electrons in the body. However, the cordus theory explains mass differently, as the number of discrete force emissions convoluted with the frequency of the particle. The underlying principles are inferred as:

- §1 Heavier particles are those with more discrete forces to emit.
- §2 The theory also anticipates that having more discrete forces requires that the particle adopt a higher frequency to service the emissions, so the addition of discrete forces, e.g. antielectron vs. proton [58], has a disproportionately large effect on mass.

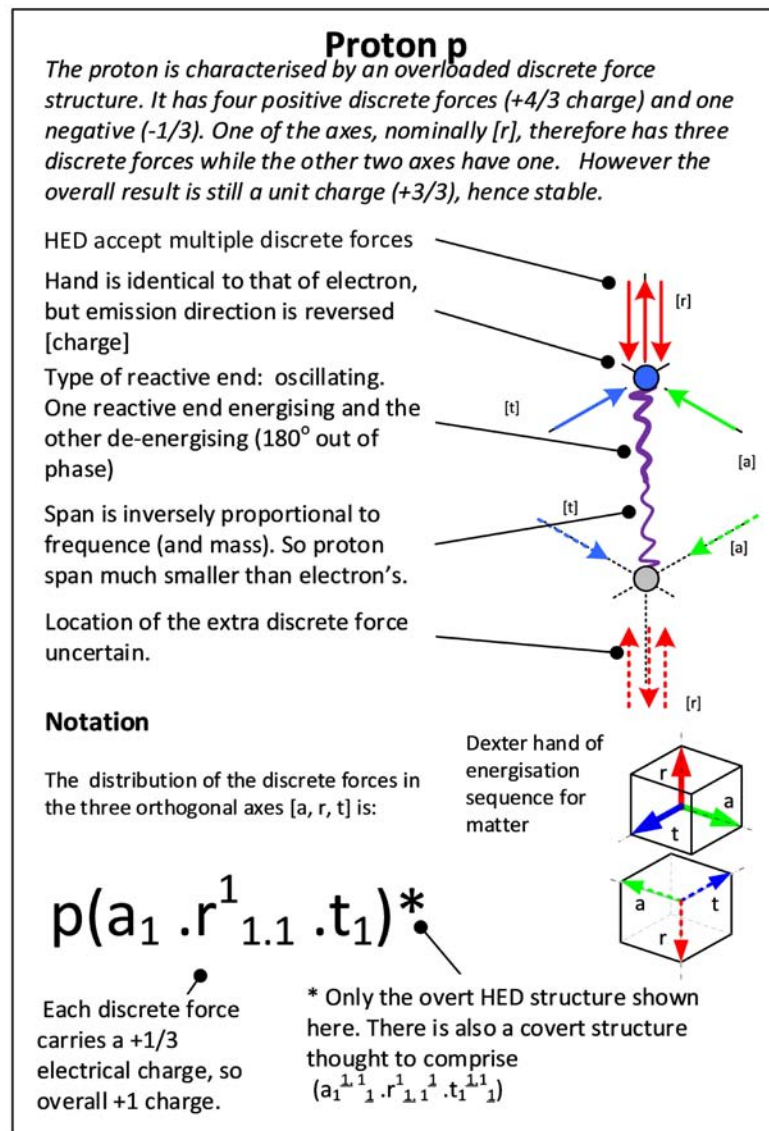
- §3 Furthermore, mass is predicted not to be fixed, but rather dependent on fabric density. Particle B seeks opportunity to emit its own discrete forces into the fabric, and the rate at which it does this is its frequency. The fabric density $\emptyset$ affects its ability to emit. This is elaborated below. Consequently in this theory the mass of an object is an intrinsic property, but not fixed, irrespective of how it is measured.

4.1.3. Identity of Motion

Quantum theory explains force as the emission and receipt of bosons, but as both bosons and fundamental particles are assumed to be 0-D points, it is unclear how these are intercepted. Nor is a QM mechanism apparent for how the receipt of a boson causes force/displacement/velocity of the particle. In contrast the cordus theory does not suffer those limitations, because of its spatial dispersed and non-local structures, and is able to propose explanations for how a



(a)



(b)

Figure 1. Particle sub-structures for the electron and proton. (a) The electron, as a fundamental particle, has a clean architecture of discrete field emissions symbolised by sequential emissions in the $[a, r, t] = [1, 1, 1]$ plane, where $[a, r, t]$ is the Cartesian coordinate system for a reactive end. All these emissions are of the same sign. Image adapted from [108] Creative Commons Attribution 4.0 International license. (b) The proton emissions are more complex as they include both positive and negative charged elements, and covert emission components (not illustrated). Image adapted from [58].

particle detects and moves in a field [4]. The results show that under this theory each of the two reactive ends moves in an inclined orbit around its nominal central location. This has two important consequences. First, it allows the reactive end to sample the space around it and hence intercept & interact with external fields. This addresses the targeting problem of theories that rely on 0-D point particle exchange. It also explains why locality is not strictly preserved at small scales. Second, the reactive end emits forces during its orbit. These may be con-

considered discrete pulses (the earlier interpretation of the cordus theory) or sinusoidally variable fields – both interpretations are correct depending on the perspective. The orbital motion thus corresponds to the handed emission of fields/discrete forces. In turn this handed emissivity underpins the gravitational interaction. Importantly, incoming fields/handed discrete forces retard or advance the reactive end's own emissions, and hence also its orbital motion. The orbital motion and emissions are coupled phenomena. This orbit is circular in shape for a particle at rest, but becomes warped into an asymmetrical spiral when the particle moves or is in the presence of a field, see **Figure 2**. In turn this causes the reactive end (hence the particle as a whole) to move in the direction of the favourable field gradient [4]. As a consequence, motion and velocity of the particle arise [84], which is relevant to the present discussion because this corresponds to the inertial response of the particle.

The reactive end moves to the degree that it is de-energised, and is stationary when fully energised. Consequently, for linear motion, each reactive end moves forward in turn. Hence motion is intermittent at the level of an individual reactive end, but continuous when considering the overall effect of both reactive ends. The presence of external discrete forces, whether from fields or the background fabric, affects this ability to emit. This provides the reactive end with a mechanism to sample the region of space around its nominal location [4], and

Cumulative displaced reactive end position [red] vs undisplaced [green]

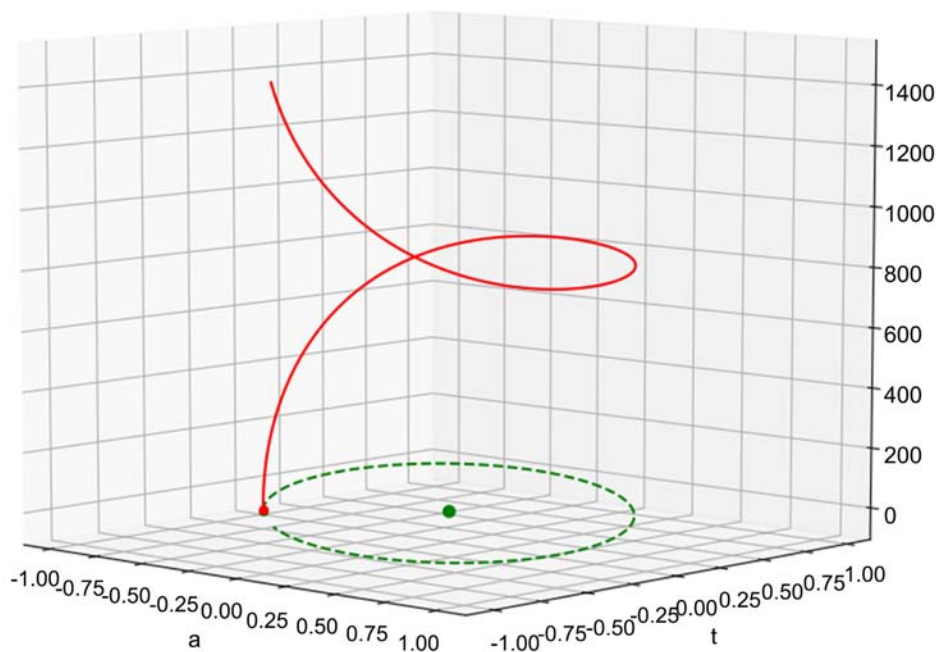


Figure 2. Locus of a single reactive end, undisturbed and under the effect of an electrostatic force in the r direction. Dashed (green) curve shows motion of the undisturbed reactive end. This is a circle in the $[a, r, t] = [1, 1, 1]$ plane, where $[a, r, t]$ is the Cartesian coordinate system for a reactive end. Solid (red) curve shows cumulative locus of the reactive end under the effect of a force in the r direction. This is not a regular spiral. Image reproduced from [4] with permission.

hence both detect and respond to field gradients [85]. This also provides a mechanism for the particle to change its own energisation in response to fabric density, and it is proposed that frequency slows in response to higher fabric density.

4.1.4. Equivalence from a Velocity Perspective

Consider a small test mass B as a satellite in a gravitational interaction with another larger body A. We assume that both bodies are physical objects at our macroscopic level of existence and hence comprise internally decoherent assemblies of matter, as opposed to having quantum coherence between the particles. This is a reasonable assumption to make at the level at which relativity and gravitation occur. We are primarily interested in small body B in the following analysis. We assume B is of negligibly small mass compared to A, *i.e.* we do not consider the reciprocal gravitational effect of B on A.

From the perspective of particle B, its identity is expressed by the emissions of its discrete forces, with the frequency of emissions moderated by the fabric density in which it finds itself (described below). Those emissions are intricately linked to both its tangential motion and its gravitational response. Previous work in the cordus theory has established the theoretical grounds for both. The forward motion has been shown to arise from periodic movement of the reactive ends during their de-energised phase [84]. The motion is not continuous, but inversely related to the strength of energisation, which is sinusoidal. Other work showed that the gravitational response arises from distortion of the locus of the reactive end, which also occurs to the extent that the reactive end is not energised [4] [85]. The gravitational response is therefore one of discrete displacements, the linear motion likewise, and both are affected by frequency. This is an important finding because it implies, from the perspective of self-consistency of the particle's emissions, that both its linear and orbital motion are affected by frequency. This relationship is not anticipated by either GR or QM.

This line of reasoning implies that the centripetal acceleration relation is applicable, which relates radial and tangential motion Hence:

$$a_{B2} = \frac{v_{B2}^2}{r_B} = \frac{F_{rB}}{m_B} \quad (7)$$

where a_{B2} is the radial acceleration of body B moving at velocity v_{B2} at orbital radius r_B , with radial force F_{rB} and mass m_B .

In the orbit case the radial force is the gravitational force $F_{rB} = F_{gB}$ and the tangential force is the inertial.

For clarity, note that in the cordus theory per §2 the discrete force interactions of the ranged forces are fundamentally unidirectional [4]. The fact that remote particle B moves in response to the discrete forces emitted by basal particle A does not change the behaviour of A. This is a logical necessity to preserve temporal causality: the discrete forces received by B were emitted by A some time in the past and have travelled out at the speed of light. If retrocausality is to be denied, as physical realism expects, then it is a logical necessity that the ranged

forces must all be unilateral. It is only when considering discrete forces *en-masse* that the familiar bi-directional force interaction emerges. Under the cordus theory the gravitational interaction experienced by B is a consequence of discrete forces it receives from A. The interaction only changes B, the recipient, not A. The corollary is that B can only change A by sending discrete fields back to A. The above comments apply to decoherent bodies. Coherent assemblies of matter have access to the synchronous interaction [85] [95] and thus to entanglement which can result in superluminal conveyance of information [77], but this is not relevant to the mass equivalence question.

Also, note that per §3 discrete forces are not consumed in the interactions with other particles (in contrast to the colour change of QCD or the virtual bosons of QED), but instead continue to propagate away from their emitting particle. Therefore they act progressively on *every* particle within a macroscopic body as they travel through that body. Hence whether body B comprises one electron, or a whole star, the results are the same, *i.e.* the gravitational interaction generalises for decoherent assemblies of matter.

To address the equivalence question, it is necessary to substitute the gravitational force F_{gB} in the above. However before doing so it is necessary to introduce the fabric variable.

4.2. Fabric Properties

The *fabric* refers in this theory to space between matter, which contains the field emissions of other particles [93]. The nature of these emissions is described in [4] and [84]. The fabric density $\emptyset$ depends on the spatial distribution of matter in the accessible universe. The universe itself expands by spatial divergence, and bodies move about, hence $\emptyset$ is non-isotropic and spatially and temporarily variable across the universe [105]. At any one location in the universe, the fabric takes a local value that may be different to other locations because of the different exposure to the emissions of remote masses.

The fabric is not the same as an aether. The Michelson-Morley experiment [109] disproved the existence of a directional aether wind affecting the speed of light due to the relative movement between the Earth and the medium. However such experiments do not disprove the present proposition, where the speed is isotropic but varies with fabric density [105]. The fabric is not a static medium or fluid such that it has motion of its own to contribute to photons traversing it. This fabric is discrete at the fundamental scale, but approximately smooth at coarser scales. The local fabric density determines the electrical and magnetic constants of the vacuum, which by this theory would not be universally constant [105]. Per this theory the speed of light depends on the fabric density [105]. This may be unconventional but is consistent with the observation that general relativity already accepts that the speed of light is dependent on gravitational field strength, and likewise optics accepts a speed of light that is dependent on refractive index (which is a density dependent property). The emissions from a par-

ticle propagate out at the local speed of light to affect others elsewhere in space. The fabric thus provides the mechanism whereby remote regions are causally connected temporally [96]. Consequently it may be shown that time in this theory is an emergent property of the fabric rather than a dimension [2] [96].

We use the term “*situation*” to refer to a location in space with its local fabric. This is similar to “*frame*” in relativity, but with the inclusion of the fabric parameter. GR does not admit the possibility that inertial frames of reference may not be identically alike. In contrast the cordus theory asserts that equivalence of locations only arises when *both* the inertial conditions and fabric density are the same.

4.2.1. Fabric Properties: Density, Vector, and Gradient

There are multiple attributes to the fabric. It has a scalar magnitude (“density”) value, and also a vector attribute. The magnitude of the fabric density $\emptyset$ at a location x is the summed magnitude of all contributions of masses m_i at range r_i in the observable universe, hence:

$$\emptyset = \sum_x \left| \frac{m_i}{r_i^2} \right| \quad (8)$$

where the computation is situationally-centric to the location, *i.e.* the radial separation r_i is measured from the location under examination. The reason for the r_i^2 dependency is due to spatial dilution of the field. Discrete forces travel radially outwards and are diluted over a front comprising an expanding spherical area [85].

The fabric vector $\vec{\emptyset}$ is the vector sum of the mass contributions:

$$\vec{\emptyset} = \sum_i \frac{m_i}{r_i^2} \quad (9)$$

This vector arises because of the asymmetrical distribution of masses around situation i . The magnitude of the resultant vector is $|\vec{\emptyset}| = \nabla \emptyset$, which is the gravitational field intensity at situation i . This gradient is not a flow of discrete forces – there is no wind or movement of any aether.

The equations for $\vec{\emptyset}$ and $\emptyset$ appear to have a singularity for vanishingly small r . This does not occur as the expressions are only for the far field when the masses m_i are far from each other, e.g. in a galaxy. In the near field the fabric effect becomes one of individual negotiation between particles for emission rights, and this corresponds to alignment of particle orientations (spin parameters). This results in bonding and the formation of crystalline structures. So the m_i/r_i^2 formulation does not hold within such bodies, and the response of the body to external fabric is shared by the crystalline structure. In the extreme near field of coincident particles, the effect becomes one of reactive ends of different particles being co-located (at least at one reactive end each) and synchronising their emissions. The smallest r can be is the size of a single particle A, which in this theory still has a span and occupies volume due to its internal structure [83].

4.2.2. Fabric Effects on Frequency

The Lorentz transformation has been derived using the cordus theory [2], and the results show that frequency and velocity depend also fabric density, which the conventional general relativity does not include. Therefore the fabric density is termed a *covert variable*. The fabric density affects the ability of a particle to complete its own emissions, with greater fabric density retarding the emission [2], hence changing the frequency, velocity and time of the particle. Frequency determines the rate at which the particle can interact with other particles, hence affecting rates of nuclear, chemical & physiological reactions, and this corresponds to time as experienced by the particle [96].

4.2.3. Migration: Motion into Situations of Different Fabric Density

Changes to fabric density, or movement of the particle into situations of different fabric density, affect the rate of time of the particle, hence also time dilation [2]. Consider test particle B with velocity v_{B1} and frequency f_{B1} in situation 1 where the fabric density is ϕ_1 . Subsequently B moves to a new situation of ϕ_2 . Per [2] its intrinsic velocity changes to v_{B2} and frequency to f_{B2} as viewed by Observer A remaining in ϕ_1 .

Observation from originating station

In what follows, $\rightarrow A1$ is used in ambiguous cases to indicate the situation of observation:

$$f_{B2 \rightarrow A1 \phi_2} = f_{B1 \phi_1} \quad (10)$$

and

$$v_{B2 \rightarrow A1 \phi_2} = v_{B1 \phi_1} \quad (11)$$

Hence also

$$v_{B2 \rightarrow A1} = v_{B1} \frac{\phi_1}{\phi_2}$$

Thus if $\phi_2 < \phi_1$ then the velocity of B increases. Examples of where these intrinsic changes would occur are where a star in a galaxy moves distally, or out the plane of the galaxy, or outside of the bar/spiral arm: all these changes are towards reduced fabric density.

From the perspective of an observer A positioned in situation 1 with ϕ_1 the changes to B are detectable. That observer would see a change in the frequency of emission of radiation from B, and the velocity of B against a backdrop of markers of known position.

Co-moving Observation

The changes are not detectable by a co-moving observer since the particles making up the observer will also change in frequency and velocity. However an observer co-moving with B would perceive the rest of the universe as having changed by the inverse relationships.

These changes relate to difference in fabric density, and are separate to relativistic considerations. We refer to these as *intrinsic* changes as they occur via

conservation mechanisms internal to the particle [4] [84], in contrast to *extrinsic* changes such as contact forces or electro-magneto-gravitational fields. Unconventionally, the theory predicts the changes do not require extrinsic energy supply. They are the response of the energisation mechanism of the particle to the changed external constraints imposed by the fabric. The rate of time in this theory is not a dimension, nor is it universally constant, rather it is a property of the fabric density, hence of the distribution of mass.

Observation from a third situation

Relative to observer E in a different situation with ϕ_3 , possibly in a different galaxy at a later epoch of time, the relative local passage of time, measured by frequency f_3 is affected by the fabric density per Equation (10), hence all else being equal $f_3\phi_3 = f_2\phi_2$. If $\phi_3 < \phi_2$ then time passes faster for E, hence E perceives B to have slower frequency $f_{B2 \rightarrow E3}$ and velocity $v_{B2 \rightarrow E3}$. Thus:

$$v_{B2 \rightarrow E3} = v_{B2} \frac{\phi_3}{\phi_2} \quad (12)$$

And if B migrates from ϕ_1 to ϕ_2 then

$$v_{B2 \rightarrow E3} = v_{B2 \rightarrow A1} \frac{\phi_3}{\phi_1} \frac{\phi_3}{\phi_2} = v_{B1} \frac{\phi_1}{\phi_2} \frac{\phi_3}{\phi_1} \frac{\phi_3}{\phi_2} = v_{B1} \left(\frac{\phi_3}{\phi_2} \right)^2 \quad (13)$$

In the case where E is an observer on Earth (within the Milky Way), looking back at a denser prior epoch of the universe, and B is a star in the outer reaches of a distant galaxy, then Equation (12) implies observed galaxy rotation curves would not be reliable indicators of actual velocities. The correction is complex because ϕ_2 depends on orbital radius, size and morphology of the galaxy. Equation (12) applies for stars formed in place, whereas Equation (13) is for migratory stars.

4.2.4. Intrinsic Changes in Velocity

The above changes in velocity occur due to the particle experiencing changed fabric density. This occurs without the application of a direct force, at least not an extrinsic force acting on the particle at the time. It is unhelp to consider this an acceleration in the conventional sense. Rather the body natural speeds up or slows down depending on fabric density encountered. Elsewhere in the universe some force must previously have been applied to rearrange matter to cause the fabric density to change, but there need be no extrinsic force on B itself. Hence there is an intrinsic change in velocity Δv given by:

$$\Delta v = v_{B2} - v_{B1} = v_{B1} \left(\frac{\phi_1}{\phi_2} - 1 \right) \quad (14)$$

If a sudden boundary in ϕ were encountered, then the change in velocity would be immediate, *i.e.* the acceleration would be infinite. However this is not a realistic case because there is always some minimum time involved, which is the energisation period of the particle. The time t_ϕ taken for transition from situation 1 to 2 is expected to be large when dealing with cosmological scales and

weak gradients in fabric density, and $\frac{\phi_1}{\phi_2}$ will be close to unity at the scale of a solar system, so the acceleration $a_\varnothing = \Delta v/t_\phi$ is expected to be miniscule. As the acceleration does not require a direct extrinsic force, it is necessary to modify Newton's equation of motion to:

$$a_{B1} = \frac{F_{iB}}{m_{B1}} + a_\varnothing = \frac{F_{iB}}{m_{B1}} + \frac{v_{B1}}{t_\phi} \left(\frac{\phi_1}{\phi_2} - 1 \right) \quad (15)$$

where a_{B1} is acceleration as observed from situation 1, and F_{iB} is the inertial force applied. The first component is the conventional extrinsic part, and the second is the intrinsic part due to the fabric. The latter goes to zero when there is no difference in fabric density, hence recovering Newton's inertial equation. This completes the derivation of how the fabric density parameter affects frequency (hence time) and velocity.

4.3. Inertial Mass

Having established the fabric density effect, it is now necessary to consider the implications for inertial mass.

Intrinsic Changes in Inertial Mass

We assume mass is proportional to the energisation frequency f_B of the particle. (It may instead be a squared or other function—this is not settled within the cordus theory). When the frequency of B increases, it emits discrete forces at a higher rate. Hence there are more discrete gravitational interactions with another body, and the gravitational force increases in strength.

When particle B of mass m_{B1} moves from a situation with ϕ_1 into a new situation of ϕ_2 its frequency changes, and hence the mass changes to m_{B2} as viewed by an Observer A remaining in ϕ_1 :

$$m_{B2 \rightarrow A1} = m_{B1} \frac{\phi_1}{\phi_2} \quad (16)$$

This equation describes a case where a mass moves from ϕ_1 to ϕ_2 . This might be a satellite that leaves the Solar system, or a star that migrates through a galaxy. If $\phi_2 < \phi_1$, *i.e.* when the mass moves into a region of lower fabric density, then the inertial mass increases. This is an intrinsic increase, since mass is intricately linked to energisation frequency in this theory. Hence this theory predicts that mass is not strictly constant, but depends on fabric density.

The changes in mass (Equation (17)) and velocity (Equation (11)) are in the same direction (both increase or decrease). Hence the cordus theory predicts that the conservation of momentum equation must also be modified to include the fabric density term:

$$m_{B1} v_{B1} \phi_1^2 = m_{B2} v_{B2} \phi_2^2 \quad (17)$$

Another way of looking at this is to state that the particle changes to a more energised state, hence also becomes heavier, when the fabric density decreases.

Having examined the effect of fabric density on inertial mass, we next address gravitational mass. As identified above, for reasons of self-consistency of discrete force identity, it is assumed that the inertial and gravitational masses are the same. However this does not necessarily mean that their relationship with fabric density is trivial.

4.4. Gravitation

The operation of the gravitational field has previously been developed in the cordus theory [85], and relevant parts are briefly introduced here. Consider a central mass A comprising one or more particles, with a satellite mass B. We are only concerned here with the gravitational effect on B. In this theory the gravitational field of a particle arises from torsion in its emitted flux tube [85]. This torsion arises from the sequence (and hence handedness) of emission of its three discrete forces. For the particle B receiving and responding to a gravitational field, the interaction arises as a constraint on displacement during the energisation cycle of the particle [4] [85]. The particle needs to emit its own discrete forces, detects the gradient in the fabric, and attempts to move in a direction that maximises its opportunity to emit its own discrete forces [4]. Hence the reactive ends of the unconstrained particle move along the field gradient to maximise the mutual compatibility. Compatibility here relates to the handedness of emissions sequence. In a three dimensional spatial system with three discrete forces, there are only two ways this energisation may be accomplished, hence two hands, dexter and sinister, and these are attributed to the matter-antimatter species respectively [97]. The interaction is not continuous but rather follows a sinusoidal function over time, with the reactive end undertaking a torsional displacement during its de-energised phase [84] (see **Figure 2** above). Typically the remote particle B is of the same species as body A, *i.e.* both are matter. Then B finds it easier to make its own emissions if they are synchronised with those it receives from A. This increases the compatibility, and hence B moves up the field gradient, *i.e.* matter-matter gravitation is attractive. Matter-antimatter is predicted to be gravitationally repulsive, and antimatter-antimatter attractive [85]. This concept of gravitation being underpinned by a torsional mechanism is consistent with other developments in the field, such as Einstein-Sciama-Kibble torsion gravity developments [110], but originates from a different line of reasoning.

4.4.1. Gravitational Dependency on Fabric

The strength of the gravitational interaction depends proportionally on the mass of B and the gravitational field intensity, hence the effect is proposed to be multiplicative:

$$F_{BG} \approx \left| \overline{\phi}_2 \right| m_B \quad (18)$$

The gravitational mass of B (m_B) is the product of quantity of discrete forces it emits determined by particle identity [58], total number of particles, and frequency (which is moderated by the fabric density ϕ_2).

For example, a satellite body B in a situation of lower fabric density ϕ_2 has faster re-energisation of its reactive ends than A, *i.e.* time passes faster [96], compared to a reference situation with ϕ_1 . Its frequency is per Equation (10)

$$f_{B2 \rightarrow A1} = f_{B1} \frac{\phi_1}{\phi_2}. \text{ Consequently B has greater receptivity to gravitational interaction,}$$

hence it experiences a stronger gravitational force than does A from B, by a receptivity enhancement of the ratio of fabric densities.

Note that this frequency effect applies irrespective of whether or not B originated in the same location as A, *i.e.* this is not a migration effect per se, but rather a difference in the fabric densities. This can also be viewed as a time dilation effect [96]—the perspectives are complementary. Time dilation exists in the presence of a gravitational field, and hence by inference the two bodies cannot experience the same temporal summation of force. This asymmetry is unconventional. This finding is not accessible to conventional relativity which assumes forces are continuous parameters, nor to conventional quantum mechanics which assumes forces arise from bilateral exchange of infinitely compact bosons. The cordus theory proposes that forces have a sinusoidal nature, and it is this characteristic that makes the difference.

4.4.2. Gravitation Formalism

Per Equation (9) the gravitational field vector experienced by B at location 2 is

$$\overline{\phi}_2 = \sum_B \frac{m_i}{r_i^2} \text{ where the computation is B-centric. Hence combining the above}$$

Eqns, the gravitational force experienced by B at location 2 is:

$$F_{BG \rightarrow 2} = |\overline{\phi}_2| m_B \frac{\phi_1}{\phi_2} k_G = \left| \sum_B \frac{m_i}{r_i^2} \right| m_B \frac{\phi_1}{\phi_2} k_G \quad (19)$$

where k_G is a factor to account for the conversion of the torsional effect of the discrete forces into gravitational force. Hence the conventional gravitational constant G comprises:

$$G = \frac{\phi_1}{\phi_2} k_G \quad (19a)$$

This factor is not explored further here.

Another way to look at this is that gravitation is an emergent property of the fabric, rather than an invariant attribute of space-time. This idea of gravitation being emergent is somewhat similar to the position taken by entropic gravity, though that approaches it from the different direction of assuming dark matter arises from dark energy effects at the particle level.

4.5. The Inference of Equivalence

This theory expects that a body will behave in internally consistent ways in both inertial and gravitational cases, because the underlying mechanisms of mass are related to a common cause in the emission of discrete forces. Superficially this

means that the cordus theory supports the equivalence of inertial and gravitational mass. Now, having established the fabric dependencies, the question of what the equivalence looks like from a formulaic perspective may be addressed.

Assume a circular gravitational orbit. Assume that the centripetal acceleration a_{B2} applies in the radial direction per Equation (7) ($a_{B2} = \frac{v_{B2}^2}{r_B} = \frac{F_{rB}}{m_B}$). For radial force F_{rB} substitute F_{BG} from Equation (19) as the gravitational force is radially directed for a circular orbit. Hence:

$$\frac{v_B^2}{r_B} = \frac{F_{BG}}{m_B} = \left| \sum_B \frac{m_i}{r_i^2} \right| \frac{\phi_1}{\phi_2} k_G \quad (20)$$

For a gravitational field dominated by a single body A (*i.e.* not a multi-body problem), the gravitational field $|\vec{\varnothing}| = \nabla \varnothing$ which is determined simply by m_A and the orbital radius:

$$\left| \sum_B \frac{m_i}{r_i^2} \right| = \frac{m_A}{r_B^2} \quad (21)$$

Hence:

$$v_B^2 = \frac{m_A}{r_B} \frac{\phi_1}{\phi_2} k_G$$

Thus

$$v_B = \sqrt{\frac{m_A}{r_B} \frac{\phi_1}{\phi_2} k_G} \propto \frac{1}{\sqrt{r_B}} \quad (22)$$

Thus the ratio of fabric densities appears within the equivalence. If the fabric densities are assumed to be the same, which is what Newtonian gravitation assumes, then the conventional equivalence is recovered.

5. Discussion

5.1. Interpretation

5.1.1. Disjointed Equivalence

The surprise in the equivalence formulation of Equation (22) is the covert term $\frac{\phi_1}{\phi_2}$. Our interpretation is that the equivalence holds in any one case under

examination, in terms of the relationship between m_A, r_B, v_B , but a different equivalence holds when the fabric densities change. We refer to this as a *dis-jointed* equivalence. In terms of Newtonian gravitation this corresponds to a variable gravitational constant. This implies the need to abandon the concept of a gravitational constant that is constant in time and place. This is unconventional, but is consistent with the observation that G has been difficult to determine with the precision expected of fundamental constants.

5.1.2. Composition of the ϕ Term

In general the $\varnothing$ terms may be simplified into several summed components:

1) The fabric density of the background universe ϕ_0 at that location in space and temporal epoch. Assuming an expanding universe, the fabric density was greater at early epochs. The universe term may also include the super-structures of galaxies, and the galaxy in which the star is located. The latter is a complex relationship of galaxy morphology, galaxy mass, and orbital radius of the star. The morphology is especially complex with disk galaxies with their cores, disks and arms.

2) The contribution of massy body A to the fabric density, ϕ_A . Depending on the context, body A may refer to a satellite orbiting a star (which is a relatively simple case of m_A/r^2), or a multibody galactic core. At the centre of solid body A (a star) there is a self-contribution to the fabric determined as:

$$\phi_A = \int_0^{R_A} \frac{4\pi r^2 \rho dr}{r^2} = 4\pi \rho R_A = \frac{3m_A}{R_A^2} \quad (23)$$

where this assumes a homogenous spherical body of outer radius R_A and mass m_A and density ρ . This shows that the fabric density is primarily a mass density property, and no singularity arises at the centre of a massy body (or cluster of masses). Separately A also contributes to the gradient via the $\frac{m_i}{r_i^2}$ term.

3) In addition body B makes a fabric contribution which we ignore as small. However in some cases, such as binaries, this will not be a safe assumption.

Hence the fabric density at location A is:

$$\phi_1 = \phi_0 + \frac{3m_A}{R_A^2} \quad (24a)$$

And for location B:

$$\phi_2 = \phi_0 + m_A/r_B^2 \quad (24b)$$

Note that R_A is the outer radius of the mass surface of A, whereas r_B is the orbital radius of body B.

5.2. Gravitation Special Cases

There are several cases with specific assumptions relating to the epoch of the universe. Fabric density varies with the temporal epoch of the universe, being denser in the past. Hence more distant galaxies are predicted in general to have greater background fabric density. Fabric density also varies with galaxy size, with larger galaxies having greater fabric density. In addition, fabric density varies across any one galaxy, being denser and even uniform in the centre (because of many-many mass interactions), and lighter at the periphery. Furthermore this is complicated by the shape of the galaxy, being different in- and out- of the galactic plane. Hence the theory implies that gravitation over cosmological scales is a much more complex interaction than experienced in Earth environs. Several specific cases are enumerated below.

1) Earth location and epoch: Newtonian gravitation

As the universe evolves over time t , its mass becomes spatially distributed, and a

fabric density $\phi_0(t)$ develops. For observations in the vicinity of Earth, cocooned among other stars and galaxies, ϕ_0 is assumed to be large, and approximately homogeneous and static. Hence $\phi_0 \gg \frac{3m_A}{R_A^2}$ where A refers to the Sun, thus per Equation (24), $\phi_1 \equiv \phi_0$ and likewise $\phi_2 \equiv \phi_0$ which means the receptivity enhancement tends to unity, and the gravitational field intensity is dominated by body A (the Sun), giving for Equation (19):

$$F_{BG \rightarrow 2} = \frac{m_A}{r_B^2} m_B k_G \quad (25)$$

which recovers the form of Newtonian gravitation. Hence we propose that the gravitational constant is identified with $G = k_G$ but only in this special case. More generally $G = \phi_2 / \phi_2 k_G$ hence is not constant but rather has a covert fabric dependency. The equivalence for this special case becomes, per Equation (20):

$$\frac{v_B^2}{r_B} = \frac{m_A}{r_B^2} k_G \quad (26a)$$

Hence:

$$v_B \propto \sqrt{\frac{1}{r_B}} \quad (26b)$$

This recovers the conventional Newtonian formulation, where orbital velocity decreases with separation.

2) Galaxy rotation curves

Fabric density will be reduced for a star in a more distal part of a galaxy, or out the plane of the galaxy, or outside of the bar/spiral arm. We assume an approximately spherical galaxy shape, with star B at radius r_B from the galactic centre, though we acknowledge that disk galaxies have a more complex distribution of mass. At the centre of galaxy A, $\phi_1 = \phi_0 + \frac{3m_A}{R_A^2}$, where R_A is approximated as the outer edge of the galaxy, and m_A is the mass of the galaxy. Assume the fabric density of the background universe is some fraction $\frac{1}{a}$ of the galaxy,

$$\phi_0 = \frac{1}{a} \frac{3m_A}{R_A^2} \text{ hence}$$

$$\phi_1 = \frac{3m_A}{R_A^2} \left(\frac{1}{a} + 1 \right) \quad (27a)$$

$$\phi_2 = \phi_0 + \frac{m_A}{r_B^2} = m_A \left(\frac{3}{aR_A^2} + \frac{1}{r_B^2} \right) \quad (27b)$$

Thus:

$$\frac{\phi_1}{\phi_2} = \frac{3 \left(\frac{1}{a} + 1 \right)}{\frac{3}{a} + \frac{R_A^2}{r_B^2}} \cong a + 1 \quad (27c)$$

For example, with $a = 1$, *i.e.* the fabric density of the galaxy at its core is the same strength as the universe at large, $\frac{\phi_1}{\phi_2} = 2$.

Thus the gravitational formulation in Equation (19) becomes:

$$F_{BG \rightarrow 2} = \frac{m_A}{r_B^2} m_B (a + 1) k_G \quad (28)$$

The equivalence for this special case becomes, per Equation (20):

$$\frac{v_B^2}{r_B} = \frac{F_{BG \rightarrow 2}}{m_B} = \frac{m_A}{r_B^2} (a + 1) k_G \quad (29a)$$

Hence:

$$v_B \propto \sqrt{\frac{a + 1}{r_B}} \quad (29b)$$

This predicts that more peripheral stars will orbit faster than the Newtonian prediction, because they experience a stronger gravitation. This is consistent with known observations, though the consistency is conceptual rather than quantified at this stage, because we cannot yet suggest a way for parameter a to be determined. The orthodox cosmological interpretation attributes dark matter as the cause of anomalous galaxy rotation curves, though no evidence of such matter has yet been found. The MOND family of theories [69, 70] are successful at modelling the galaxy rotation curves, and hence obviate the need for dark matter, by assuming that gravitation get stronger with distance. While MOND is quantitatively accurate, it has no underlying ontological explanation of why gravitation should increase in strength with distance. The cordus theory offers such an explanation, and further proposes that the effect is not solely distance, but also fabric density. Increased gravitational force is also provisioned in the non-geodesic theories, *i.e.* an additional force is involved [111], involving both vector and scalar fields [112]. The cordus theory also involves vector and scalar fields in gravitation. Hence there is a possibility that the cordus theory might provide an underlying rationale based on physical realism for the MOND and vector-scalar field theories. There is a conceptual consistency, though additional work would be necessary to check this.

Galaxy rotation velocities are determined from Doppler shift in spectral lines. Our interpretation is that these frequencies would need to be corrected for differences in fabric density those velocities can be determined. Given the difficulty of determining fabric density for remote locations, it is currently unclear how this may be done.

3) Genesis epoch for the universe

Expansion of the universe

Evidence of the accelerated expansion of the universe has been accumulated from multiple studies, e.g. based on redshifts [113], or supernovae [114], considering also anisotropies [115], and these results generally support the idea of a cosmological constant [116]. This constant may represent a vacuum energy den-

sity [117], which is commonly believed to be created by a dark energy. The conventional interpretation of the expansion of the universe is that the metric changes scale, *i.e.* that the spatial dimensions themselves expand. This is premised on there being nothing for space to expand into, *i.e.* nothing outside the universe. There are numerous models for the expansion, and theories for its mechanisms. The dominant conventional model is Lambda cold dark matter (Λ CDM), which assumes the existence of dark energy to drive the cosmological expansion, and of cold dark matter to explain the galaxy rotation curves. The model shows good quantitative agreement with empirical measurements [118]. However the identities of neither dark energy nor dark matter are known. Alternative theories for the expansion include modifications to general relativity, modifications to gravitation such as MOND [119] and similar such as entropic gravity, and biometric gravity.

In contrast the cordus theory supports a different interpretation. For a start, time is not a dimension in the cordus theory, but an emergent property of matter, and is communicated through the fabric [96]. The theory also proposes the primacy of the spatial dimensions, so that there is something for the universe to expand into. Consequently there is also a frontier of expansion, which is the cosmological boundary [93]. At genesis the baryogenesis is proposed to have occurred via pair production [104] and the remanufacture of the anti-electron into the proton [58], resulting also in domain warfare between competing matter and antimatter pathways. The matter and antimatter bodies are expected to have comprised coherent neutron-species, and the interaction between them is predicted to be repulsive by the synchronous/strong force [95] for coherent conditions, and also repulsive by gravitational for the decoherent state [85]. Hence this provides a mechanism for a rapid explosion (“inflation”) of the primal massy material, followed by a momentum-driven expansion of the universe. Consequently the cordus theory disfavours the idea of metric expansion of space, and instead proposes it to be a more conventional expansion by movement of matter. Hence when the universe expands, other masses move outwards, their separations increase, and the overall fabric density drops, *i.e.* a change in the r_i

components within $\varnothing = \sum_x \left| \frac{m_i}{r_i^2} \right|$ (Equation (8)). This causes an increase in

intrinsic velocity per Equations (14)-(15). Thus the cordus theory does not conceptually need dark energy, but instead proposes that the expansion of the universe is an intrinsic change in velocity caused by the reduction in fabric density that arises from the expansion itself.

The inertial-gravitational relationship at the genesis epoch is predicted as follows. Assume the universe comprised a single mass A located at 1, and a small remote test body B located at 2. Also assume body A being in a decoherent state at least in its far-field emissions. In this extreme case, with no other mass in the universe the fabric density is entirely created by A, hence $\varnothing_1 = \frac{3m_A}{R_A^2}$, $\varnothing_2 = \frac{m_A}{r_B^2}$,

and with $\left| \sum_B \frac{m_i}{r_i^2} \right| = \frac{m_A}{r_B^2}$ substituting into Equation (19):

$$F_{BG \rightarrow 2} = \left| \sum_B \frac{m_i}{r_i^2} \right| m_B \frac{\phi_1}{\phi_2} k_G = m_B \frac{3m_A}{R_A^2} k_G \quad (30)$$

Note there is no surviving r_B term representing the orbital radius of body B. Hence the theory makes the unusual prediction of a genesis gravitational force that is initially constant with separation. The equivalence for this special case becomes, per Equation (20):

$$\frac{v_B^2}{r_B} = \frac{3m_A}{R_A^2} k_G \quad (31a)$$

Hence

$$v_B = \sqrt{r_B \frac{3m_A}{R_A^2} k_G} \propto \sqrt{r_B} \quad (31b)$$

Hence orbital velocity would increase slowly as separation increased. In contrast for the Newtonian case $v_B \propto \frac{1}{\sqrt{r_B}}$.

The current prediction of stronger gravitation at earlier epochs of the universe is consistent with other findings. First, there is empirical evidence for the accelerating expansion of the universe. This is conventionally attributed to dark energy or a variety of other causes [120]. The dark energy formulation of Chevallier-Polarski-Linder has a linear dependency on redshift z of the form $z/(1+z)$ which predicts an increase in dark matter for greater z . In the present model, greater redshift corresponds to earlier phases when the universe was denser and hence gravitation stronger. Second, the dark matter findings may also be interpreted as consistent with the current model. Per the Navarro, Frenk & White (NFW) model [121], the density of a Lambda cold dark matter (Λ CDM) halo depends on the mean density of the universe at its location, *i.e.* later formed haloes have lower density [122]. Furthermore, the Einasto power law has coefficient α which is not constant but rather increases with redshift and mass [122]. An increase in dark matter corresponds to a stronger gravitational force.

Future evolution of the universe

As the matter of the universe expands further apart in the far future, there is expected to be a reduction in the fabric density of the universe as a whole, so ϕ_0 tends to smallness. For body B orbiting A, the fabric density will be dominated by A, hence $\phi_1 = \frac{3m_A}{R_A^2}$ and $\phi_2 = \frac{m_A}{r_B^2}$ hence $\frac{v_B^2}{r_B} = \frac{3m_A}{R_A^2} k_G$ as per the genesis epoch. Thus the gravitational force is predicted to be stronger at genesis, weaker during middle epochs, and strong again at the end of the universe. We are unsure whether this implies an open or closed universe.

5.3. Critique

With inductive reasoning the conclusions are not logically certain. This is because there remains an element of doubt about the starting principles, and not all the empirical evidence (in this case from cosmology and particle physics) has been considered. Nonetheless the inductive approach has the benefit of amplifying a topic into a broader set of principles, *i.e.* the theory is further developed and new principles discovered.

In the case of the present paper those new principles are the prediction of a disjointed equivalence conditional on fabric densities, a variable gravitational constant G , and a gravitational formulation that changes with fabric density and the evolution of the universe. These findings are consistent with empirical observations of galaxy rotation curves.

Furthermore the present findings are logically consistent with the other published work on the cordus theory. This is a useful quality test for conceptual development of a candidate new theory of physics like this.

5.3.1. Falsifiable Predictions

Falsifiable feature of the theory are:

- 1) A unidirectional causality of the gravitational forces exists at the fundamentally level.
- 2) In the current epoch and location, the fabric density contributed by the observable universe is comparatively larger than that contributed by local gravitational effects (e.g. Sun, Earth).
- 3) A body will have faster velocity than predicted by gravitation alone, when it moves into situations of lower fabric density.

5.3.2. Limitations and Future Research

The cordus theory developed here is a conceptual and logical work, built on a starting conjecture for the structure of matter. The current paper has provided a theoretical formalism for mass equivalence, but has not tested this against empirical results. Left for future work is the task of analysing galaxy rotation curves to determine how well this theory fits those observations. This would seem to be the most direct route to test the theory, since the fabric density—which is otherwise invisible—can perhaps be computed from mass and morphology of galaxies. We do not underestimate the potential difficulty in determining absolute values of fabric densities for various astronomical objects. If this cordus theory is true, determining these densities would be a necessity to understand the evolution of the universe. This would seem to require new mathematical methods and empirical approaches, hence a large potential future area of cosmology research.

6. Conclusions

The principles of the cordus non-local hidden-variable theory have been extra-

polated by inductive reasoning to examine the identity of mass, and explore the inertial-gravitational equivalence. Key findings of this theory are as follows. A disjointed equivalence is predicted, whereby inertial and gravitational masses are equivalent in any one situation, but a different equivalence holds when the fabric densities change. This is interesting, because it means that the equivalence holds in any one situation, but the nature of the equivalence changes with location. Consequently this theory predicts that the gravitational constant G varies with fabric density, and hence would be different across the universe and across time. Other theories also propose a variation in gravitation, such as the Brans-Dicke theory [123] where the gravitational constant is replaced with a variable scalar field (also called ϕ), though the approach and formulation differ from the present theory. Similarly MOND [69] [70] proposes that the gravitational parameters are non-constant.

In this cordus theory, not only is the gravitational constant non-constant, but the formulation of gravitation changes with fabric density. Specifically, the theory predicts gravity is stronger at genesis (and the end of the universe) such that orbit velocity $v_B \propto \sqrt{r_B}$ (where r_B is orbit radius), compared to weaker gravitation at middle life epochs such that $v_B \propto \frac{1}{\sqrt{r_B}}$. The current Earth location and epoch

correspond to the latter case, *i.e.* Newtonian gravitation is recovered by a special-case simplification of the cordus theory. This is an original finding.

Furthermore, a novel explanation is provided for the accelerating expansion of the universe. The theory proposes that the expansion of the universe is an intrinsic change in velocity caused by the reduction in fabric density that arises from the expansion itself.

If true, this has implications for understanding the evolution & expansion of the universe, and the interpretation of galaxy rotation curves. The findings disfavour the existence of both dark energy and dark matter, and instead attribute these effects to differences in the fabric density.

In summary the original contribution of this work is the elucidation of fabric density as a covert variable in the mass equivalence formulations, the identification that mass equivalence is disjointed—inertial and gravitational masses are equivalent in any one situation, but a different equivalence holds when the fabric densities change, and the identification that the gravitation formulation itself changes with relative fabric densities.

Conflicts of Interest

The author declares that there are no financial conflicts of interest regarding this work.

Acknowledgements

Thanks for hospitality to concentrate on writing this paper: Jas, Jodi, Dave, Stef, Hends, Owen, Cath, Wa, Jono, Helen, Andrew, Liesl, Di.

References

- [1] Pons, D.J., *et al.* (2012) *Physics Essays*, **25**, 132-140.
<https://doi.org/10.4006/0836-1398-25.1.132>
- [2] Pons, D.J., Pons, A.D. and Pons, A.J. (2018) *Journal of Modern Physics*, **9**, 500-523.
<https://doi.org/10.4236/jmp.2018.93035>
- [3] Pons, D.J., Pons, A.D., Pons, A.M. and Pons, A. (2011) Cordus in Extremis: Part 4.3 Gravitation, Mass and Time. viXra:1104.0029, 1-14. <http://vixra.org/abs/1104.0029>
- [4] Pons, D.J. (2020) *Journal of Modern Physics*, **11**, 1560-1575.
<https://doi.org/10.4236/jmp.2020.1110097>
- [5] Davoudiasl, H. and Huber, P. (2005) *Physical Review Letters*, **95**, Article ID: 191302.
<https://doi.org/10.1103/PhysRevLett.95.191302>
- [6] Hirano, K. (2014) Constraints on Neutrino Masses from Future Cosmological Observations. *AIP Conference Proceedings*, **1594**, 35-40.
<https://doi.org/10.1063/1.4874041>
- [7] Murayama, H. (2006) *Progress in Particle and Nuclear Physics*, **57**, 3-21.
<https://doi.org/10.1016/j.pnpnp.2006.02.001>
- [8] Wang, K. (2008) Testing Origin of Neutrino Masses at the CERN Large Hadron Collider. PhD Thesis, University of Wisconsin, Madison, WI.
<https://ui.adsabs.harvard.edu/abs/2008PhDT.....110W/abstract>
- [9] Akhmedov, E.K. and Frigerio, M. (2006) *Physical Review Letters*, **96**, Article ID: 061802. <https://doi.org/10.1103/PhysRevLett.96.061802>
- [10] Froggatt, C.D. (2003) *Surveys in High Energy Physics*, **18**, 77-99.
<https://doi.org/10.1080/01422410310001610419>
- [11] Kwong, K.-Y. (2001) Quantum Induction and Higgs Mass. PhD Thesis, The University of Arizona, Tucson, AZ. <https://repository.arizona.edu/handle/10150/289767>
- [12] Popovic, M.B. (2002) Dynamical Generation of the Top Quark Mass. PhD Thesis, Boston University, Boston, MA.
<https://ui.adsabs.harvard.edu/abs/2002PhDT.....27P/abstract>
- [13] Vaucher, A.R. (2001) Quark Mass Hierarchy and Dynamical Symmetry Breaking. PhD Thesis, University of California, Riverside, CA.
<https://ui.adsabs.harvard.edu/abs/2001PhDT.....28V/abstract>
- [14] Higgs, P.W. (1964) *Physical Review Letters*, **13**, 508-509.
<https://doi.org/10.1103/PhysRevLett.13.508>
- [15] Englert, F. and Brout, R. (1964) *Physical Review Letters*, **13**, 321-323.
<https://doi.org/10.1103/PhysRevLett.13.321>
- [16] Anderson, P.W. (1963) *Physical Review*, **130**, 439-442.
<https://doi.org/10.1103/PhysRev.130.439>
- [17] Djouadi, A. (2009) The Higgs Mechanism and the Origin of Mass. In: Blanchet, L., Spallicci, A. and Whiting, B., Eds., *Mass and Motion in General Relativity*, Springer, Dordrecht, 1-23. https://doi.org/10.1007/978-90-481-3015-3_1
- [18] Berbeco, R.I. (2002) In Search of the Standard Model Higgs Boson. PhD Thesis, University of Michigan, Ann Arbor, MI.
<https://deepblue.lib.umich.edu/handle/2027.42/129133>
- [19] Gabriel, S. (2010) New Ideas in Higgs Physics. PhD Thesis, Oklahoma State University, Stillwater, OK. <https://shareok.org/handle/11244/6888>
- [20] Murdock, Z. (2011) New Ideas for Yukawa Interactions and the Origin of Mass. PhD Thesis, Oklahoma State University, Stillwater, OK.

- <https://shareok.org/handle/11244/6906>
- [21] Iordanidou, K. (2014) *Acta Physica Polonica B, Proceedings Supplement*, **7**, 639-644.
<https://doi.org/10.5506/APhysPolBSupp.7.639>
 - [22] Gianotti, F. and Virdee, T.S. (2015) *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **373**, 20140384.
<https://doi.org/10.1098/rsta.2014.0384>
 - [23] Palmer, C.A. (2014) Discovery of the Higgs Boson Decaying to Two Photons. PhD Thesis, University of California, San Diego, CA. <https://cds.cern.ch/record/1951331>
 - [24] Cerio, B.C. (2015) Evidence of Higgs Boson Production through Vector Boson Fusion. PhD Thesis, Duke University, Durham, NC.
<https://dukespace.lib.duke.edu/dspace/handle/10161/9820>
 - [25] Field, B.J. (2005) Higgs Phenomenology in the Standard Model and Beyond. PhD Thesis, State University of New York at Stony Brook, Stony Brook, NY.
<https://cds.cern.ch/record/1044919/>
 - [26] Conley, J.A. (2009) Expecting the Unexpected: Signals for New Physics. PhD Thesis, Stanford University, Stanford, CA.
<https://www.proquest.com/openview/672bb381eee0f5852f75745a36c0724e/1?pq-origsite=gscholar&cbl=18750>
 - [27] He, B. (2015) Higgs Phenomenology and New Physics beyond the Standard Model. PhD Thesis, University of Delaware, Newark, DE.
<https://udspace.udel.edu/items/f4734407-37f1-4308-9728-f1b7f1468218>
 - [28] Jiang, Y. (2015) Higgs Boson Physics beyond the Standard Model. MSc Thesis, University of California, Davis.
<https://www.proquest.com/openview/cff8689942094aae381f637f83567612/1?pq-origsite=gscholar&cbl=18750>
 - [29] Lee, G. (2014) Applications of Effective Field Theory to the Higgs Mass Problem and the Proton Radius Puzzle. PhD Thesis, University of Chicago, Chicago, IL.
<https://www.proquest.com/openview/802c8e94756b95473db045ddc001cb8f/1?pq-origsite=gscholar&cbl=18750>
 - [30] Jaiswal, P. (2012) Higgs Physics in Supersymmetric Models. PhD Thesis, State University of New York at Stony Brook, Stony Brook, NY.
<https://inspirehep.net/files/9fd5583049700c33da0bf45bf64dcea8>
 - [31] Wacker, J.G. (2003) Little Higgs and the Hierarchy Problem. PhD Thesis, University of California, Berkeley, CA.
<https://www.proquest.com/openview/599b0d3e67b6bd5d4f9808d97d8fa828/1?pq-origsite=gscholar&cbl=18750&diss=y>
 - [32] Virdee, T.S. (2016) *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **374**, 20150259.
<https://doi.org/10.1098/rsta.2015.0259>
 - [33] Babu, K.S. and Mohapatra, R.N. (1990) *Physical Review Letters*, **64**, 2747-2750.
<https://doi.org/10.1103/PhysRevLett.64.2747>
 - [34] Breskvar, M., Lukman, D. and Mankoc Bortnik, N.S. (2006) On the Origin of Families of Fermions and Their Mass Matrices—Approximate Analyses of Properties of Four Families within Approach Unifying Spins and Charges. *9th Workshop on What Comes beyond the Standard Models*, Bled, 16-26 September 2006, 25-50.
 - [35] Gallas, S., Giacosa, F. and Rischke, D.H. (2009) About the Origin of the Mass of the Nucleon in a Linear SIGMA Model. *Acta Physica Polonica Series B Proceedings Supplement*, **2**, 319-324. <https://www.actaphys.uj.edu.pl/S/2/2/319/pdf>

- [36] Ioffe, B.L. (2006) *Physics-Uspekhi*, **49**, 1077-1078.
<https://doi.org/10.1070/PU2006v049n10ABEH006030>
- [37] Hanif, T. (2010) Generation of the Higgs Mass through Radiative Effects. PhD Thesis, University of Western Ontario, London, ON.
<https://ui.adsabs.harvard.edu/abs/2010PhDT.....257H/abstract>
- [38] Jegerlehner, F. (2014) *Acta Physica Polonica B*, **45**, 1167-1214.
<https://doi.org/10.5506/APhysPolB.45.1167>
- [39] Sarkar, U. (2000) *Pramana: Journal of Physics*, **54**, 101-118.
<https://doi.org/10.1007/s12043-000-0010-8>
- [40] Shima, K. and Tsuda, M. (2008) Nonlinear Supersymmetric General Relativity and Origin of Mass. *18th Workshop on General Relativity and Gravitation in Japan, JGRG 2008*, Higashi-Hiroshima, 17-21 November 2008, 1-4.
<https://inspirehep.net/files/41a0e624ce5581d65cbb9d087c3a70db>
- [41] Ajaib, M.A. (2014) Supersymmetry at the Crossroads: Higgs, Dark Matter and Collider Signals. PhD Thesis, University of Delaware, Newark, DE.
<https://udspace.udel.edu/items/6cabadb1-7e63-4cb8-8339-572c03bfcafe>
- [42] Cai, H. (2011) Aspects of Physics beyond the Standard Model. PhD Thesis, University of California, Davis, CA.
<https://www.proquest.com/openview/10002ea620d53757fc4c04eef241c7ea/1?pq-origsite=gscholar&cbl=18750>
- [43] Evans, J.L. (2009) Higgs Exempt No-Scale Supersymmetry. PhD Thesis, University of Michigan, Ann Arbor, MI.
https://deepblue.lib.umich.edu/bitstream/handle/2027.42/63863/jayevans_1.pdf?sequence=1
- [44] Huang, P. (2013) Natural Supersymmetry and the Higgs Boson. PhD Thesis, University of Wisconsin-Madison, Madison, WI.
<https://www.proquest.com/openview/932e9a56cbb3d37b34bb317c98c4a789/1?pq-origsite=gscholar&cbl=18750>
- [45] Walker, D.G.E. (2006) Theories on the Origin of Mass and Dark Matter. PhD Thesis, Harvard University, Cambridge, MA.
<https://www.proquest.com/openview/c89d998efe3c241012efa73bd3a1787e/1?pq-origsite=gscholar&cbl=18750&diss=y>
- [46] BenTov, J. (2013) Neutrino Oscillations, the Higgs Boson, and the Private Higgs Model. PhD Thesis, University of California, Santa Barbara.
<https://www.proquest.com/openview/728c33406815fb177769ae72de4926f7/1?pq-origsite=gscholar&cbl=18750>
- [47] Blechman, A.E. (2006) Topics in High-Energy Physics: The Standard Model and Beyond. PhD Thesis, Johns Hopkins University, Baltimore, MD.
<https://jscholarship.library.jhu.edu/handle/1774.2/7358>
- [48] Bonilla, C., Romao, J.C. and Valle, J.W.F. (2016) *New Journal of Physics*, **18**, Article ID: 033033. <https://doi.org/10.1088/1367-2630/18/3/033033>
- [49] Rashed, A.M.M. (2014) Higgs Effects in Neutrino Physics and Heavy Quark Systems. PhD Thesis, University of Mississippi, University, MS.
<https://egrove.olemiss.edu/etd/26/>
- [50] Blake, M., Tong, D. and Vegh, D. (2014) *Physical Review Letters*, **112**, Article ID: 071602. <https://doi.org/10.1103/PhysRevLett.112.071602>
- [51] Henning, B.Q. (2015) Precision Higgs Physics, Effective Field Theory, and Dark Matter. PhD Thesis, University of California, Berkeley, CA.

- <https://www.proquest.com/openview/5aef7dbb1061568df5486b2574215628/1?pq-origsite=gscholar&cbl=18750>
- [52] Jegerlehner, F. (2014) *Acta Physica Polonica B*, **45**, 1215-1254.
<https://doi.org/10.5506/APhysPolB.45.1215>
 - [53] Krauss, L.M. and Dent, J.B. (2013) *Physical Review Letters*, **111**, Article ID: 061802.
<https://doi.org/10.1103/PhysRevLett.111.061802>
 - [54] Servant, G. and Tulin, S. (2013) *Physical Review Letters*, **111**, Article ID: 151601.
<https://doi.org/10.1103/PhysRevLett.111.151601>
 - [55] Ma, E. (2014) *Physical Review Letters*, **112**, Article ID: 091801.
<https://doi.org/10.1103/PhysRevLett.112.091801>
 - [56] Krishnaswami, G.S. and Phatak, S.S. (2015) *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **471**, 20140803.
<https://doi.org/10.1098/rspa.2014.0803>
 - [57] Love, J.R. (2012) A Search for Technicolor at the Large Hadron Collider. PhD Thesis, Boston University, Boston, MA. <https://open.bu.edu/handle/2144/31586>
 - [58] Pons, D.J., Pons, A.D. and Pons, A.J. (2014) *Journal of Modern Physics*, **5**, 1980-1994.
<https://doi.org/10.4236/jmp.2014.517193>
 - [59] Fosdick, R. and Serrin, J. (2014) *Continuum Mechanics and Thermodynamics*, **26**, 287-302. <https://doi.org/10.1007/s00161-013-0301-1>
 - [60] Herman, P. (2004) *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, **218**, 731-736.
<https://doi.org/10.1243/0954406041319518>
 - [61] Wayne, R. (2012) *Turkish Journal of Physics*, **36**, 165-177.
 - [62] Broekaert, J. (2008) *Foundations of Physics*, **38**, 409-435.
<https://doi.org/10.1007/s10701-008-9210-8>
 - [63] Abdo, A.A., *et al.* (2009) *Nature*, **462**, 331-334.
 - [64] Fradkin, E.S. and Tseytlin, A.A. (1985) *Physics Reports*, **119**, 233-362.
[https://doi.org/10.1016/0370-1573\(85\)90138-3](https://doi.org/10.1016/0370-1573(85)90138-3)
 - [65] Smolin, L. (2015) Lessons from Einstein's 1915 Discovery of General Relativity. arXiv:1512.07551, 1-14. <http://arxiv.org/abs/1512.07551>
 - [66] Zwicky, F. (1937) *Astrophysical Journal*, **86**, 217-246.
<https://doi.org/10.1086/143864>
 - [67] Williams, T.B. (1975) *Astrophysical Journal*, **199**, 586-590.
<https://doi.org/10.1086/153726>
 - [68] Cai, Y.-F., *et al.* (2016) *Reports on Progress in Physics*, **79**, Article ID: 106901.
<https://doi.org/10.1088/0034-4885/79/10/106901>
 - [69] Milgrom, M. (1983) *Astrophysical Journal*, **270**, 371-389.
<https://doi.org/10.1086/161131>
 - [70] Moffat, J.W. (1979) *Physical Review D*, **19**, 3554-3558.
<https://doi.org/10.1103/PhysRevD.19.3554>
 - [71] Bahamonde, S., Böhmer, C.G. and Krššák, M. (2017) *Physics Letters B*, **775**, 37-43.
<https://doi.org/10.1016/j.physletb.2017.10.026>
 - [72] Bamba, K., Odintsov, S.D. and Sáez-Gómez, D. (2013) *Physical Review D*, **88**, Article ID: 084042. <https://doi.org/10.1103/PhysRevD.88.084042>
 - [73] Everett, H. (1973) The Theory of the Universal Wave Function: The Many-Worlds interpretation of Quantum Mechanics. In: DeWitt, B.S. and Graham, N., Eds., *Princeton Series in Physics*, Princeton University Press, Princeton, NJ, 266.

- [74] Einstein, A., Podolsky, B. and Rosen, N. (1935) *Physical Review*, **47**, 777-780.
<https://doi.org/10.1103/PhysRev.47.777>
- [75] Ellis, B. (2005) *Ratio*, **18**, 371-384. <https://doi.org/10.1111/j.1467-9329.2005.00300.x>
- [76] Giustina, M., Versteegh, M. A. M., *et al.* (2015) Significant-Loophole-Free Test of Bell's Theorem with Entangled Photons. *Physical Review Letters*, **115**, Article ID: 250401. <http://arxiv.org/abs/1511.03190>
<https://doi.org/10.1103/PhysRevLett.115.250401>
- [77] Pons, D.J., Pons, A.D. and Pons, A.J. (2017) *Journal of Modern Physics*, **8**, 1257-1274.
<https://doi.org/10.4236/jmp.2017.88082>
- [78] Bell, J.S. (1964) *Physics Physique Fizika*, **1**, 195-200.
<https://doi.org/10.1103/PhysicsPhysiqueFizika.1.195>
- [79] Colbeck, R. and Renner, R. (2011) *Nature Communications*, **2**, Article No. 411.
<https://doi.org/10.1038/ncomms1416>
- [80] Leggett, A. (2003) Nonlocal Hidden-Variable Theories and Quantum Mechanics: An Incompatibility Theorem. *Foundations of Physics*, **33**, 1469-1493.
<https://doi.org/10.1023/A:1026096313729>
- [81] de Broglie, L. (1929) The Wave Nature of the Electron. In: *Nobel Lecture*, Nobel Prize in Physics.
- [82] Bohm, D. and Bub, J. (1966) *Reviews of Modern Physics*, **38**, 453-469.
<https://doi.org/10.1103/RevModPhys.38.453>
- [83] Pons, D.J., Pons, A.D. and Pons, A.J. (2019) *Journal of Modern Physics*, **10**, 835-860.
<https://doi.org/10.4236/jmp.2019.107056>
- [84] Pons, D.J. (2021) *Journal of Modern Physics*, **12**, 1931-1953.
<https://doi.org/10.4236/jmp.2021.1214110>
- [85] Pons, D.J. (2020) *Journal of Modern Physics*, **11**, 1598-1619.
<https://doi.org/10.4236/jmp.2020.1110100>
- [86] Krššák, M., *et al.* (2019) *Classical and Quantum Gravity*, **36**, Article ID: 183001.
<https://doi.org/10.1088/1361-6382/ab2e1f>
- [87] Wanas, M.I. (2002) Absolute Parallelism Geometry: Developments, Applications and Problems. arXiv:gr-qc/0209050. <https://arxiv.org/abs/gr-qc/0209050>
- [88] Wanas, M.I., Melek, M. and Kahil, M.E. (1995) *Astrophysics and Space Science*, **228**, 273-276. <https://doi.org/10.1007/BF00984980>
- [89] Wanas, M.I. and Kahil, M.E. (1999) *General Relativity and Gravitation*, **31**, 1921-1929.
<https://arxiv.org/abs/gr-qc/9912007>
<https://doi.org/10.1023/A:1026743007086>
- [90] Wanas, M.I. (1997) *Astrophysics and Space Science*, **258**, 237-248.
<https://doi.org/10.1023/A:1001747710135>
- [91] Pons, D.J. (2015) *Applied Physics Research*, **7**, 14-26.
<https://doi.org/10.5539/apr.v7n2p1>
- [92] Pons, D.J. (2020) *Optics*, **1**, 243-254. <https://doi.org/10.3390/opt1030018>
- [93] Pons, D.J. and Pons, A.D. (2013) *The Open Astronomy Journal*, **6**, 77-89.
- [94] Pons, D.J., Pons, A.D. and Pons, A.J. (2013) *Applied Physics Research*, **5**, 145-174.
<https://doi.org/10.5539/apr.v5n6p145>
- [95] Pons, D.J., Pons, A.D. and Pons, A.J. (2013) *Applied Physics Research*, **5**, 107-126.
<https://doi.org/10.5539/apr.v5n5p107>
- [96] Pons, D.J., Pons, A.D. and Pons, A.J. (2013) *Applied Physics Research*, **5**, 23-47.
<https://doi.org/10.5539/apr.v5n6p23>

- [97] Pons, D.J., Pons, A.D. and Pons, A.J. (2014) *Physics Essays*, **27**, 26-35.
<https://doi.org/10.4006/0836-1398-27.1.26>
- [98] Pons, D.J., Pons, A.D. and Pons, A.J. (2014) *Applied Physics Research*, **6**, 28-46.
<https://doi.org/10.5539/apr.v6n2p28>
- [99] Pons, D.J., Pons, A.D. and Pons, A.J. (2014) *Applied Physics Research*, **6**, 50-63.
<https://doi.org/10.5539/apr.v6n3p50>
- [100] Pons, D.J., Pons, A.D. and Pons, A.J. (2015) *Physics Research International*, **2015**, Article ID: 651361. <https://doi.org/10.1155/2015/651361>
- [101] Pons, D.J., Pons, A.D. and Pons, A.J. (2015) *Applied Physics Research*, **7**, 1-13.
<http://vixra.org/abs/1412.0279>
<https://doi.org/10.5539/apr.v7n2p1>
- [102] Pons, D.J., Pons, A.D. and Pons, A.J. (2015) *Applied Physics Research*, **7**, 1-11.
<https://doi.org/10.5539/apr.v7n1p1>
- [103] Pons, D.J., Pons, A.D. and Pons, A.J. (2015) *Applied Physics Research*, **7**, 18-29.
<https://doi.org/10.5539/apr.v7n3p18>
- [104] Pons, D.J., Pons, A.D. and Pons, A.J. (2015) *Journal of Nuclear and Particle Physics*, **5**, 58-69. <https://doi.org/10.1155/2015/651361>
- [105] Pons, D.J., Pons, A.D. and Pons, A.J. (2016) *Applied Physics Research*, **8**, 111-121.
<https://doi.org/10.5539/apr.v8n3p111>
- [106] Pons, D.J., Pons, A.D. and Pons, A.J. (2016) *Journal of Modern Physics*, **7**, 1277-1295.
<https://doi.org/10.4236/jmp.2016.710113>
- [107] Pons, D.J., Pons, A.D. and Pons, A.J. (2016) *Journal of Modern Physics*, **7**, 1049-1067.
<https://doi.org/10.4236/jmp.2016.710094>
- [108] Pons, D.J. (2015) Internal Structure of the Electron (Image Licence Creative Commons Attribution 4.0). Wikimedia Commons, (Creative Commons Attribution 4.0 International License).
https://commons.wikimedia.org/wiki/File:Internal_structure_of_the_electron.jpg
- [109] Michelson, A.A. and Morley, E.W. (1887) *American Journal of Science*, **34**, 333-345.
<https://www.ajsonline.org/content/s3-34/203/333>
<https://doi.org/10.2475/ajs.s3-34.203.333>
- [110] Fabbri, L. (2021) *Universe*, **7**, Article 305. <https://doi.org/10.3390/universe7080305>
- [111] Kahil, M.E. and Harko, T. (2008) Is Dark Matter an Extra-Dimensional Effect? arXiv:0809.1915. <https://arxiv.org/abs/0809.1915>
- [112] Kahil, M.E (2019) *Gravitation and Cosmology*, **25**, 268-276.
<https://arxiv.org/abs/1801.08396>
<https://doi.org/10.1134/S020228931903006X>
- [113] Schrabback, T., *et al.* (2010) *Astronomy and Astrophysics*, **516**, A63.
<https://doi.org/10.1051/0004-6361/200913577>
- [114] Goobar, A., *et al.* (2000) *Physica Scripta*, **2000**, 47-58.
<https://doi.org/10.1238/Physica.Topical.085a00047>
- [115] Schwarz, D.J. and Weinhorst, B. (2007) *Astronomy and Astrophysics*, **474**, 717-729.
<https://doi.org/10.1051/0004-6361:20077998>
- [116] Cao, S., Ryan, J., *et al.* (2021) *Monthly Notices of the Royal Astronomical Society*, **501**, 1520-1538.
- [117] Straumann, N. (1999) *European Journal of Physics*, **20**, 419-427.
<https://doi.org/10.1088/0143-0807/20/6/307>

- [118] Tu, Z.L., Hu, J. and Wang, F.Y. (2019) *Monthly Notices of the Royal Astronomical Society*, **484**, 4337-4346.
- [119] Sanders, R.H. (1998) *Monthly Notices of the Royal Astronomical Society*, **296**, 1009-1018. <https://doi.org/10.1046/j.1365-8711.1998.01459.x>
- [120] Li, X.-D., *et al.* (2018) *The Astrophysical Journal*, **856**, 88.
- [121] Navarro, J.F., Frenk, C.S. and White, S.D.M. (1997) *The Astrophysical Journal*, **490**, 493-508. <https://doi.org/10.1086/304888>
- [122] Gao, L., Navarro, J.F., *et al.* (2008) *Monthly Notices of the Royal Astronomical Society*, **387**, 536-544. <https://doi.org/10.1111/j.1365-2966.2008.13277.x>
- [123] Brans, C. and Dicke, R.H. (1961) *Physical Review*, **124**, 925-935. <https://doi.org/10.1103/PhysRev.124.925>

A Physical Explanation for the Formation of Auroras

WeiHong Qian^{1,2*}

¹Guangzhou Institute of Tropical and Marine Meteorology, CMA, Guangzhou, China

²School of Physics, Peking University, Beijing, China

Email: *qianwh@pku.edu.cn

How to cite this paper: Qian, W.H. (2023)

A Physical Explanation for the Formation of Auroras. *Journal of Modern Physics*, 14, 271-286.

<https://doi.org/10.4236/jmp.2023.143018>

Received: November 15, 2022

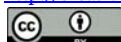
Accepted: February 17, 2023

Published: February 20, 2023

Copyright © 2023 by author(s) and Scientific Research Publishing Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

What mechanism causes the symmetrical distribution of two oval-ring auroras approximately around the geomagnetic poles but uneven brightness distribution of them? To answer this question, we firstly describe the charged particles like electrons or protons emitted from the Sun and the magnetic ions formed in the Earth's atmosphere. Then, the interaction dynamics of two-type particles between electron and ion is given under the non-relativistic limit. Finally, under the relativistic frame, auroras deduced are higher energy density formed by the orthogonal interaction of solar charged particles and geomagnetic ions in the narrow regions centered on the upper geomagnetic poles. The physical nature of ideal oval-ring auroras with uneven brightness distribution is an optical phenomenon that occurs when solar charged particles collide orthogonally with ions and the magnetic axis inclines to the solar radiation. For actual aurora distribution, the impact of multiple factors is discussed. Therefore, the aurora is a natural illustration of an orthogonal collider in the Earth's upper atmosphere.

Keywords

Aurora, Electron, Ion, Orthogonal Collision, Space Weather

1. Introduction

One of the most spectacular shows that the natural light offers in the northern latitudes is the so-called Aurora Borealis [1]. Aurora is a colorful plasma phenomenon that occurs due to the flow of charged particles (solar wind) from the Sun entering the Earth's magnetic field so that the brilliant and beautiful glow appears at clear-sky night near the north and south poles of the Earth. They are respectively referred as the Aurora Borealis (northern lights) near the north pole

and Aurora Australis (southern lights) near the south pole. As early as the 13th century to the middle of the 18th century, 1538 days of Aurora Borealis were recorded below 55 degrees at the northern mid-latitudes [2]. At the Space-Weather website of USA [3], the northern lights are thought to be the result of accelerated electrons or strong electrically charged solar wind particles colliding with air molecules (or oxygen and nitrogen atoms) in the Earth's upper atmosphere. In these collisions, the electrons transfer their energy to the upper atmosphere thus exciting the atoms and molecules to higher energy states, so that beautiful color lights can be observed by human naked eyes in the high latitudes. The auroras typically form about 80 to 500 km above the Earth's surface and usually occur between 60 and 75 degrees of the latitude. Therefore, the first question is why do auroras occur frequently only in these ranges of latitude and altitude?

In the above process of forming the northern and southern lights, a key physical event is the acceleration of electrons and collision with other particles. It is generally believed that the acceleration of electrons is associated with the cause of geomagnetic storms during strong solar activities [4]. The Earth's magnetic fields converge and guide the electrons or other charged particles at the upper atmosphere such that auroras form two bright ovals approximately centered at the geomagnetic poles [5] [6]. The geomagnetic intensity is varied in the northern and southern hemispheres during geomagnetic quiet and disturbed times [7]. The geomagnetic activity reaches the maximum around the magnetic latitude of 70 degrees [8]. Auroral oval boundary dynamics is associated with geomagnetic storm intensity [9]. When stronger geomagnetic storms occur, auroras extend frequently equator-ward and increase the appearance of space weather activity [10] [11]. Therefore, auroras are considered as one of space weather phenomena. The best time to observe the aurora is at the mid-night in a cold winter because the skies must be clear, dark and free of clouds [12] [13]. Two questions can be asked here. Why the best time to observe the aurora is in clear skies and cold winter midnight? And how do high-energy electrons or other charged particles collide only with upper atmospheric molecules (or atoms) to excite new high-energy states during a solar activity?

Historically, auroras can only be observed by human naked eyes at the ground. Recent decades, satellites play an important role in the observation of auroras and the distribution of their ion-electron energies [14]. The Aurora Borealis and Aurora Australis have been observed by special all-sky sensor ultraviolet scanning imagers on board satellites. The global positioning system (GPS) phase scintillation caused by high-latitude ionospheric irregularities during an intense high-speed stream of the solar wind from April 29 to May 5, 2011, was observed using arrays of the GPS ionospheric scintillation and total electron content monitors in the Arctic and Antarctica [15]. Sequences of scintillation events were observed as the aurora brightened and/or drifted across the field of view of the all-sky imager. In the northern and southern hemispheres, the image captured

an auroral arc brightening at the pole-ward edge of the auroral oval, with an arc being detached. Two successive auroral image scans showed that the oval-ring brightness was not uniform. Three oval-ring events of current aurora were also observed by satellite and ground imagers in the two hemispheres on May 16, 2005, August 26, 2018 and September 5, 2005 respectively [16]. The oval-ring current auroras appeared as auroral spots or arcs in the day, dusk or night side. Similarly, ring-oval auroras were also observed by the Galilean satellites at the Jupiter and its moons [17] [18]. The last question can be raised from the auroral image scans: what causes the pattern of oval-ring auroral arc with its distribution of inhomogeneous brightness?

It can be summarized that the formation of auroras is caused by high-energy electrons from the Sun entering the Earth's upper atmosphere in the narrow regions centered on the magnetic poles. The energy of molecules (atoms) in the upper atmosphere is limited so we need to examine whether the abnormal energy of the Sun's activity can collide with molecules (atoms) to form a new high-energy state. As a high-altitude space weather phenomenon, the first two questions are about the spatial and temporal distributions of auroras in the climate view, while the latter two questions are associated with the physical nature of auroral formation. These four questions need to be answered by a unified physical theory. Thus, this paper first describes how charged particles produced by the Sun and the Earth emit to space and then gives a microscopic description about how high-energy particles emitted from the Sun interact with the electronically charged Earth's particles in the narrow regions centered on the upper geomagnetic poles. The goal is to theoretically describe that the northern and southern lights are physically originated from the interaction of different charged particles. Finally, the paper gives conclusions and discussion.

2. Particles Produced by the Sun and the Earth

As space weather phenomena, the formation of auroras is inseparable from the space environment of the universe. In addition to the presence of various celestial bodies, including stars, planets and moons, the universe is also filled with radiation and reflection of particles from these celestial bodies. The Sun that can shine is the only star in the solar system. The Earth is a planet in the solar system with a geomagnetic field. The Moon is a natural satellite of the Earth without a magnetic field. We first discuss why three objects the Sun, Earth and Moon can emit or reflect radiation. The Moon is already a solid ball so that different layers can no longer be circled inside. The bright Moon can be seen by human naked eyes from the Earth at night because the Moon reflects the sunlight. Similarly, the Apollo astronauts on the Moon can also see our own beautiful Earth reflecting the sunlight. Stars outside our solar system also emit light beams and particles from far away, but their luminous and particle fluxes are too small, and their reflections on the Earth and Moon are also small.

As a planet in the solar system, the Earth is composed of multiple concentric

layers of matter in different states. In the simplest form, the Earth is composed like an egg, with a core as its central part, a middle yolk as fluid layer, and a thin shell covered in the whole outside. The Earth's core is made up of heavy metals (iron-nickel core). The Earth's fluid layer is composed of mantle magma, and the Earth's crust is the continental-oceanic shell which is composed of lighter silicon-magnesium material and water. The outer part of the Earth's crust is an atmosphere up to 1000 km thick. Since large amounts of water vapor are concentrated in the layer more than 10 kilometers called the troposphere above the surface, daily weathers such as normal clouds and extreme storms occur in the troposphere. Auroras occur in the ionosphere that is about 80 kilometers above the surface, where temperatures of thermosphere are higher and there are fewer water vapor molecules and dust particles. Auroras formed over the Earth's poles are associated with the interaction between energetic ions in the upper atmosphere and the charged particles coming to there.

We now analyze possible magnetic fields of the above three celestial objects from the Earth to the Moon and the Sun. The compass used by ancient people for navigation has proved that the Earth is a huge magnet with a geomagnetic field. Electricity and magnetism are closely linked. There is a relative motion between the mantle magma layer and the solid core inside the Earth. The relative motion between the adjacent layers acts as electrical generators to produce the Earth's electric field and magnetic field. The Earth rotates around its axis. The geomagnetic field direction or polarity is reversed depending on whether the mantle rotates relative to the core fast or slow, which is called the geomagnetic field reversal. In the geological history of the Earth, the geomagnetic field has been inverted many times [19]. Geomagnetic inversion was recorded in sediments distributed parallel on both sides of the mid-ocean ridge [20] [21].

The remanent magnetization of lunar soils suggests that the formation and evolution of the Moon are similar to those of the Earth [22]. In the early days of the formation of the Moon, the relative motion between the lunar iron core and the lunar mantle magma layer also formed a magnetic field, which recorded on rocks and soils of the lunar crust. But the Moon is smaller in both size and mass than Earth, so it also evolves geologically faster than the Earth. Today's lunar core, mantle and crust have all solidified together so that there is no atmosphere and magnetic field surrounding the current Moon. The remanence on the Moon now is the lunar historical record of the relative motion of the lunar inner layers over geological time [23].

There is also relative motion between multiple layers in the interior of the Sun. Therefore, the Sun is a huge electrical generator, and there is also a huge solar magnetic field [24]. The relative motion of the different layers of the Sun forms a convective circulation in the latitudinal zone inside each circle and a motion of many vortexes in the horizontal plane of the same circle. These convective and vortex motions are local generators which form a local solar magnetic field that manifests itself as solar activities, such as the appearance of sunspots [25] [26]

[27]. In effect, these convective and vortex motions form nuclear fusion on the Sun. Scientists and engineers have continued to study the Sun's fusion process in hopes of one day using nuclear fusion to generate electricity [28] [29].

These charged particles excited from the solar activities have a high velocity. The momentum of a solar particle is the ability to break free from the mass inertia that originally moved in the same way as the Sun as a whole [30]. From Newton's point of view, it is to break free from the Sun's huge gravitational field. To use a formula, a particle with a little mass m on the Sun wants to leave the Sun. Thus, its acceleration a_m must be equal to or greater than the gravitational acceleration $g_{\odot}$ of the Sun on it

$$a_m \geq g_{\odot}. \quad (1)$$

If a particle can break free from the Sun, it is an electron or proton. Electrons are constantly breaking free from the Sun's corona and radiate into space, with one beam reaching the Earth and the other reaching the Moon. The flux and intensity of solar particles reaching the Earth and Moon varies with the solar activities.

The Earth's particles, especially those in the Earth's atmosphere, are also trying to break free from the same inherent inertia as the Earth [30]. In Newton's gravitational discourse, the particles on the Earth also want to leave the geo-gravitational field and become particles heading into space. But the low-velocity relative motion between the inner iron core and the mantle magma layer is not powerful enough to get these particles out of the geo-gravitational field. Otherwise, the Earth's atmosphere would have disappeared long ago. For this reason, the low-velocity relative motion between the Earth's inner layers formed the geomagnetic field.

The Earth's materials can be roughly divided into three types: solid, liquid and gaseous. The Earth's core and crust are solid layer structures. The Earth's mantle, especially in the early days of Earth's formation, is a liquid magma layer. The global ocean above the Earth's crust occupies liquid water, but separated by continents so that it is not entirely circle-structured. The Earth's outer layer is structured by the atmosphere. Although the low-velocity relative motion between the Earth's inner layers cannot produce the ability of particles to leave the Earth, the electromagnetic field formed by the relative motion between the circle layers will change the motion and distribution of particles (ions) in the Earth's upper atmosphere.

The geomagnetic poles do not coincide completely with the geographical poles. There is a geomagnetic declination. The physical nature of the geomagnetic poles is the distribution of the Earth's electromagnetic field formed by the overall low-velocity fluid of the inner mantle relative to the inner core. In addition to the overall movement of the mantle layer relative to the inner core, there is also a local movement relative to the inner core and outer crust in the mantle fluid, which also forms local geomagnetic disturbances. Moreover, the local movement of the mantle relative to the Earth's crust can lead to earthquakes and

volcanic activities. Volcanic activities on the mid-ocean ridges record the past information about changes in local geomagnetic fields and geomagnetic field reversals [31].

In **Figure 1**, the range that the magnetic field lines can reach is mainly in the atmosphere within 1000 km. Within this layer, the relatively dense magnetic field lines are located at the narrow areas over the north and south poles, while sparse lines are found in the vast areas over the mid-low latitudes. The density of magnetic field lines can be seen as places in the atmosphere where charged particles (ions) are dense and moving fast. Dense magnetic field lines around the north and south magnetic poles form a narrow area like a funnel-shaped distribution. It is an area in where exists a relatively high density of ions in the mid-upper atmosphere.

From the above analysis, we have shown that the Sun as a star and the Earth as a planet excite different particles. The relative motion between the different layers in the Sun's interior excites solar particles and radiates them into the surrounding space. The relative motion between the different layers in the Earth's interior excites ions in the Earth's atmosphere and moves around the Earth with magnetic field lines. The relative motion between the inner layers of the Moon has stopped, so that there is no magnetic field on the Moon now.

3. Interaction Dynamics of Two-Type Particles

According to the current understanding, auroras are the result from the solar charged particles interacting with the geomagnetic ions in the Earth upper atmosphere. Therefore, auroras are optical phenomena exhibited by the interaction of high-energy particles. We now analyze the dynamics of high-energy particles to see how this phenomenon is excited. Any particle in the universe has

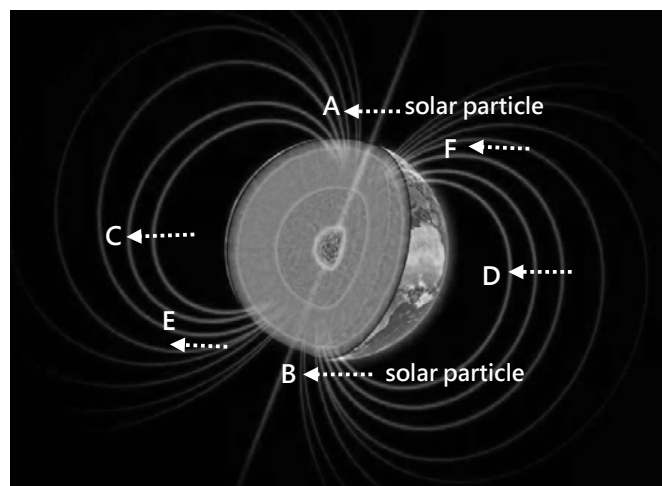


Figure 1. An ideal schematic pattern of the distribution of solar particles (dotted-line arrows) and geomagnetic field lines (white lines). Letters A, B, C, and D indicate four places where anomalous solar particles collide orthogonally with geomagnetic ions moved in the geomagnetic field lines. Letters E and F indicate two places where particles and ions respectively collide with their angles of 0 and 180 degrees along a straight line.

two forms of motion: curved motion and straight motion of acceleration or deceleration. A particle forms a curved motion as well as a motion of acceleration or deceleration, which is affected by the centripetal force and the mass force. We examine the interaction that occurs when two low-velocity particles α and β have a centripetal force $\frac{m}{r}v^2\mathbf{n}$ and mass force $m\mathbf{a}$. The centripetal and mass forces of the two low-velocity particles are respectively

$$\mathbf{F}_\alpha = \frac{m_\alpha}{r_\alpha}v_\alpha^2\mathbf{n}_\alpha + m_\alpha\mathbf{a}_\alpha, \quad (2)$$

$$\mathbf{F}_\beta = \frac{m_\beta}{r_\beta}v_\beta^2\mathbf{n}_\beta + m_\beta\mathbf{a}_\beta. \quad (3)$$

For the collision of two particles, their interaction can be expressed as the vector product of two forces, which is the shear stress [32]

$$\boldsymbol{\tau}_C = \left(\frac{m_\alpha}{r_\alpha}v_\alpha^2\mathbf{n}_\alpha + m_\alpha\mathbf{a}_\alpha \right) \times \left(\frac{m_\beta}{r_\beta}v_\beta^2\mathbf{n}_\beta + m_\beta\mathbf{a}_\beta \right). \quad (4)$$

If the motion of acceleration or deceleration of particles is not taken into account, *i.e.*, $\mathbf{a}_\alpha = \mathbf{a}_\beta = 0$, we have

$$\boldsymbol{\tau}_C = \left(\frac{m_\alpha}{r_\alpha}v_\alpha^2\mathbf{n}_\alpha \right) \times \left(\frac{m_\beta}{r_\beta}v_\beta^2\mathbf{n}_\beta \right). \quad (5)$$

We use Equation (5) in **Figure 1** for the collision of a particle from the Sun with a particle in the Earth's upper atmosphere. It is well known that the vector-product of the two forces is another vector, namely $\mathbf{n}_\alpha \times \mathbf{n}_\beta = \mathbf{n}_\gamma$ so that the shear stress is usually defined as a force per surface area. This vector $\mathbf{n}_\gamma$ is perpendicular to the plane formed by those two vectors between $\mathbf{n}_\alpha$ and $\mathbf{n}_\beta$. The modulus of vector-product of the two forces based on Equation (5) is

$$\tau_C = \left(\frac{m_\alpha}{r_\alpha}v_\alpha^2 \right) \cdot \left(\frac{m_\beta}{r_\beta}v_\beta^2 \right) \sin \theta. \quad (6)$$

where r_α and r_β are the radius of two moving particles with their masses m_α and m_β as well as their velocities v_α and v_β , respectively. The letter θ is the angle at which two particles collide. The modulus reaches the maximum

$$\tau_C = 4 \left(\frac{1}{2} m_\alpha v_\alpha^2 \right) \left(\frac{1}{2} m_\beta v_\beta^2 \right) / (r_\alpha r_\beta) = 4E_\alpha E_\beta / (r_\alpha r_\beta) \quad (7)$$

only when the collision angle between the two particles is orthogonal (90 degrees). Otherwise, when the collision angle between them is larger or less than 90 degrees, the modulus of shear stress will be reduced. Since the Sun rotates, it also rotates with the Milky Way. The Earth rotates and revolves around the Sun. Therefore, particles emitted by a star and the ions in a planet's atmosphere do not move in a straight line, but there is a centripetal motion with a radius.

Although the shear stress is another vector, we only concern its modulus which is really associated with whether it can produce light phenomena (new physical

state) or not. In Equation (7), $E_\alpha = \frac{1}{2}m_\alpha v_\alpha^2$ and $E_\beta = \frac{1}{2}m_\beta v_\beta^2$ are energies of two particles. Before the collision, the two energies are separated. At the moment of the orthogonal collision, the energy of particle β has increased by $4E_\alpha$ times. The term $4E_\alpha E_\beta$ is a total energy at the collision moment. This total energy is originally scattered over a rectangular area by two radius r_α and r_β . Therefore, the term $4E_\alpha E_\beta / (r_\alpha r_\beta)$ can be simply referred to as energy density. Actually, after the orthogonal collision, two particles have been completely crushed into a new physical state, exhibiting a huge energy density. Before the collision of two particles, the total energy $4E_\alpha E_\beta$ on the area ($r_\alpha \times r_\beta$) is known and should be conserved with time. The total energy $4E_\alpha E_\beta$ will be concentrated on a very small three-dimensional space point A after the collision in **Figure 1**. The energy density of the orthogonal collision of two particles is extremely large and can be called as a new high-energy state or a plasma phenomenon.

The actual interaction between particles and ions differs from the above theoretical derivation in two differences. The first is the number difference between particle pairs. An actual aurora is produced by a large number of continuous solar particles interacting with geomagnetic ions. When a large number of solar particles collide orthogonally with a large number of geomagnetic ions, the energy density excited is huge, forming a new substance, which represents an aurora. In the above theory, we only give the interaction of a pair between a solar particle and a geomagnetic ion as an illustration. But the latter has theoretically given a clear explanation for the dynamics of the former's aurora formation.

The second is the difference in the velocity of particles. The particles used in the theoretical derivation process belong to the low-velocity state and use the non-relativistic limit. In fact, the velocity of particles emitted by the Sun to the Earth's poles is very high. The velocity of ions in the Earth's atmosphere along the geomagnetic field lines is also high. Thus, the relativistic momentum must be considered in the formation of auroras.

The relativistic momentum for a particle mass m_i and velocity v_i is

$$p_i = \gamma_i m_i v_i, \quad (8)$$

and the relativistic energy is

$$E_i = \gamma_i (m_i v_i^2), \quad (9)$$

while

$$\gamma_i = \frac{1}{\sqrt{1 - v_i^2/c^2}}. \quad (10)$$

is the Lorentz factor related to the speed of light c [33].

Thus, the high energy density $Q_{s,G}$ collided orthogonally by a solar particle and a geomagnetic ion is concentrated at the point A or B in a three-dimensional space. From Equation (7),

$$Q_{s,G} = \gamma_s (m_s v_s^2) \gamma_G (m_G v_G^2) = 4\gamma_s E_s \gamma_G E_G \quad (11)$$

where m_s and m_G are respectively masses of solar particle and geomagnetic ion with their velocity v_s and velocity v_G . Equation (11) indicates that the high energy density is the multiplication of energies between two-type particles. The space (or area by a particle thickness) after orthogonal collision of two particles is very small so that the energy density is very high. There are many points of orthogonal collisions between two-type particles over the polar atmosphere. These points form bright areas where they show Auroras. Here, only relativistic energy of particle interaction is considered. The wave-particle interaction will be discussed late.

4. Orthogonal Interaction between Solar Particles and Geomagnetic Ions

The upper magnetosphere of the Earth's atmosphere has higher temperatures so that it is called as the thermosphere. However, due to the asymmetry of the thermal distribution, the high-level atmospheric circulation around the poles is often not symmetrical relative to the north and south poles. Also, produced major total particle content varies at different latitudes and longitudes during enhanced magnetospheric storm processes, which showed the strong longitudinal and hemispherical asymmetry [34]. Therefore, the movement of various charged particles and ions around the polar region will also be relatively polar asymmetrical on the planetary scale, and even local disturbances will vertically occur. For example, regional temperature inversion can vertically form varied ion content. But they should still have a fundamental large-scale feature from the distribution of geomagnetic ions.

The spatial distribution of the geomagnetic field is not isolated and is affected by the solar wind (solar radiated or charged particles). How do solar particles affect the structure of the geomagnetic field? We first give an ideal structure while the actual asymmetrical or modified structure will be discussed late. Solar particles arriving near the Earth can be watched as moving in a straight line at a certain fractional speed of light. In **Figure 1**, solar particles reach six different points (A, B, C to F) in the geomagnetic field. When the solar particles ideally reach point A near the Earth's North Pole, they collide orthogonally with ions in the geomagnetic field. If the Sun is in an active phase, or the geomagnetic field is in an enhanced period, or both solar and geomagnetic activities are at an enhanced moment, an orthogonal collision at point A produces a higher energy state. The high-energy state there can present strong lights or colorful plasma phenomenon, which are the northern lights (Aurora Borealis). When other solar particles ideally reach point B near the South Pole, ions in the Earth's upper atmosphere collided orthogonally with solar particles, produce a new high-energy state symmetrical to point A and the excited the southern lights (Aurora Australis).

The interaction of solar particles reaching points E and F with geomagnetic ions in the Earth's upper atmosphere is not orthogonal, but two positions are

symmetrical. We look at the linear collision between two particles happened in **Figure 1**. The linear collision is commonly used in all current colliders [33]. The rear-end collision and head-on collision of the two-type particles are set at the points E and F, respectively. The total energies of linear collision between a solar particle and a geomagnetic ion are addition or subtraction of their energies. The energy of rear-end collision at the point E is the subtraction.

$$E_E = \gamma_S E_S - \gamma_G E_G. \quad (12)$$

This is at the point E where the solar particle's speed is faster than that of the geomagnetic ion at the same direction when they collide.

The total energy of head-on collision at the point F is the sum or addition of their energies [33].

$$E_F = \gamma_S E_S + \gamma_G E_G. \quad (13)$$

It indicates that the solar particle is collided with the geomagnetic ion at the point F at the opposite direction.

The head-on collision and the rear-end collision only have total energy at the points E and F because their energy density is zero when the collision angle of two particles is 0 or 180 degrees based on Equation (6). Therefore, there should theoretically form two symmetrical oval rings of brightness over the two polar geomagnetic fields but the brightness along each oval ring is not symmetrically bright in one half than the other half. The oval light ring with a distribution of asymmetrical brightness is formed by because the Earth's axis is tilted relative to the solar radiation. The energy density of two-type particles collided at the points A and B reaches the maximum which formed new physical state while the energy density is zero at the points E and F without formed new physical state. This answers the question why the intensity distribution of an aurora is asymmetry, as observed by satellites. At the points C and D in **Figure 1**, the energy density of solar particles colliding orthogonally to the upper atmospheric ions along the geomagnetic field lines is low because the ion density is sparse there. Solar radiation can reach the point D on the front of the Earth, where solar particles collide with few ions orthogonally. Since this happens during the daytime, people on the Earth cannot see the abnormal energy formed by the collision. Actually, near magnetic local noon, the diffuse aurora was often observed in structured forms [35]. The point C located on the back of the Earth is in the nighttime sky, where no solar particles interact with geomagnetic ions.

A comprehensive investigation of the interaction in **Figure 1** shows that solar particles interact with geomagnetic ions on the dense band of geomagnetic field lines at the poles to form a high-energy distribution or new physical state, which is where the aurora appears. As the Earth rotates, the strong aurora also changes with the position of the orthogonal collision between solar particles and geomagnetic ions. When their collision angle is less than or larger than 90 degrees, the aurora is weak according to Equation (6) for their modulus.

The density and velocity of polar-area ions vary with particle flow and the

strength of geomagnetic field. When the strength and density of ions in the geomagnetic field are abnormal and, at the same time, the solar winds are abnormal, the energy density excited by the orthogonal collision between solar particles and geomagnetic ions will be intensified, forming abnormal auroras. Therefore, the spatial distribution of auroras in the two polar areas is neither completely symmetrical, nor periodic in time.

Whether auroras can be observed on the ground is affected by many weather conditions. In a rainy weather day, auroras are not visible there over the two polar areas. The abnormal distribution of ions and other particles in the ionosphere can also affect the formation of auroras, and even change the color of auroras through reflection and refraction. Therefore, clear skies and cold winter are possibly better conditions to see auroras in the two polar areas.

5. Conclusions

On the basis of collecting people's cognition of auroras, this paper uses a unique dynamic theory to reveal the physical nature of auroral formation. Auroras are optical phenomena in which atmospheric ions interact with solar charged particles at the height of the Earth's polar ionosphere. When the geomagnetic field is abnormal or/and an active solar phase, solar charged particles and geomagnetic ions will collide more violently in the polar space-time domain, forming a super strong auroral phenomenon. The theoretical distribution of high energy density or new physical state formed by the orthogonal collision of two-type particles can be applied to explain auroras occurring at the two poles of the Earth.

Theoretically, the maximum energy density can only occur at the points A and B from **Figure 1**, but high-energy rays produced by the new physical state, known as auroras, can only be observed over the polar areas at clear night. There is no auroral light at the points E and F because only the rear-end collision and the head-on collision appear there respectively. Outside the points E and F, auroras are not as strong as at the points A and B. The reason is that the collision angle between the solar particles and the Earth's atmospheric ions is larger or less than 90 degrees.

The theory in this paper and **Figure 1** are sufficient to answer the four questions raised. The fact that auroras often appeared at 60 - 75 degree latitudes and 80 - 500 km altitudes is determined by the condition of orthogonal interaction between the solar charged particles and geomagnetic ions. The auroras are the optical phenomena or the high-energy or new physical state of two types of high-speed particles interacting in the upper atmosphere. The oval-ring aurora observed by satellites shows that one part is brighter than the other with a reason because the Earth's axis is tilted relative to the solar radiation. Clear skies and cold winter midnight are the best chances to observe auroras because the tropospheric weather environment sometimes influences the observer's perspective, not a physical one.

In each of planetary atmosphere of the solar system, and even in the atmos-

phere of natural moons, auroras at their two poles can often appear as long as they have their own magnetic fields. Another common condition for forming auroras is the solar activity. The Jupiter and its moons also have distributed auroras at their two polar areas. All auroras on the two polar ionospheres of planets and moons show an oval-ring distribution of brightness because their axes are tilted to the solar radiation.

The physics community has spent a lot of efforts building accelerators and colliders [36] [37] [38]. Their aim is to increase the speed and energy of known particles through accelerators. Current colliders are used to producing higher energies and tried to form new matter through a head-on collision. Since head-on collisions are linear ones, the formed energy is limited and predictable [39]. Therefore, in current linear colliders, it seems difficult to achieve the further goal that physicists are looking for new substances or new physical state. The auroras that occur in the middle and upper parts of the polar geo-ionospheres are a natural orthogonal collider. This paper has explored the formation mechanism of auroras, which implies that the dream of the scientific community to find new matter can be enhanced if orthogonal high-energy colliders can be built.

6. Discussion

This paper presents a symmetrical or idealized model of the geomagnetic field that expresses the interaction of charged particles from the Sun with geomagnetic ions to form oval-ring auroras with uneven brightness distribution near the poles. Auroras are anomalous space weather that occurs in the ionosphere of the upper atmosphere, like extreme weather in the troposphere, so that they have their corresponding anomalous patterns, but each weather event has differences in intensity, coverage, and duration. For the possible causes of aurora formation in multiple structures and multi-luminous distributions, we give the following discussion.

1) Asymmetric geomagnetic distribution under the interaction of the Sun's magnetic field with the Earth's magnetic field. As previously described in the solar system, the Earth's magnetic field is not the only one that exists. The Sun has a magnetic field by the strength which affects the flow of high-energy charged plasma particles in all directions to form the solar wind. When the solar wind reaches the vicinity of the Earth, it interferes with the magnetic field generated by the Earth, flattening it on the sunny side of the Earth facing the Sun and stretching it into a near-Earth magnetotail on the side facing away from the Sun [40]. As magnetic field lines pass through a distorted magnetic field, they break and reconnect. When the solar wind magnetic field and the Earth's magnetic field are tilted, the Earth's magnetic field will produce a deviation from the relative symmetry distribution between the north and south poles, which will also lead to a deviation of the north and south lights related to the symmetrical distribution. As a result, there are differences in the Earth's magnetic field between ionospheric conjugate points, as well as differences in solar particle exposure

between the two magnetic hemispheres. These differences lead to asymmetries in ionospheric convection, thermosphere wind, currents, magnetic field disturbances, ion outflow, electron density, and aurora emission [41]. To be precise, these asymmetries are deviations from their relative symmetrical distribution as shown in **Figure 1** by the ideal pattern.

2) Charged particles in the atmospheric inversion layer cause auroral noise. The inversion layer in the atmosphere is a perturbing environment that can cause many kinds of extreme weather, which concentrates pollutant materials and produces photochemical effects. The near-surface inversion layer in the troposphere is the environmental conditions that form polluted haze weather and ozone weather [42] [43]. In the thermosphere, atmospheric inversions are also formed on cold nights, resulting in changes of ion concentrations and abnormal ion movements there. As a result, when particles from the Sun hit the abnormally distributed and moving ions in the inversion layer, the charged particles collide and discharge at different angles, resulting in aurora noise. The inversion layer can have a significant impact on the optical phenomenon linked to the mechanism that modifies the symmetrical distribution of two auroras in an oval ring since the solar-charged particles could no longer entirely collide by a perpendicular angle with the ions. Thus, such inversion noise can cause the two oval ring auroras to deviate from the symmetrical distribution.

3) Effect of wave-particle interaction on the precipitation of energy particles. In the upper atmosphere, strong magnetosphere-ionosphere coupling plays an important role in wave-particle interactions [44]. Relativistic effects need to be considered during wave-particle interactions. Wave-particle interactions can cause relativistic electrons to simultaneously settle into isolated proton auroras. We believe that in wave-particle interactions, orthogonal collisions between them are still an important mechanism for electrons or protons to excite new energy. In the ionosphere, charged particles of the solar wind collide with oxygen and nitrogen atoms through wave-particle interactions. This release of collision energy allows people to observe colorful auroras. We speculate that only orthogonal collisions of solar wind particles with atmospheric particles are likely to yield higher anomalous energies and accelerate the velocity of electron groups during the last leg of the journey. The interaction of electrons and the Alfvén waves also plays a definitive role in the production of Earth's auroras [45] [46]. To generate energies of tens of keV and of relativistic electrons, only orthogonal interactions between electrons and Alvin waves are likely to have greater electron acceleration and auroras.

4) Possible formation causes of diffuse and discrete auroras. Auroras can be divided into diffuse and discrete (or pulsating) auroras [47]. They should be manifestation on the two magnetic conjugate regions involving complex interactions between ionospheric particles and energies. The difference between diffuse auroras and discrete auroras is the intensity of light, and its physical nature may be the difference in energy density as described from **Figure 1**. On the equator-

ward side of the dayside cusp there often appear diffuse auroras [48]. Thus, the diffuse aurora defines the extent of the auroral zone while the discrete auroras are sharply defined brightness features within the diffuse aurora. The intensity of solar activity and geomagnetic anomalies will not be exactly the same every time, so the collision angles and energy density of particle interactions or wave-particle interactions will be different. Therefore, the spatial and temporal distribution of auroras is associated with a complex space weather process, which has an intrinsic physical pattern and differences in the intensity, coverage and persistence of each process.

Acknowledgements

The author wishes to thank Dr. Du, J. for improving this paper's expression and the four anonymous reviewers for their constructive suggestions which have improved the paper. This work was supported by the innovative R&D project in Guangdong Province in China (2019ZT08G669).

Conflicts of Interest

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

References

- [1] Laroussi, M. and Akman, M.A. (2014) *The IEEE Transactions on Plasma Science*, **42**, 2662-2663. <https://doi.org/10.1109/TPS.2014.2326559>
- [2] Schlamminger, L. (1992) *The Journal of Atmospheric and Solar-Terrestrial Physics*, **54**, 989-993. [https://doi.org/10.1016/0021-9169\(92\)90065-S](https://doi.org/10.1016/0021-9169(92)90065-S)
- [3] Space Weather (2016) What Are Northern Lights? <http://www.swpc.noaa.gov/phenomena/aurora>
- [4] Cardenas, F.M., Sanchez, S.C. and Dominguez, S.V. (2016) *Advances in Space Research*, **57**, 257-267. <https://doi.org/10.1016/j.asr.2015.08.026>
- [5] Elphinstone, R.D., Murphree, J.S., Hearn, D.J., *et al.* (1995) *Journal of Geophysical Research: Space Physics*, **100**, 12075-12092. <https://doi.org/10.1029/95JA00326>
- [6] Elphinstone, R.D., Hearn, D.J., Cogger, L.L., *et al.* (1995) *Journal of Geophysical Research*, **100**, 12093-12102. <https://doi.org/10.1029/95JA00327>
- [7] Zhao, K., Jiang, Y., Men, K.P., *et al.* (2014) *Chinese Journal of Geophysics*, **57**, 3715-3728.
- [8] Viljanen, A. and Tanskanen, E. (2013) *Earth Planets and Space*, **65**, 45-49. <https://doi.org/10.5047/eps.2012.05.014>
- [9] Edemskiy, I.K. and Yasyukevich, Y.V. (2022) *Remote Sensing*, **14**, 5486. <https://doi.org/10.3390/rs14215486>
- [10] Hapgood, M.A. (2011) *Advances in Space Research*, **47**, 2059-2072. <https://doi.org/10.1016/j.asr.2010.02.007>
- [11] Tsurutani, B.T., Lakhina, G.S. and Hajra, R. (2020) *Nonlinear Processes in Geophysics*, **27**, 75-119. <https://doi.org/10.5194/npg-27-75-2020>
- [12] Pihlaja, P.M. (2012) *Acta Borealia*, **29**, 119-136. <https://doi.org/10.1080/08003831.2012.732281>

- [13] Johannesson, G.T. and Lund, K.A. (2017) *Annals of Tourism Research*, **63**, 183-190. <https://doi.org/10.1016/j.annals.2017.02.001>
- [14] Sauvaud, J.A., Barthe, H., Aoustin, C., Thocaven, J.J., Rouzaud, J., Penou, E., Popescu, D., Kovrazhkin, R.A. and Afanasiev, K.G. (1998) *Annales Geophysicae*, **16**, 1056-1069. <https://doi.org/10.1007/s00585-998-1056-z>
- [15] Prikryl, P., et al. (2013) *Annales Geophysicae*, **56**, R0216. <https://www.annalsofgeophysics.eu/index.php/annals/article/view/6227>
- [16] Zhang, Y.L. and Paxton, L.J. (2019) *The Journal of Atmospheric and Solar- Terrestrial Physics*, **184**, 1-4. <https://doi.org/10.1016/j.jastp.2019.01.005>
- [17] Mura, A., Adriani, A., Connerney, J.E.P., et al. (2018) *Science*, **361**, 774. <https://doi.org/10.1126/science.aat1450>
- [18] Belenkaya, E.S. (2020) *Solar System Research*, **54**, 85-95. <https://doi.org/10.1134/S003809462002001X>
- [19] Malinverno, A., Quigley, K.W., Staro, A. and Dymant, J. (2020) *Journal of Geophysical Research: Solid Earth*, **125**, e2020JB020034. <https://doi.org/10.1029/2020JB020034>
- [20] Rea, D.K. and Blakely, R.J. (1975) *Nature*, **255**, 126-128. <https://doi.org/10.1038/255126a0>
- [21] Fujiwara, T. and Fujimota, H. (1998) *Journal of Geophysical Research: Solid Earth*, **103**, 30335-30349. <https://doi.org/10.1029/1998JB900015>
- [22] Fuller, M. (1972) *Science*, **178**, 154-156. <https://doi.org/10.1126/science.178.4057.154>
- [23] Runcorn, S.K. (1983) *Nature*, **304**, 589-596. <https://doi.org/10.1038/304589a0>
- [24] Bumba, V. and Howard, R. (1965) *Science*, **149**, 1331-1337. <https://doi.org/10.1126/science.149.3690.1331>
- [25] Lewis, E.P. (1908) *Nature*, **78**, 569-570. <https://doi.org/10.1038/078569d0>
- [26] Cram, L.E. (1979) *Nature*, **282**, 133-134. <https://doi.org/10.1038/282133a0>
- [27] Livingston, W.C. (1982) *Nature*, **297**, 208-209. <https://doi.org/10.1038/297208a0>
- [28] Petrescu, F. and Petrescu, R. (2020) *Independent J. Management & Production*, **11**, 1835-1855. <https://doi.org/10.14807/ijmp.v11i3.930>
- [29] Baramsai, B., Benyo, T., Forsley, L. and Steinetz, B. (2022) *IEEE Spectrum*, **59**, 32-37. <https://doi.org/10.1109/MSPEC.2022.9729961>
- [30] Qian, W (2022) *Journal of High Energy Physics, Gravitation and Cosmology*, **8**, 184-194. <https://www.scirp.org/journal/jhepgc>
<https://doi.org/10.4236/jhepgc.2022.81014>
- [31] Valet, J.P., Fournier, A., Courtillot, V. and Herrero-Bervera, E. (2012) *Nature*, **490**, 89-93. <https://doi.org/10.1038/nature11491>
- [32] Qian, W. (2022b) *Journal of High Energy Physics, Gravitation and Cosmology*, **8**, 417-431. <https://doi.org/10.4236/jhepgc.2022.82032>
- [33] Workman, et al. (2022) *Progress of Theoretical and Experimental Physics*, **2022**, 083C01.
- [34] Ngwira, C.M., Habarulema, J., Astafyeva, E., et al. (2019) *Journal of Geophysical Research: Space Physics*, **124**, 7123-7139. <https://doi.org/10.1029/2018JA026172>
- [35] Han, D.S. (2021) *Frontiers in Astronomy and Space Sciences*, **8**, Article ID: 725677. <https://doi.org/10.3389/fspas.2021.725677>
- [36] Litvinenko, A.G. (2007) *Particle and Nuclei*, **38**, 204-231. <https://doi.org/10.1134/S1063779607020037>

- [37] Bruning, O., Burkhardt, H. and Myers, S. (2012) *Progress in Particle and Nuclear Phys.*, **67**, 705-735. <https://doi.org/10.1016/j.pnpnp.2012.03.001>
- [38] Normile, D. and Cho, A. (2019) *Science*, **363**, 911-912. <https://doi.org/10.1126/science.363.6430.911>
- [39] Qian, W. (2022) *Journal of Modern Physics*, **13**, 1440-1451. <https://doi.org/10.4236/jmp.2022.1311089>
- [40] Man, H.Y., Zhou, M., Deng, X.H., *et al.* (2018) *Geophysical Research Letters*, **45**, 8729-8737. <https://doi.org/10.1029/2018GL079778>
- [41] Ambrosino, D., Amata, E., Marcucci, M.F., *et al.* (2009) *Annales Geophysicae*, **27**, 2423-2438. <https://doi.org/10.5194/angeo-27-2423-2009>
- [42] Qian, W.H. and Huang, J. (2019) *Science of the Total Environment*, **647**, 878-887. <https://doi.org/10.1016/j.scitotenv.2018.07.408>
- [43] Qian, W.H., Xu, M. and Ai, Y. (2022) *Air Quality, Atmosphere & Health*, **15**, 1699-1711. <https://doi.org/10.1007/s11869-022-01208-w>
- [44] Ni, B.B. and Thorne, R.M. (2012) Recent Advances in Understanding the Diffuse Auroral Precipitation: The Role of Resonant Wave-Particle Interactions. In: Summers, D., Mann, I.R., Baker, D.N. and Schulz, M., Eds., *Dynamics of the Earth's Radiation Belts and Inner Magnetosphere*, American Geophysical Union, Washington DC, 291-302. <https://doi.org/10.1029/2012GM001337>
- [45] Watt, C.E.J. and Rankin, R. (2010) *Journal of Geophysical Research: Space Physics*, **115**, A07224. <https://doi.org/10.1029/2009JA015185>
- [46] Mottez, F. (2012) The Role Alfvén Waves in the Generation of Earth Polar Auroras. *Workshop on Waves and Instabilities in Space and Astrophysical Plasmas*, Vol. 1439, 3-25.
- [47] Nishimura, Y., Lessard, M.R., Katoh, Y., *et al.* (2020) *Space Science Reviews*, **216**, Article No. 4. <https://doi.org/10.1007/s11214-019-0629-3>
- [48] Taguchi, S., Hosokawa, K. and Ogawa, Y. (2019) *Journal of Geophysical Research: Space Physics*, **124**, 10778-10793. <https://doi.org/10.1029/2019JA026895>

The Speed of Light Is Not Constant in Basic Big Bang Theory

Jef Van Royen

Edegem, Belgium

Email: j.van.royen@scarlet.be

How to cite this paper: Van Royen, J. (2023) The Speed of Light Is Not Constant in Basic Big Bang Theory. *Journal of Modern Physics*, 14, 287-310.

<https://doi.org/10.4236/jmp.2023.143019>

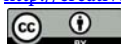
Received: November 15, 2022

Accepted: February 20, 2023

Published: February 23, 2023

Copyright © 2023 by author(s) and Scientific Research Publishing Inc. This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

Starting from the basic assumptions and equations of Big Bang theory, we present a simple mathematical proof that this theory implies a varying (decreasing) speed of light, contrary to what is generally accepted. We consider General Relativity, the first Friedmann equation and the Friedmann-Lemaître-Robertson-Walker (FLRW) metric for a Comoving Observer. It is shown explicitly that the Horizon and Flatness Problems are solved, taking away an important argument for the need of Cosmic Inflation. A decrease of 2.1 cm/s per year of the present-day speed of light is predicted. This is consistent with the observed acceleration of the expansion of the Universe, as determined from high-redshift supernova data. The calculation does not use any quantum processes, and no adjustable parameters or fine tuning are introduced. It is argued that more precise laboratory measurements of the present-day speed of light (and its evolution) should be carried out. Also it is argued that the combination of the FLRW metric and Einstein's field equations of General Relativity is inconsistent, because the FLRW metric implies a variable speed of light, and Einstein's field equations use a constant speed of light. If we accept standard Big Bang theory (and thus the combination of General Relativity and the FLRW metric), a variable speed of light must be allowed in the Friedmann equation, and therefore also, more generally, in Einstein's field equations of General Relativity. The explicit form of this time dependence will then be determined by the specific problem.

Keywords

General Relativity, Friedmann Equation, Big Bang, Cosmic Microwave Background, CMB, Varying Speed of Light, Flatness Problem, Horizon Problem, Cosmic Inflation

1. Introduction

At present, Big Bang theory is about one century old, and it is still our best

framework to describe the long term evolution of the Universe. It is considered the dominant theory to describe the evolution of our Universe (however dissident opinions exist). An important observation was the redshift of far-away galaxies, suggesting that we live in an expanding Universe. Furthermore, 3D-space seems to be homogeneous and isotropic. To describe these observations, Big Bang Cosmology (Λ -CDM) was developed, starting from General Relativity [1]. The calculations are based on the Friedmann-Lemaître-Robertson-Walker (FLRW) metric [2]-[8], which says that the scale of space itself increases uniformly, as a function of time. We call Big Bang the total process (continuing today) describing the expansion of space. The physical mechanisms that started this expansion, and the causes of Big Bang in general, are not understood. In Big Bang theory, galaxies are carried away from each other by the expanding space, and the longer the distance, the higher the recession speed. The divergent movement is only observed for very large distances. Nearby galaxies will attract each other gravitationally, and thus they will show motion not related to the universal expansion.

Big Bang Cosmology describes satisfactorily a number of experimental observations, such as the age of the Universe, the observed cosmological redshifts, the existence of a Cosmic Microwave Background (CMB), the homogeneity and isotropy of the Universe and, to a reasonable approximation, the abundance of primordial atomic nuclei. However, there are some important questions, such as the Horizon Problem and the Flatness Problem. In order to solve these difficulties, later theoretical work suggested some new physical mechanisms. At present the most popular of these proposals is Cosmic Inflation [9] [10] [11] [12]. Other proposals are a Bouncing Universe [13], and a Conformal Cyclic Cosmology (CCC) [14]. Also there is the suggestion of Varying Speed of Light (VSL) [15] [16]. Note that at present none of these ideas are generally accepted. They all generate a lot of discussions and new questions. They are additions to basic Big Bang theory that were added later, to fix problems that were not understood from the elementary calculation.

In the present calculation we will start from the standard assumptions and equations of basic Big Bang theory. We will then give a simple mathematical proof that the standard version of this theory implies a variable speed of light, contrary to what is generally accepted. We will show that this solves some of the basic problems of Big Bang Theory. Also we will point out that this has important implications for General Relativity as a whole.

2. General Relativity and the Expanding Universe: The FLRW Metric

The aim of Big Bang theory is to describe the evolution of the Universe as a function of time. To do this the path is studied of light reaching us from a very long time ago. The propagation of this light is influenced by the distribution of matter and other forms of energy in the Universe. Therefore we must start from General Relativity. We will have to solve a differential equation, and in order to obtain the proper solution of this equation we will have to introduce a number

of experimentally measured parameters.

In General Relativity the three spatial dimensions and time are taken together to describe one four-dimensional world, space-time. To calculate distances in this space-time we need to introduce a metric. For this purpose, the traditional approach is the Friedmann-Lemaître-Robertson-Walker (FLRW) metric [2]-[8].

The FLRW metric combines time and a homogeneous and isotropic 3D-space with uniform curvature. Light travels through space-time following a radial path along a null geodesic. In the FLRW metric we use the present day speed of light c_0 . A time-dependent scale factor $a(t)$ is introduced to allow for uniform expansion (or contraction) of 3D-space. The FLRW metric postulates that the motion of light through space-time is determined by the (zero) line-element:

$$ds^2 = 0 = c_0^2 dt^2 - a(t)^2 \frac{dr^2}{1 - kr^2} \quad (1)$$

Here $a(t)$ has the dimension of length and k is a dimensionless constant that describes the overall curvature of space. It can take the values 0, +1 or -1. These values correspond to a flat Euclidian space, a spherical space with constant positive curvature and a hyperbolic space with constant negative curvature respectively. Note that r is a (dimensionless) comoving radial coordinate. It doesn't change when the Universe expands.

From Equation (1) it follows, for non-zero $a(t)$:

$$\frac{a_0 c_0}{a(t)} dt = \frac{a_0 dr}{\sqrt{1 - kr^2}} \quad (2)$$

Here $a_0 = a(t_0)$, the value of the scale factor at present time.

The distance traveled by a light signal between time t_1 and t_0 , and between position r_1 and r_0 , can therefore be calculated as

$$\Delta s(t_1, t_0) = \int_{r_1}^{r_0} \frac{a_0 dr}{\sqrt{1 - kr^2}} = \int_{t_1}^{t_0} \frac{a_0 c_0}{a(t)} dt \quad (3)$$

Here proper distances are calculated. Obviously $\Delta s(t_1, t_0)$ is the distance between the starting and the end point of the light signal, measured by an observer at time $t = t_0$, using fixed (non-expanding) length units.

In order to obtain actual distances from Equation (3), the scale factor $a(t)$ must be determined from General Relativity. This is done by calculating $a(t)$ from the (first) Friedmann equation [2] [3]. This will be illustrated in Section 4.

Note that the FLRW metric is not part of General Relativity. It was added later, by imposing geometric properties on 3D-space (homogeneity and isotropy).

3. Frame of Reference

An important aspect of the General Relativity/FLRW calculation is the status of the observer. The Friedmann equation was developed for an observer on Earth. However every "local" motion of this observer has to be neglected. This means an observer moving only because he/she is carried away from distant galaxies by

the expansion of space, *i.e.* a so-called Comoving Observer.

Also note that it is generally accepted that at present our material world is not expanding (and has not expanded) with the Universe, due to the forces keeping material objects together. Therefore we will make all calculations from the point of view of a Comoving Observer, and we will use nonexpanding units along the spatial axes. These non-expanding length units should be determined by the length of (non-expanding) material objects. Note that this is also the frame of reference used by astronomers determining the distance of far-away galaxies, based on the “Cosmic Distance Ladder”.

Since 1983 the definition of the meter is based on a (constant) speed of light c_0 [17]. Therefore, in the present calculation we will return to the old definition of the meter, based on the size of material objects, in order to avoid confusion.

4. Calculation of $a(t)$ from General Relativity

The derivation of the FLRW metric is based on a “perfect liquid” model for the Universe, such that a (continuous) density and pressure can be defined for mass and other forms of energy. Furthermore the Universe is supposed to be homogeneous and isotropic. These assumptions are considered acceptable as a lowest order approximation. The variable t appearing in all expressions here is simply the time as measured on the clock of the Comoving Observer.

We will now recall the (first) Friedmann equation. This is a differential equation that was first derived by Alexander Friedmann [2] [3]. The equation has to be solved for the scale factor $a(t)$, describing the expansion of the Universe:

$$\dot{a}^2 - \frac{8\pi G}{3} \rho a^2 = -k c_0^2 \quad (4)$$

Here the dot indicates the derivative with respect to time. G is Newton’s gravitational constant and c_0 is the present day speed of light. The curvature parameter k is the same as in the FLRW metric, Equation (1). The time-dependent variable ρ describes the average density of matter and other forms of energy in the Universe. Essentially, the Friedmann equation expresses energy (density) conservation: kinetic energy density plus potential energy density plus energy density related to the overall curvature of space always equals zero.

The Friedmann equation, Equation (4), was obtained by substituting the FLRW metric, Equation (1), in the (0, 0) component of the standard Einstein field equations. Note that the speed of light occurs only in the rhs of the Friedmann equation, in a term proportional to the curvature factor k .

The solution of the Friedmann equation can be found in standard cosmology textbooks (see also [18], where the equation is solved, not only for a constant speed c_0 , but also for a time dependent speed of light $c(t)$ in the Friedmann equation). We now illustrate the most elementary solution of the Friedmann equation. This is the basis of standard Big Bang theory. We first introduce some variables and constants that make the calculations more transparent.

The (time dependent) Hubble variable is defined as:

$$H(t) = \frac{\dot{a}(t)}{a(t)} \quad (5)$$

At present time ($t = t_0$), this reduces to the Hubble constant H_0 :

$$H_0 = \frac{\dot{a}_0}{a_0} \quad (6)$$

The (time dependent) critical density ρ_c is defined as:

$$\rho_c = \frac{3H^2}{8\pi G} \quad (7)$$

ρ_c serves to make the distinction between a positively and a negatively curved geometry. When the total density ρ is exactly equal to the critical density ρ_c , the spatial geometry is flat. At present ($t = t_0$), the critical density reduces to

$$\rho_{c,0} = \frac{3H_0^2}{8\pi G} \quad (8)$$

In the Friedmann equation, the total density is written as the sum of three terms, describing radiation (ρ_R), matter (ρ_M) (normal and dark) and dark energy (ρ_Λ). The curvature term (ρ_K) is incorporated in the total density:

$$\rho = \rho_R + \rho_M + \rho_\Lambda + \rho_K \quad (9)$$

The variation of ρ_R , ρ_M , ρ_Λ , and ρ_K with $a(t)$ is then described as:

$$\rho_R = \rho_{R,0} \frac{a_0^4}{a^4} \quad (10)$$

$$\rho_M = \rho_{M,0} \frac{a_0^3}{a^3} \quad (11)$$

$$\rho_\Lambda = \rho_{\Lambda,0} \quad (12)$$

$$\rho_K = \rho_{K,0} \frac{a_0^2}{a^2} \quad (13)$$

with

$$\rho_{K,0} = -\rho_{c,0} \frac{kc_0^2}{H_0^2 a_0^2} \quad (14)$$

The Friedmann equation can then be rewritten as

$$H^2(t) = \frac{\dot{a}^2}{a^2} = \frac{H_0^2}{\rho_{c,0}} \left[\rho_{R,0} \frac{a_0^4}{a^4} + \rho_{M,0} \frac{a_0^3}{a^3} + \rho_{\Lambda,0} + \rho_{K,0} \frac{a_0^2}{a^2} \right] \quad (15)$$

Finally, we introduce the relative present-day densities:

$$\Omega_{R,0} = \frac{\rho_{R,0}}{\rho_{c,0}} \quad (16)$$

$$\Omega_{M,0} = \frac{\rho_{M,0}}{\rho_{c,0}} \quad (17)$$

$$\Omega_{\Lambda,0} = \frac{\rho_{\Lambda,0}}{\rho_{c,0}} \quad (18)$$

$$\Omega_{K,0} = \frac{\rho_{K,0}}{\rho_{c,0}} = -\frac{kc_0^2}{H_0^2 a_0^2} \quad (19)$$

The Friedmann equation can then be rewritten as:

$$H^2(t) = \frac{\dot{a}^2}{a^2} = H_0^2 \left[\Omega_{R,0} \frac{a_0^4}{a^4} + \Omega_{M,0} \frac{a_0^3}{a^3} + \Omega_{\Lambda,0} + \Omega_{K,0} \frac{a_0^2}{a^2} \right] \quad (20)$$

The standard method to solve the Friedmann equation from Equation (20) is then to calculate t as a function of a by integration (the normalized integration variable $a' = a/a_0$ has been introduced):

$$t(a) = \frac{1}{H_0} \int_0^{\frac{a}{a_0}} \frac{a' da'}{\sqrt{\Omega_{R,0} + \Omega_{M,0} a' + \Omega_{\Lambda,0} a'^4 + \Omega_{K,0} a'^2}} \quad (21)$$

The Friedmann equation, Equation (20), must hold at all times, including $t = t_0$:

$$H_0^2 = H_0^2 \left[\Omega_{R,0} + \Omega_{M,0} + \Omega_{\Lambda,0} + \Omega_{K,0} \right] \quad (22)$$

and thus

$$1 = \Omega_{R,0} + \Omega_{M,0} + \Omega_{\Lambda,0} - \frac{kc_0^2}{H_0^2 a_0^2} \quad (23)$$

This means that, at present, the total energy density, including the curvature of the FLRW Universe, must be equal to the critical density.

Recent experimental values for the parameters used in the present calculation can be taken from the Planck 2015 data [19]:

$$H_0 = 67.3 \text{ km} \cdot \text{s}^{-1} \cdot \text{Mpc}^{-1} \quad (24)$$

$$\Omega_{R,0} = 9.24 \times 10^{-5} \quad (25)$$

$$\Omega_{M,0} = 0.315 \quad (26)$$

$$\Omega_{\Lambda,0} = 0.685 \quad (27)$$

$$\Omega_{K,0} = 0 \quad (28)$$

Also note that the Hubble time t_{Hubble} is then found to be:

$$t_{\text{Hubble}} = \frac{1}{H_0} = 14.53 \times 10^9 \text{ years} \quad (29)$$

From Equations (19) and (28) it is obvious that, with the present-day experimental parameters, space must be almost perfectly flat: $\Omega_{K,0}$ must be equal to zero to a very good approximation, and the boundary condition Equation (23) reduces to

$$1 = \Omega_{R,0} + \Omega_{M,0} + \Omega_{\Lambda,0} \quad (30)$$

which is in good agreement with the experimental results.

We then obtain the solution of the Friedmann equation, for a flat geometry:

$$t(a) = \frac{1}{H_0} \int_0^{\frac{a}{a_0}} \frac{a' da'}{\sqrt{\Omega_{R,0} + \Omega_{M,0} a' + \Omega_{\Lambda,0} a'^4}} \quad (31)$$

where the experimental parameters $H_0, \Omega_{R,0}, \Omega_{M,0}$ and $\Omega_{\Lambda,0}$ are given by Equations (24)-(27). Note that $t(a)$ is a function of the ratio a/a_0 only. The absolute value of a is unimportant. Also, since $k=0$, the time function $t(a)$, and thus also the scale factor function $a(t)$, do not depend on c_0 .

For the time since the start of Big Bang, t_{BB} , one then finds, by numerical integration:

$$\begin{aligned} t_{BB} = t(a_0) = t_0 &= \frac{1}{H_0} \int_0^1 \frac{a' da'}{\sqrt{\Omega_{R,0} + \Omega_{M,0} a' + \Omega_{\Lambda,0} a'^4}} \\ &\cong \frac{0.9506}{H_0} = 0.9506 t_{Hubble} = 13.81 \times 10^9 \text{ years} \end{aligned} \quad (32)$$

As a result, we can rewrite Equation (31) as

$$\frac{t}{t_0} = \frac{1}{0.9506} \int_0^{\frac{a}{a_0}} \frac{a' da'}{\sqrt{\Omega_{R,0} + \Omega_{M,0} a' + \Omega_{\Lambda,0} a'^4}} \quad (33)$$

In Equation (33), time and length only appear in the ratios $t' = t/t_0$ and $a' = a/a_0$. From the start of Big Bang until today these relative values vary from 0 to 1. Therefore we will use a_0 and t_0 as units of length and time whenever we perform a numerical calculation.

Note that all these results are purely classical. Very close to the origin of Big Bang, distances between material objects become very small, and a more correct calculation should also take into account non-classical (quantum) processes, that are not included in General Relativity.

5. The Calculated Scale Factor Function $a(t)$ Must Be Independent of the Time t_0 When It Is Determined by the Comoving Observer

The calculation of the scale factor $a(t)$, as illustrated here, is based on experimental measurements made at time $t = t_0$, with scale factor a_0 . The resulting formulas contain experimental parameters (Hubble constant, densities, ...) that vary with time. It is obvious that, in order to be consistent, a theory that calculates the deformation of space-time must always obtain the same scale factor function $a(t)$, independent of the time of the observations. We will illustrate now that this is indeed the case for the standard Big Bang results. We will start from Equation (21), giving the function $t(a)$ as derived by an observer at time $t = t_0$, and we will transform this expression to find the result for $t(a)$, as derived by the same observer at a different time $t = t_1$.

First note that, from Equations (10)-(13):

$$\rho_{R,0} = \rho_{R,1} \frac{a_1^4}{a_0^4} \quad (34)$$

$$\rho_{M,0} = \rho_{M,1} \frac{a_1^3}{a_0^3} \quad (35)$$

$$\rho_{\Lambda,0} = \rho_{\Lambda,1} \quad (36)$$

$$\rho_{K,0} = \rho_{K,1} \frac{a_1^2}{a_0^2} \quad (37)$$

Also, from Equations (7) and (8):

$$\rho_{c,0} = \frac{3H_0^2}{8\pi G} = \frac{H_0^2}{H_1^2} \frac{3H_1^2}{8\pi G} = \frac{H_0^2}{H_1^2} \rho_{c,1} \quad (38)$$

Therefore, Equations (16)-(19) can be rewritten as:

$$\Omega_{R,0} = \frac{\rho_{R,0}}{\rho_{c,0}} = \frac{a_1^4}{a_0^4} \frac{H_1^2}{H_0^2} \frac{\rho_{R,1}}{\rho_{c,1}} = \frac{a_1^4}{a_0^4} \frac{H_1^2}{H_0^2} \Omega_{R,1} \quad (39)$$

$$\Omega_{M,0} = \frac{\rho_{M,0}}{\rho_{c,0}} = \frac{a_1^3}{a_0^3} \frac{H_1^2}{H_0^2} \frac{\rho_{M,1}}{\rho_{c,1}} = \frac{a_1^3}{a_0^3} \frac{H_1^2}{H_0^2} \Omega_{M,1} \quad (40)$$

$$\Omega_{\Lambda,0} = \frac{\rho_{\Lambda,0}}{\rho_{c,0}} = \frac{H_1^2}{H_0^2} \frac{\rho_{\Lambda,1}}{\rho_{c,1}} = \frac{H_1^2}{H_0^2} \Omega_{\Lambda,1} \quad (41)$$

$$\Omega_{K,0} = \frac{\rho_{K,0}}{\rho_{c,0}} = \frac{a_1^2}{a_0^2} \frac{H_1^2}{H_0^2} \frac{\rho_{K,1}}{\rho_{c,1}} = \frac{a_1^2}{a_0^2} \frac{H_1^2}{H_0^2} \Omega_{K,1} \quad (42)$$

From Equation (42) it follows that, if $\Omega_K = 0$ at time t_0 , it is zero at all other times.

The results of Equations (39)-(42) can also be formulated as:

$$\Omega_R(t) = \frac{a_0^4}{a^4} \frac{H_0^2}{H^2} \Omega_{R,0} \quad (43)$$

$$\Omega_M(t) = \frac{a_0^3}{a^3} \frac{H_0^2}{H^2} \Omega_{M,0} \quad (44)$$

$$\Omega_{\Lambda}(t) = \frac{H_0^2}{H^2} \Omega_{\Lambda,0} \quad (45)$$

$$\Omega_K(t) = \frac{a_0^2}{a^2} \frac{H_0^2}{H^2} \Omega_{K,0} \quad (46)$$

Also, from Equation (20) one obtains:

$$H_1^2 = H_0^2 \left[\Omega_{R,0} \frac{a_0^4}{a_1^4} + \Omega_{M,0} \frac{a_0^3}{a_1^3} + \Omega_{\Lambda,0} + \Omega_{K,0} \frac{a_0^2}{a_1^2} \right] \quad (47)$$

We can conclude that, as soon as the scale factor a_1 is known, all other parameters $H_1, \Omega_{R,1}, \Omega_{M,1}, \Omega_{\Lambda,1}$ and $\Omega_{K,1}$ can be deduced from their values at time $t = t_0$.

Substituting the expressions Equations (34)-(42), (47) in the solution of the Friedmann equation, Equation (21), we obtain:

$$t(a) = \frac{1}{H_1} \int_0^a \frac{a'' da''}{\sqrt{\Omega_{R,1} + \Omega_{M,1} a'' + \Omega_{\Lambda,1} a''^4 + \Omega_{K,1} a''^2}} \quad (48)$$

Here the new integration variable $a'' = \frac{a_0}{a_1} a'$ was introduced.

It is immediately clear that Equation (48) is the solution of the Friedmann equation, as derived by an observer at time $t = t_1$, with scale factor a_1 , Hubble

constant H_1 and radiation/mass/dark energy/curvature relative densities $\Omega_{R,1}$, $\Omega_{M,1}$, $\Omega_{\Lambda,1}$ and $\Omega_{K,1}$. We can conclude that the function $t(a)$ as obtained at time t_1 is identical to the function $t(a)$ as obtained at time t_0 .

Finally, substituting Equations (39)-(42) in Equation (47) we find immediately that

$$H_1^2 = H_1^2 (\Omega_{R,1} + \Omega_{M,1} + \Omega_{\Lambda,1} + \Omega_{K,1}) \quad (49)$$

and thus

$$\Omega_{R,1} + \Omega_{M,1} + \Omega_{\Lambda,1} + \Omega_{K,1} = 1 \quad (50)$$

Therefore, the total energy density, including the energy density related to the overall curvature, remains equal to the critical density at all times.

We can conclude that, in the FLRW Universe, the Comoving Observer always obtains the same time function $t(a)$, independent of the value of the scale factor $(a_0, a_1, a_2, \dots)$ at which the measurements are made. This is also true for the scale factor $a(t)$ since it is calculated as the inverse function of $t(a)$. Obviously this property is necessary for internal consistency. Otherwise observers at different epochs would obtain different evolution functions for the Universe. This result will be used explicitly in the following sections.

Note that this condition, that the scale factor function $a(t)$ must be independent of the time at which it is calculated by the Comoving Observer, is valid, not only for a flat 3D-space, but also for arbitrary curvature.

6. Time Dependence of the Speed of Light

We will now perform a simple Gedankenexperiment. Consider a Comoving Observer who can carry out measurements of the parameters relevant for cosmology, throughout the history of the Universe. After each measurement session the observer uses General Relativity and the FLRW metric to calculate the scale factor function, and the distance covered by a light signal. Note that an identical evolution function of the Universe must be obtained by the Comoving Observer at all times. At time $t = t_i$ the observer determines the speed of light c_i and the corresponding parameters H_i , $\Omega_{R,i}$, $\Omega_{M,i}$, $\Omega_{\Lambda,i}$ and $\Omega_{K,i}$. From these parameters the observer then calculates the universal scale factor function $a(t)$. Consider the case of a flat 3D-space at $t = t_0$, corresponding to the actual experimentally observed world in which we live. At time t_i the observer will then obtain the distance covered by a light signal since the start of Big Bang, as given by Equation (3):

$$\Delta s(0, t_i) = \int_0^{t_i} \frac{a_i c_i}{a(t)} dt = a(t_i) c(t_i) \int_0^{t_i} \frac{dt}{a(t)} \quad (51)$$

The observer can determine $\Delta s(0, t_i)$ with infinitesimal time intervals dt_i . The speed of light, as seen by the observer at $t = t_i$, can then be found as the derivative of $\Delta s(0, t_i)$ with respect to t_i . Note that $a(t)$ is independent of t_i , as discussed before: the derivative of $a(t)$ with respect to t_i is zero.

Taking the derivative of both sides of Equation (51) with respect to t_i we then find the speed of light $c(t_i)$ as obtained by the Comoving Observer at t_i :

$$c(t_i) = \frac{d\Delta s(0, t_i)}{dt_i} = c(t_i) + \frac{d[a(t_i)c(t_i)]}{dt_i} \int_0^{t_i} \frac{dt}{a(t)} \quad (52)$$

Therefore, since $\int_0^{t_i} \frac{dt}{a(t)}$ is a positive quantity, we obtain

$$\frac{d[a(t_i)c(t_i)]}{dt_i} = 0 \quad (53)$$

We can conclude that $a_i c_i$ is a constant quantity as a function of time.

This means that Equation (51) describes light moving at a speed of

$$c(t) = \frac{a_i c_i}{a(t)} = \frac{a_0 c_0}{a(t)} \quad (54)$$

Therefore, for a Comoving Observer, standard Big Bang theory describes light traveling at a variable speed $\frac{a_0 c_0}{a(t)}$, where $a_0 c_0$ is a universal constant.

We will now prove the same fact from internal consistency arguments.

First consider a Comoving Observer at time $t = t_0$. This observer makes all relevant experimental measurements (c_0 , H_0 , ...) and determines the scale factor function $a(t)$. The observer then calculates the distance covered by a light signal from time $t = 0$ to an arbitrary time $t = t_1$. The result is

$$\Delta s(0, t_1) = \int_0^{t_1} \frac{a_0 c_0}{a(t)} dt \quad (55)$$

However the same observer can also make all relevant measurements (c_1 , H_1 , ...) at time $t = t_1$ and then, using these results, calculate the distance covered by the light signal from time $t = 0$ to this time $t = t_1$. The result is

$$\Delta s(0, t_1) = \int_0^{t_1} \frac{a_1 c_1}{a(t)} dt \quad (56)$$

In order for the theory to be self-consistent, the results from Equation (55) and (56) must be identical. Therefore, again, we find $a_0 c_0 = a_1 c_1$. $a_i c_i$ is a universal constant. The actual value of $a_i c_i$ is not known, because only the relative value of the scale factor appears in the numerical formulas, not the absolute value.

From Equation (54), the variation of the speed of light for a Comoving Observer can now be calculated by taking the derivative of $c(t)$ with respect to time:

$$\frac{dc(t)}{dt} = -\frac{1}{a^2} \dot{a} a_0 c_0 = -\frac{\dot{a}}{a} \frac{a_0 c_0}{a} = -H(t) c(t) \quad (57)$$

Equations (54) and (57) have been suggested before [18] [20] [21] [22]. However, the calculation given above is a simple mathematical proof that, for a Comoving Observer, they follow directly from basic Big Bang theory.

As $a(t)$ increases with time, it is clear from Equation (54) that the speed of light is a decreasing function of time. At present ($t = t_0$), Equation (57) reduces

to

$$\frac{dc}{dt}(t_0) = -H_0 c_0 \cong -2.1 \text{ cm} \cdot \text{s}^{-1} \cdot \text{year}^{-1} \quad (58)$$

The present-day variation of $c(t)$ is extremely small. Therefore a constant speed of light, as postulated in Special Relativity, is a very good approximation for all “normal” experiments. However, for longer timespans, e.g. in cosmology, the variation of the speed of light becomes important.

For light starting not at $t = 0$, but at a later time t_1 , we can write:

$$\Delta s(t_1, t_0) = \int_0^{t_0} \frac{a_0 c_0 dt}{a(t)} - \int_0^{t_1} \frac{a_0 c_0 dt}{a(t)} = \int_{t_1}^{t_0} \frac{a_0 c_0 dt}{a(t)} \quad (59)$$

This means that a light signal starting at a later time must be described using the same function $a(t)$ as before. Therefore, for light generated today, $c(t)$ varies at the same rate as $c(t)$ for light originating from far away and long ago.

7. Analytical Approximations for the Scale Factor Function $a(t)$, with Error Smaller than 1% over the Complete Range $[0, 1]$. Speed of Light as a Function of Time

In order to calculate actual distances traveled by a light signal, we will now present sufficiently accurate approximations for $a(t)$. In this section all values of a and t will be expressed in normalized units a/a_0 and t/t_0 respectively.

The function $t(a)$, as given by Equation (33), can be considered in three different regimes, depending on the dominant term under the square root sign. At high values of t the dark energy contribution ($\Omega_{\Lambda,0}$) dominates. At intermediate t matter ($\Omega_{M,0}$) is more important. At early t the radiation contribution ($\Omega_{R,0}$) is dominant. These three regimes also appear in the scale factor function $a(t)$. Using these approximations for $a(t)$ we then calculate the speed of light as a function of time.

7.1. Approximation for $a(t)$: “Dark Energy and Matter” Regime

A useful expression for $a(t)$, for the regime where matter and dark energy dominate, can be found in the literature [23]. Translated to our units this approximation can be written as

$$a(t) = 0.7947t^{2/3} + 0.0623(e^{t/0.6785} - 1) \quad (60)$$

7.2. Approximations for $a(t)$: “Radiation and Matter” Regime

For low values of t and a , it is obvious from Equation (33) that the term proportional to $\Omega_{\Lambda,0}$ under the square root sign can be neglected. One then obtains

$$t(a) \cong \frac{1}{0.9506} \int_0^a \frac{a' da'}{\sqrt{\Omega_{R,0} + \Omega_{M,0} a'}} \quad (61)$$

This integral can be solved analytically. The result is

$$t(a) = \frac{2}{3} \frac{1}{0.9506\Omega_{M,0}^2} \left[(\Omega_{R,0} + \Omega_{M,0}a)^{3/2} - \Omega_{R,0}^{3/2} \right] - \frac{2\Omega_{R,0}}{0.9506\Omega_{M,0}^2} \left[(\Omega_{R,0} + \Omega_{M,0}a)^{1/2} - \Omega_{R,0}^{1/2} \right] \quad (62)$$

This can be rewritten as

$$(\Omega_{R,0} + \Omega_{M,0}a)^{3/2} - 3\Omega_{R,0}(\Omega_{R,0} + \Omega_{M,0}a)^{1/2} + 2\Omega_{R,0}^{3/2} - \frac{3}{2} \cdot 0.9506\Omega_{M,0}^2 t = 0 \quad (63)$$

This is a cubic equation for the unknown $(\Omega_{R,0} + \Omega_{M,0}a)^{1/2}$. The equation can be solved using standard methods. The discriminant of this equation is

$$D = \frac{3}{2} \times 0.9506\Omega_{M,0}^2 t \left(\frac{3}{8} \times 0.9506\Omega_{M,0}^2 t - \Omega_{R,0}^{3/2} \right) \quad (64)$$

Finally $a(t)$ can be calculated from the resulting solutions for $(\Omega_{R,0} + \Omega_{M,0}a)^{1/2}$.

Depending on the sign of the discriminant different expressions are obtained:

For $D > 0$ (i.e. for $t > t_D = \frac{8}{3} \frac{\Omega_{R,0}^{3/2}}{0.9506\Omega_{M,0}^2} \cong 2.51107 \times 10^{-5}$) one finds:

$$a(t) = -\frac{\Omega_{R,0}}{\Omega_{M,0}} + \frac{1}{\Omega_{M,0}} \left[\left(\frac{3}{4} \times 0.9506\Omega_{M,0}^2 t - \Omega_{R,0}^{3/2} + \left(\frac{9}{16} \times 0.9506^2 \Omega_{M,0}^4 t^2 - \frac{3}{2} \times 0.9506\Omega_{M,0}^2 \Omega_{R,0}^{3/2} t \right)^{1/2} \right)^{1/3} + \left(\frac{3}{4} \times 0.9506\Omega_{M,0}^2 t - \Omega_{R,0}^{3/2} - \left(\frac{9}{16} \times 0.9506^2 \Omega_{M,0}^4 t^2 - \frac{3}{2} \times 0.9506\Omega_{M,0}^2 \Omega_{R,0}^{3/2} t \right)^{1/2} \right)^{1/3} \right] \quad (65)$$

For $D < 0$ (i.e. for $t < t_D = \frac{8}{3} \frac{\Omega_{R,0}^{3/2}}{0.9506\Omega_{M,0}^2} \cong 2.51106 \times 10^{-5}$) the result is:

$$a(t) = -\frac{\Omega_{R,0}}{\Omega_{M,0}} + 4 \frac{\Omega_{R,0}}{\Omega_{M,0}} \cos^2 \left[\frac{1}{3} \arccos \left(\frac{3}{4} \frac{0.9506\Omega_{M,0}^2 t}{\Omega_{R,0}^{3/2}} - 1 \right) \right] \quad (66)$$

For $t < t_D$ we will use Equation (66). For $t_D < t < 0.333$ we will use Equation (65), and for $0.333 < t < 1$ we will use Equation (64). In this way, it is easy to show that the difference with the exact values of $a(t)$ is smaller than 1% over the complete time interval $[0, 1]$.

Note that time t_D is a purely mathematical quantity. It doesn't have any physical meaning.

7.3. Speed of Light as a Function of Time

Using the approximations for $a(t)$ given above, and using Equation (54), it is now straightforward to calculate the speed of light as a function of time throughout the history of the Universe. The result is shown in **Figure 1**.

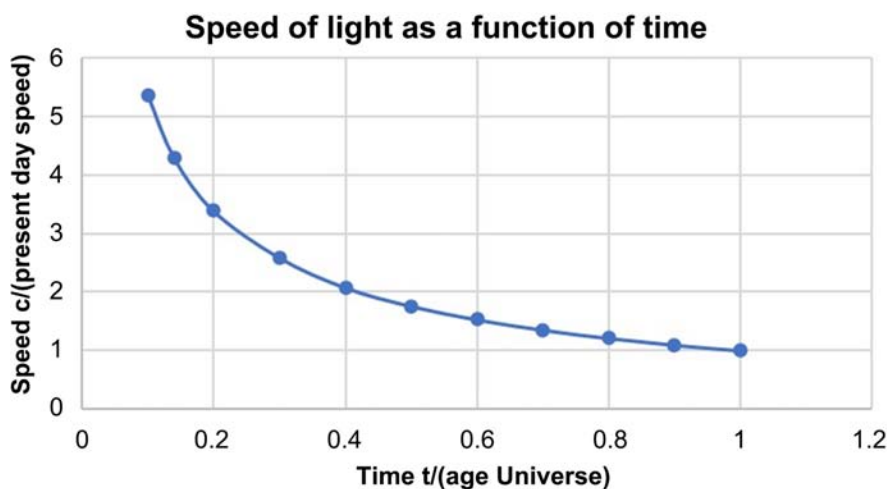


Figure 1. Speed of light (as observed by a Comoving Observer) as a function of time in the FLRW Universe. For $t = t_0$ (present time) the speed of light converges to c_0 . For very early times the speed increases strongly, and even diverges for $t = 0$.

8. The Horizon Problem

Our visual horizon is determined by the Cosmic Microwave Background. This is the oldest electromagnetic radiation we can observe. It originated in the first stages of Big Bang, when the temperature of the Universe had cooled down to about 3000 K. At that temperature electrons and nuclei (mainly protons) started to combine to form atoms. After this primordial deionization event, light was free to travel through space, without being scattered, along null geodetics, as described by General Relativity. Some of these photons can still be observed today, and they constitute the CMB. It was found experimentally that the present-day CMB is a nearly perfect example of Black Body radiation, with a temperature of approximately 2.725 K. As the original temperature of the CMB was about 3000 K, it is straightforward to calculate the redshift that produced the present day CMB from the original radiation:

$$z_{CMB} = \frac{3000}{2.725} \cong 1101 \quad (67)$$

Then the corresponding scale factor at CMB time t_{CMB} is (relative to a_0):

$$\frac{a_{CMB}}{a_0} = \frac{1}{1 + z_{CMB}} \cong \frac{1}{1102} \quad (68)$$

From Equation (33) we can calculate the time at which the CMB light started, by numerical integration, using the experimental parameters mentioned before:

$$t_{CMB} = \frac{t_0}{0.9506} \int_0^{\frac{a_{CMB}}{a_0}} \frac{a' da'}{\sqrt{\Omega_{R,0} + \Omega_{M,0} a' + \Omega_{\Lambda,0} a'^4}} \quad (69)$$

$$\cong 2.644 \times 10^{-5} t_0 \cong 365000 \text{ years}$$

It is found experimentally that the CMB background radiation is extremely isotropic. Variations in CMB temperature across the sky are smaller than 0.001 Kelvin. This means that all points of our Visible Universe must once have been

in thermal equilibrium with each other. If a constant speed of light c_0 is accepted, this is impossible to understand. This is called the Horizon Problem.

To solve this problem we will first calculate the distance covered by a photon from the start of CMB until today, using the approximations introduced in Section 7. From Equation (59) it then follows that this distance can be calculated as

$$\begin{aligned} d_{CMB}(t_0) &= \Delta s(t_{CMB}, t_0) = \int_{t_{CMB}}^{t_0} \frac{a_0 c_0}{a(t)} dt = c_0 t_0 \int_{t_{CMB}/t_0}^1 \frac{dt'}{a'(t')} \\ &= c_0 t_0 \int_{2.644 \times 10^{-5}}^1 \frac{dt'}{a'(t')} \cong 3.28 c_0 t_0 \end{aligned} \quad (70)$$

Therefore, from Equation (70), $3.28 c_0 t_0$ is, according to standard Big Bang theory, the distance traveled by a CMB photon observed today. Therefore this is the radius of our present-day Visible Universe (corresponding to $3.28 \times 13.81 \times 10^9 \cong 45.3$ billion light years, where a light year is defined as the distance covered by a light signal in one year at the “traditional” speed of 299,792,458 km/s).

Taking into account the expansion of space we can now calculate what was the value of the radius of our present-day Visible Universe at $t = t_{CMB}$. According to the definition of the scale factor this is simply

$$\begin{aligned} d_{CMB}(t_{CMB}) &= \frac{a_{CMB}}{a_0} d_{CMB}(t_0) = \frac{1}{1102} \times 3.28 c_0 t_0 \cong 0.00298 c_0 t_0 \\ &= 0.00298 \times 13.81 \times 10^9 \text{ light years} \cong 41.2 \times 10^6 \text{ light years} \end{aligned} \quad (71)$$

Before the time of CMB, at temperatures above 3000 K, photons could not travel freely through space, due to scattering by the primordial plasma. However we can calculate the cumulative distance traveled by a photon that is continually scattered from $t = 0$ to $t = t_{CMB}$. This distance is given by the sum of the individual steps between scattering events (obviously these steps are not in a straight line). The cumulative distance traveled by a photon between time $t = 0$ and $t = t_{CMB}$ can then be calculated numerically as

$$\Delta s(0, t_{CMB}) = \int_0^{t_{CMB}} \frac{a_0 c_0}{a(t)} dt = c_0 t_0 \int_0^{t_{CMB}/t_0} \frac{dt'}{a'(t')} \cong 0.0657 c_0 t_0 \quad (72)$$

We can now address the Horizon Problem. According to standard Big Bang theory, the ratio of the cumulative distance covered, since the start of Big Bang, by a photon at $t = t_{CMB}$, to the corresponding radius of our present-day Visible Universe at that time, is

$$\frac{\Delta s(0, t_{CMB})}{3.28 c_0 t_0 \times \frac{a_{CMB}}{a_0}} = \frac{0.0657 c_0 t_0}{0.00298 c_0 t_0} \cong 22.0 \quad (73)$$

This large ratio shows that CMB photons were indeed able to make thermal contact between opposite ends of our present-day Visible Universe, evening out temperature fluctuations in the CMB radiation. This is a purely classical process: no quantum gravity considerations are involved.

We will now study the evolution of the ratio (distance traveled by a light signal)/(radius of our Visible Universe) for times before t_{CMB} .

At very early times, the radiation contribution is dominant. Equation (33) reduces to:

$$\frac{t}{t_0} = \frac{1}{0.9506} \int_{a_0}^a \frac{a' da'}{\sqrt{\Omega_{R,0}}} = \frac{1}{0.9506 \times 2\sqrt{\Omega_{R,0}}} \left(\frac{a}{a_0} \right)^2 \quad (74)$$

and one finds

$$\frac{a}{a_0} = 2^{\frac{1}{2}} 0.9506^{\frac{1}{2}} \Omega_{R,0}^{\frac{1}{4}} \left(\frac{t}{t_0} \right)^{\frac{1}{2}} \quad (75)$$

Therefore, the cumulative distance traveled by a light signal is:

$$\Delta s(0, t) = \int_0^t \frac{a_0 c_0}{a(t)} dt = c_0 t_0 \int_0^{\frac{t}{t_0}} \frac{dt'}{a'(t')} = \frac{2^{\frac{1}{2}} c_0 t_0}{0.9506^{\frac{1}{2}} \Omega_{R,0}^{\frac{1}{4}}} \left(\frac{t}{t_0} \right)^{\frac{1}{2}} \quad (76)$$

From Equations (75) and (76) it follows that, in the zero time limit, the ratio (distance traveled by light)/(radius of our Visible Universe) will tend to a constant value:

$$\frac{\Delta s(0, t)}{3.28 c_0 t_0 \times \frac{a}{a_0}} = \frac{1}{3.28 \times 0.9506 \Omega_{R,0}^{\frac{1}{2}}} \cong 33.4 \quad (77)$$

The zero time limit of this ratio is clearly higher than the t_{CMB} value. We will now study the evolution of this ratio in the time interval $[0, t_{CMB}]$. Using the results of Section 7, the cumulative distance traveled by a light signal from time zero to time t can be calculated numerically as

$$\Delta s(0, t) = c_0 t_0 \int_0^{\frac{t}{t_0}} \frac{dt'}{a'(t')} \quad (78)$$

The results for the ratio of this distance to the corresponding radius of our present-day visible Universe are shown in **Figure 2**. It can be seen that the ratio decreases monotonously, from 33.4 to 22.0, during the time interval from the start of Big Bang to t_{CMB} .

We can conclude that, according to basic Big Bang theory, photons were able to make thermal contact inside our present Visible Universe during the complete time interval before the deionization event at t_{CMB} . A decreasing speed of light gives a simple intuitive explanation of this phenomenon: the light cone “opens up” at early times: photons can cover very large distances due to their very large speeds. The decreasing ratio in **Figure 2** suggests that the quality of the thermal contact in the plasma decreases as a function of time. This may explain the appearance of (small) fluctuations in the CMB temperature.

We have used here the concept of varying speed of light to explain the Horizon Problem. However all formulas and calculations used are standard Big Bang expressions. Therefore there is no Horizon Problem in standard Big Bang theory

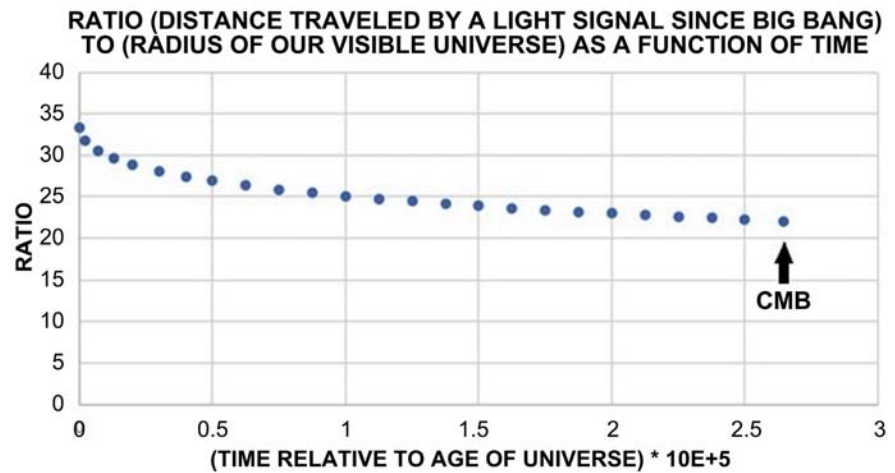


Figure 2. Ratio of the (cumulative) distance traveled by a light signal to the radius of our present-day Visible Universe for the time interval from zero to t_{CMB} . Ratios from 33.4 to 22.0 explain the smoothness of the CMB horizon.

to start with: the formulas show that thermal contact, producing a very smooth CMB, was possible in the very early Universe because of the very high distances covered by light signals at early times. Note that a variable speed of light makes the apparent contradictions intuitively acceptable.

9. For Non-Zero Curvature, the Combination of the FLRW Metric and the Friedmann Equation (and More Generally the Einstein Field Equations) Is Internally Inconsistent

The Friedmann equation, Equation (4), explicitly contains a constant speed of light c_0 . However, as shown in the previous sections, the FLRW metric implies a variable speed of light $c(t) = \frac{a_0 c_0}{a(t)}$. Obviously, this situation is inconsistent. Indeed, if the reasoning leading to a variable speed of light is correct, a Comoving Observer at time t_1 should measure a speed of light $c_1 = \frac{a_0 c_0}{a_1}$, and thus should

start from a Friedmann equation containing a speed of light c_1 instead of c_0 . In spite of this inconsistency, the combination of General Relativity and the FLRW metric (*i.e.* standard Big Bang theory) leads to remarkably good agreement with experimental results. To understand this we should note that the inconsistency noted here does not lead to problems for standard Big Bang theory immediately, because, as we know from experiment, we live in a flat 3D-space ($k = 0$) and therefore the term proportional to c_0 in the Friedmann equation is zero. Thus, as shown in the previous sections, the scale factor function $a(t)$ does not depend on the speed of light, and using the experimentally measured values (c_1 , H_1 , $\Omega_{R,1}$, $\Omega_{M,1}$, $\Omega_{\Lambda,1}$, $\Omega_{K,1}$) will automatically lead to the same (correct) solution $a(t)$ of the Friedmann equation as for the $t = t_0$ case. Problems only arise when we ask questions such as “How can the distance covered by a light signal be larger than $c_0 t_0$?” (here the variation of the speed of

light is important). The Horizon Problem belongs to this category.

However, for a curved 3D-space, ($k \neq 0$) there is a more direct problem: the standard Friedmann equation, Equation (4), uses the speed of light c_0 at all times, and therefore contradicts the requirements of the FLRW metric. This will lead to problems. A Comoving Observer at time $t_1 \neq t_0$, using the Friedmann equation containing c_0 , and not c_1 , will obtain an incorrect scale factor function. We can conclude that the evolution of the Universe, *i.e.* the function $a(t)$, as obtained for $k \neq 0$ by an observer at $t_1 \neq t_0$ using a constant speed of light c_0 , will be incorrect. The Flatness Problem belongs to this category.

The obvious solution for this problem is to modify the Friedmann equation in such a way that c_0 is replaced by a variable speed $c(t) = \frac{a_0 c_0}{a(t)}$, as required by

the FLRW metric. This was done in Ref. [18]. In the following section we will show that this procedure solves the Flatness Problem.

Also note that the Friedmann equation follows from the standard Einstein field equations of General Relativity. Therefore, if we accept a variable speed of light in the Friedmann equation, the speed of light must be allowed to be variable in the Einstein field equations in general as well.

10. The Flatness Problem

Experimentally, the curvature term $\Omega_{K,0}$ today is found to be very close to zero. This is very unexpected in standard Big Bang theory. First, we will now describe the Flatness Problem in more detail, and then we will show how the variable speed of light $c(t) = \frac{a_0 c_0}{a(t)}$ offers a simple and straightforward solution.

10.1. Flatness Problem: Traditional Big Bang Theory

In standard non-zero flatness ($k \neq 0$) Big Bang theory, the evolution of the flatness parameter $\Omega_K(t)$ can be obtained from Equation (46):

$$\Omega_K(t) = \frac{a_0^2}{a(t)^2} \frac{H_0^2}{H(t)^2} \Omega_{K,0} \quad (79)$$

The time derivative of $\Omega_K(t)$ can then be written as:

$$\frac{d\Omega_K}{dt} = a_0^2 H_0^2 \Omega_{K,0} \left[\frac{-2\dot{H}}{a^2 H^3} - \frac{2\dot{a}}{a^3 H^2} \right] = -2\Omega_K \left[\frac{\dot{H}}{H} + \frac{\dot{a}}{a} \right] = -2\Omega_K \left[\frac{\dot{H}}{H} + H \right] \quad (80)$$

Now, taking the time derivative of the definition $H = \frac{\dot{a}}{a}$ it is easy to show that

$$\frac{\dot{H}}{H} + H = \frac{\ddot{a}}{aH} \quad (81)$$

Also, from the Friedmann equation, Equation (4), we know that

$$\dot{a}^2 + kc_0^2 = \frac{8\pi G}{3} a^2 \left[\rho_{R,0} \frac{a_0^4}{a^4} + \rho_{M,0} \frac{a_0^3}{a^3} + \rho_{\Lambda,0} \right] \quad (82)$$

Taking the time derivative of both sides of Equation (82) we then find:

$$\frac{\ddot{a}}{aH} = -\frac{4\pi G}{3H} [2\rho_R + \rho_M - 2\rho_\Lambda] = -\frac{H}{2} [2\Omega_R + \Omega_M - 2\Omega_\Lambda] \quad (83)$$

Combining Equations (80), (81) and (83) we obtain the time derivative of Ω_K :

$$\frac{d}{dt}\Omega_K(t) = \Omega_K H [2\Omega_R + \Omega_M - 2\Omega_\Lambda] \quad (84)$$

It is immediately clear that $\Omega_K(t) = 0$ is a solution of this differential equation. This is consistent with the experimental value $\Omega_{K,0} = 0$. However, when we focus on early times there is a problem. At early times the radiation and mass contributions Ω_R and Ω_M are dominant. The dark energy term Ω_Λ is negligible. Therefore the numerical factor $(2\Omega_R + \Omega_M - 2\Omega_\Lambda)$ must be positive at these times. Therefore it follows from Equation (84) that even a very small deviation of Ω_K from zero (positive or negative) will grow (exponentially) fast. Small deviations from zero are supposed to occur, e.g. due to quantum processes. Therefore the $\Omega_K(t) = 0$ solution is unstable. It can easily be shown that the value of $\Omega_K(t)$ at early times must have been unrealistically small in order to produce the very low value we observe today. This is known as the Flatness Problem.

10.2. Flatness Problem Solution: Big Bang Theory with a Modified Friedmann Equation Using a Time Dependent Speed of Light

As explained in Section 9, Big Bang theory can be made self-consistent by using

$c(t) = \frac{a_0 c_0}{a(t)}$ instead of c_0 in the Friedmann equation. The equation can then

be solved using the same methods as before. The modified Friedmann equation is:

$$\dot{a}^2 - \frac{8\pi G}{3} \rho a^2 = -k \frac{a_0^2 c_0^2}{a^2} \quad (85)$$

The flatness parameter $\Omega_K(t)$ can then be written as

$$\Omega_K(t) = \frac{a_0^4}{a(t)^4} \frac{H_0^2}{H(t)^2} \Omega_{K,0} \quad (86)$$

The time derivative of $\Omega_K(t)$ then is:

$$\frac{d\Omega_K}{dt} = a_0^2 H_0^2 \Omega_{K,0} \left[\frac{-2\dot{H}}{a^4 H^3} - \frac{4\dot{a}}{a^5 H^2} \right] = -2\Omega_K \left[\frac{\dot{H}}{H} + 2\frac{\dot{a}}{a} \right] = -2\Omega_K \left[\frac{\dot{H}}{H} + 2H \right] \quad (87)$$

From the modified Friedmann equation, Equation (85), it follows that

$$\dot{a}^2 a^2 + k a_0^2 c_0^2 = \frac{8\pi G}{3} a^4 \left[\rho_{R,0} \frac{a_0^4}{a^4} + \rho_{M,0} \frac{a_0^3}{a^3} + \rho_{\Lambda,0} \right] \quad (88)$$

Taking the time derivative of both sides of Equation (88) we then find:

$$\frac{\ddot{a}}{aH} + H = \frac{4\pi G}{3H} [\rho_M + 4\rho_\Lambda] = \frac{1}{2} H [\Omega_M + 4\Omega_\Lambda] \quad (89)$$

Combining Equations (81), (87) and (89) we find the new differential equation

for $\Omega_K(t)$:

$$\frac{d}{dt}\Omega_K(t) = -\Omega_K H[\Omega_M + \Omega_\Lambda] \quad (90)$$

This equation is similar to Equation (84), however now the coefficient of Ω_K is always negative. At early times an arbitrary non-zero curvature will decrease exponentially. Space will inevitably evolve to zero curvature. Due to the decreasing speed of light the Flatness Problem has disappeared. Basic Big Bang theory predicts a flat Universe.

11. Discussion and Conclusions. Special Relativity, General Relativity and the Constancy of the Speed of Light

At present, the speed of light is generally considered to be a universal constant. Light rays can change direction, but every observer will always measure the same local speed. This is indeed the basic postulate of Special Relativity [24]. The predictions of Special Relativity have been verified multiple times to a very high degree of accuracy and therefore any claim of a variation of the speed of light can only be accepted if it is confirmed by irrefutable experimental results.

Note that in this paper, to avoid confusion, we define the speed of light from measuring the time it takes for a light signal to cover the length of a material object. This is different from the “constant speed” definition of 1983 [17].

In Special Relativity, in the absence of gravitation, the speed of light is clearly constant. In the presence of gravitation the situation is more unclear. The generally accepted opinion is that the speed of light is constant in that case as well, however dissident ideas exist. Einstein himself suggested a variable speed of light [25], before producing the standard equations of General Relativity. Einstein’s variable speed of light calculation for bending of starlight by the Sun contained an error, but an improved version of this calculation, leading to the correct result, was later given by Dicke [26].

As shown in Section 9, the Friedmann equation and the FLRW metric are essentially inconsistent, because the first uses a constant speed of light c_0 , and the latter implies a variable speed of light $c(t) = \frac{a_0 c_0}{a(t)}$, for a Comoving Observer. In

the case of a flat space (our experimentally observed situation) this only produces problems when we focus explicitly on the variation of the speed of light (the Horizon Problem). In the case of non-zero curvature, the inconsistency leads to problems directly (the Flatness Problem). Introducing the correct variable speed of light in the Friedmann equation solves the Horizon and Flatness Problems immediately, taking away the need for ad hoc additions to Big Bang theory, such as Cosmic Inflation.

The problem is more general, however. The Friedmann equation was derived from Einstein’s field equations of General Relativity. Therefore, if we accept the introduction of the FLRW metric as a legitimate procedure, and distances are expressed in units based on non-expanding material objects, we now have an

example where it is necessary to introduce a varying speed of light in the Einstein field equations. This suggests a modification of the Einstein field equations in general: they should allow for a variable speed of light. The explicit form of the time dependence of the speed of light (and possibly the spatial dependence as well) will then be determined by the specific problem.

Experimental observations of the speed of light have been made as a function of time, at different locations, and fluctuating results have been obtained [27]. It has been suggested that these fluctuations are related to (tidal) variations of the local gravitational acceleration g . If such effects are real indeed, it is obvious that they are small variations compared to the large scale, long term evolution of the speed of light described here. The FLRW metric starts from a homogeneous Universe, and thus ignores perturbations due to local mass concentrations.

We can conclude that, to verify the decrease of the speed of light predicted here, new, more precise, measurements of the speed of light, and its evolution, are needed (the latest experimental measurements, in the 1970s, had an accuracy of approximately 1 m/s). Also it would be interesting to perform such measurements in gravitational environments different from the surface of the Earth. If the speed of light is found to be invariable in deep intergalactic space, we can conclude that the combination of General Relativity and the FLRW metric does not give an acceptable description of our Universe.

The concepts “expanding space” and “variable speed of light” are complementary. This, then, seems to be the physical meaning of the FLRW metric: the expansion of the Universe, and the simultaneous decrease of the speed of light for a Comoving Observer. This explains in a simple way how the distance covered by a light signal since the start of Big Bang can be much larger than $c_0 t_0$, as expected classically from a constant speed c_0 .

A natural question is: “Is indeed the speed of light time dependent, or is what we observe just a consequence of gravitational time dilation?” It should be noted that a Comoving Observer doesn’t have an objective method to judge his cosmic flow of time. Therefore it seems reasonable to define the speed of light from the time it takes a light signal to cover the length (supposedly constant) of a material object. Future developments of physics will tell us whether this is indeed the best way to describe the reality of our Universe.

Note that, if we write the Planck radiation formula in the frequency form, the speed of light only appears as a multiplicative factor $\frac{1}{c^2}$. Therefore, when the speed of light is altered, the general form of the spectrum remains unchanged, and we still recognize it as blackbody radiation (albeit with a modified temperature, due to the expansion of space (and wavelength)).

The present results can be interpreted as showing that the speed of light, as measured by a Comoving Observer, is caused by the collective mass of the Universe. This is essentially Mach’s Principle [28]. Also note that in this view our present-day speed of light has no special meaning whatsoever. It just happens to be, by chance, the speed of light at the time we live. This also means that the

speed of light is a variable property of the vacuum, changing as a function of gravitation. Note that this has also consequences for notions such as Planck length and Planck time, since the speed of light appears in their definition.

Observations of high-redshift supernova data have shown that, at present, the expansion of the Universe is accelerating [29] [30]. Interestingly, it was recently noted that this observed accelerating expansion of the Universe can be simply described by a phenomenological decrease of the speed of light by 2.2 cm/s per year [31]. This confirms the present calculation.

Additional claims have been made about experimental observations of a systematic decrease of the speed of light [32]. An interesting case is the Lunar Laser Ranging experiment (LLR). During the Apollo and Lunokhod missions, in the 1970s, mirrors were installed on the Moon, and these are used to determine very precisely the distance between Earth and Moon, by measuring the round trip time of laser pulses. In this way, a systematic increase of the Earth-Moon orbit semimajor axis of 3.8 cm per year has been observed [33]. Also it is straightforward to calculate that a decrease of the speed of light of 2.1 cm/s per year results in an apparent increase of the Earth-Moon distance of 2.7 cm per year. However, this distance is also influenced by tidal effects and other geophysical processes. Another observation is the Pioneer Anomaly, related to radio data probing the position of distant spacecrafts. All these observations, as described in [32] and [33], might suggest a speed of light decrease of the order of 2 cm/s per year, however the results are insufficient to make a final conclusion.

It might be argued that the present results contradict Special Relativity. This is correct indeed. Lorentz invariance is broken because of the transition from a non-expanding to an expanding frame of reference. In the present formalism, $a_i c_i$ is a universal constant. In Special Relativity, where a_i is a constant, this reduces to the standard result that c_i is a universal constant. Also note that, in the calculation presented here, the speed of light remains the maximum speed of communication at any time t . Therefore the speed of light remains the speed of causality, which is an essential aspect of Special Relativity.

Obviously General Relativity is a classical, non-quantum theory, producing divergences for time $t = 0$. These divergences should disappear in a quantum version of General Relativity. However, today we do not have such a quantum theory of gravitation yet.

In “normal” experiments, times are short, and it is perfectly acceptable to use a constant speed of light. For longer times, however, such as in cosmology and astronomy, a varying speed of light will lead to new physics. An obvious example is the solution of the Horizon Problem. This is essentially a consequence of the fact that the light cone is no longer made up of straight lines: it “opens up” at times closer to the start of Big Bang. Similar remarks can be made about the Flatness Problem. As we have shown, basic Big Bang theory with a decreasing speed of light predicts a flat Universe. Note that it is no longer necessary to make claims of extreme fine tuning to explain the observed flatness. On the contrary,

basic Big Bang theory makes this flatness mandatory.

Also note that the present results imply that the most important motivations for the development of Cosmic Inflation theory no longer exist. To explain the isotropy of the CMB background and the Flatness of 3D-space we do not need to start from quantum processes, they are natural consequences of classical physical mechanisms.

Recent observations suggest that the speed of light is equal to the speed of gravitational waves to a very good approximation [34]. Therefore gravitational interaction is expected to show the same time dependence as the speed of light. This might be important for our interpretation of the results of gravity, especially at very large distances.

Also the Hubble Tension, *i.e.* the conflicting results for measurements of the present day Hubble constant, as determined from “recent” and “ancient” phenomena, might be related to this variation of $c(t)$. According to the present calculation, the speed of light was extremely high, and rapidly decreasing, at the time when the CMB background radiation was released. It should be investigated how this affects the baryon acoustic oscillations of the primordial plasma, and thus the temperature variations we observe today in the CMB background. If the present calculation is correct, it is clear that standard present day physics cannot be used to describe the situation at time $t = t_{\text{CMB}}$.

A very common remark is that in our present situation, on earth, we are not subjected to an FLRW metric, but to a Schwarzschild metric, due to the earth’s gravitational field. However note that the earth’s gravitational field does not make the gravitational influence of the rest of the Universe disappear. We are subjected to a superposition of local and distant fields. As the variation of the influence of faraway matter (and other forms of energy) is very slow, compared to local effects, these two contributions are decoupled in a natural way. It is reasonable to conclude that our present-day speed of light c_0 is very well approximated by the varying speed of light in intergalactic space (where the FLRW metric is supposed to be valid). Big Bang theory provides a general background, on which local phenomena are superimposed. Also, note that the Earth, the Solar System and the Milky Way are moving through intergalactic space, through regions that are supposed to have been modified by cosmic expansion before.

Finally, a variable speed of light suggests that other physical “constants” might be variable as well. This is certainly the case for ε_0 and/or μ_0 , the permittivity and permeability of vacuum, since $c_0 = 1/\sqrt{\varepsilon_0\mu_0}$. Obviously, such variations would seriously complicate the interpretation of experimental results. In this context it is interesting to note that recently it has been argued [35] that physics can be described starting from a Unified Energetic Tension Field based on (ε_0, μ_0) of free space.

Acknowledgements

Thanks are due to Prof. Dirk Van Dijck (University of Antwerp, Belgium) for valuable suggestions about how to publish this paper.

Conflicts of Interest

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

References

- [1] Einstein, A. (1916) *Annalen der Physik*, **49**, 769-822.
<https://doi.org/10.1002/andp.19163540702>
- [2] Friedman, A. (1922) *Zeitschrift für Physik*, **10**, 377-386.
<https://doi.org/10.1007/BF01332580>
- [3] Friedmann, A. (1924) *Zeitschrift für Physik*, **21**, 326-332.
<https://doi.org/10.1007/BF01328280>
- [4] Lemaitre, G. (1927) *Annales de la Société scientifique de Bruxelles A*, **47**, 49-56.
- [5] Robertson, H.P. (1935) *Astrophysical Journal*, **82**, 284-301.
<https://doi.org/10.1086/143681>
- [6] Robertson, H.P. (1936) *Astrophysical Journal*, **83**, 187-201.
<https://doi.org/10.1086/143716>
- [7] Robertson, H.P. (1936) *Astrophysical Journal*, **83**, 257-271.
<https://doi.org/10.1086/143726>
- [8] Walker, A.G. (1937) *Proceedings of the London Mathematical Society*, **S2-42**, 90-127.
<https://doi.org/10.1112/plms/s2-42.1.90>
- [9] Guth, A.H. (1981) *Physical Review D*, **23**, 347.
<https://doi.org/10.1103/PhysRevD.23.347>
- [10] Linde, A.D. (1982) *Physics Letters B*, **108**, 389.
[https://doi.org/10.1016/0370-2693\(82\)91219-9](https://doi.org/10.1016/0370-2693(82)91219-9)
- [11] Linde, A.D. (1983) *Physics Letters B*, **129**, 177.
[https://doi.org/10.1016/0370-2693\(83\)90837-7](https://doi.org/10.1016/0370-2693(83)90837-7)
- [12] Albrecht, A. and Steinhardt, P. (1982) *Physical Review Letters*, **48**, 1220.
<https://doi.org/10.1103/PhysRevLett.48.1220>
- [13] Steinhardt, P.J. and Turok, N. (2005) *New Astronomy Reviews*, **49**, 43-57.
<https://doi.org/10.1016/j.newar.2005.01.003>
- [14] Penrose, R. (2012) *AIP Conference Proceedings*, **1446**, 233.
<https://doi.org/10.1063/1.4727997>
- [15] Moffat, J.W. (1993) *International Journal of Modern Physics D*, **2**, 351-366.
<https://doi.org/10.1142/S0218271893000246>
- [16] Albrecht, A. and Magueijo, J. (1999) *Physical Review D*, **59**, Article ID: 043516.
<https://doi.org/10.1103/PhysRevD.59.043516>
- [17] Petley, B.W. (1983) *Nature*, **303**, 373-376. <https://doi.org/10.1038/303373a0>
- [18] Van Royen, J. (2021) *Journal of Cosmology*, **27**, 15389-15406.
<https://www.researchgate.net/publication/355197312>
- [19] Ade, P.A.R., et al. (2016) *Astronomy & Astrophysics*, **594**, A13.
<https://arxiv.org/abs/1502.01589>
- [20] Pipino, G. (2019) *Journal of High Energy Physics, Gravitation and Cosmology*, **5**, 395-411. <https://doi.org/10.4236/jhepgc.2019.52022>
- [21] Pipino, G. (2021) *Journal of High Energy Physics, Gravitation and Cosmology*, **7**, 742-760. <https://doi.org/10.4236/jhepgc.2021.72043>

- [22] Smeulders, P. (2012) *Journal of Modern Physics*, **3**, 345-349.
<https://doi.org/10.4236/jmp.2012.34047>
- [23] A Useful Approximation for $a(t)$, Valid from the Start of the Matter Dominated Era until Today. <https://physics.stackexchange.com/questions/136056>
- [24] Einstein, A. (1905) *Annalen der Physik*, **17**, 891-921.
<https://doi.org/10.1002/andp.19053221004>
- [25] Einstein, A. (1911) *Annalen der Physik*, **35**, 898-908.
<https://doi.org/10.1002/andp.19113401005>
- [26] Dicke, R.H. (1957) *Reviews of Modern Physics*, **29**, 363-376.
<https://doi.org/10.1103/RevModPhys.29.363>
- [27] Ramdani, F. (2019) *International Journal of Astronomy and Astrophysics*, **9**, 265-273.
<https://doi.org/10.4236/ijaa.2019.93019>
- [28] Mach, E. (1897) *Die Mechanik in ihrer Entwicklung—historisch-kritisch dargestellt*.
- [29] Riess, A.G., *et al.* (1998) *The Astronomical Journal*, **116**, 1009-1038.
<https://doi.org/10.1086/300499>
- [30] Perlmutter, S., *et al.* (1999) *The Astrophysical Journal*, **517**, 565-586.
<https://doi.org/10.1086/307221>
- [31] Sanejouand, Y.-H. (2005). <https://arxiv.org/abs/astro-ph/0509582>
- [32] Sanejouand, Y.-H. (2009) *EPL (Europhysics Letters)*, **88**, 59002.
<https://doi.org/10.1209/0295-5075/88/59002>
- [33] Williams, J.G., *et al.* (2014) *Planetary Science*, **3**, Article No. 2.
<https://doi.org/10.1186/s13535-014-0002-5>
- [34] Abbott, B.P., *et al.* (2017) *The Astrophysical Journal Letters*, **848**, L13.
<https://arXiv.org/abs/1710.05834>
- [35] Roychoudhuri, C. (2021) *Journal of Modern Physics*, **12**, 671-699.
<https://doi.org/10.4236/jmp.2021.125044>

Photon Structure and Wave Function from the Vector Potential Quantization

Constantin Meis

National Institute for Nuclear Science and Technology, CEA-Saclay, Gif-sur-Yvette, France

Email: constantin.meis@cea.fr

How to cite this paper: Meis, C. (2023) Photon Structure and Wave Function from the Vector Potential Quantization. *Journal of Modern Physics*, 14, 311-329.
<https://doi.org/10.4236/jmp.2023.143020>

Received: November 28, 2022

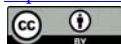
Accepted: February 24, 2023

Published: February 27, 2023

Copyright © 2023 by author(s) and Scientific Research Publishing Inc.

This work is licensed under the Creative Commons Attribution-NonCommercial International License (CC BY-NC 4.0).

<http://creativecommons.org/licenses/by-nc/4.0/>



Open Access

Abstract

A photon structure is advanced based on the experimental evidence and the vector potential quantization at a single photon level. It is shown that the photon is neither a point particle nor an infinite wave but behaves rather like a local “wave-corpuscle” extended over a wavelength, occupying a minimum quantization volume and guided by a non-local vector potential real wave function. The quantized vector potential oscillates over a wavelength with circular left or right polarization giving birth to orthogonal magnetic and electric fields whose amplitudes are proportional to the square of the frequency. The energy $\hbar\omega$ and momentum $\hbar\vec{k}$ are carried by the local wave-corpuscle guided by the non-local vector potential wave function suitably normalized.

Keywords

Photons, Photon Wave Function, Vector Potential Quantization, Photon Electric and Magnetic Fields, Photon Structure, Wave-Corpuscle Representation, Photon “Energy-Vector Potential” Equation

1. Introduction

A single photon is a particular relativistic massless wave-particle for which scientists have still major difficulties to attribute a clear physical representation. Historically, the scientific understanding of light’s properties, guided each time by the interpretation of the experiments, was a continuous balancing between the wave and particle natures. Before going ahead to the quantum description of the photons and the possible associated wave functions it is of crucial importance to understand the historical evolution of our concepts about the nature of light.

In the seventeenth century, reviving the ideas of ancient Greeks on light’s na-

ture, Newton advanced that light is composed of individual corpuscles which travel rectilinearly in a homogeneous medium [1] [2]. On the other side, Huygens developed a refined theory of light [3] [4] based on the wave representation and refuted Newton's corpuscular theory. Almost two centuries later, Young obtained experimentally interference patterns using different light sources. He also went further by explaining some polarization experiments assuming that light oscillations take particular orientations with respect to the propagation axis [4] [5]. By that time it was impossible to explain the experimental diffraction patterns using Newton's corpuscular theory, while Euler and Fresnel advanced easily precise interpretations applying the wave representation. James Clerk Maxwell published his remarkable theory on the electromagnetic waves in 1865. He established, for the first time, the relations between the electric and magnetic fields and advanced that light is composed of electromagnetic waves [4] [6]. Hertz confirmed this experimentally by discovering the long wavelength electromagnetic radiation. Thus, by the end of the nineteenth century, the scientific community replaced Newton's corpuscular representation by the wave theory of light. However, that was not for a long time.

In the beginning of the twentieth century, in order to explain the spectral energy density emitted by a black body, Max Planck reintroduced the notion of the particle nature of light with a particular sense. In fact, he assumed that bodies are composed of "oscillators" having the particularity to emit electromagnetic "packets" each with energy $h\nu$, where ν is the frequency and h Planck's constant [7] [8]. A few years later, based on Planck's works, Einstein proposed an interpretation of the photoelectric effect, first observed by Hertz [9], assuming that the electromagnetic radiation itself is composed of quanta with energy $h\nu$ [10]. Furthermore, in 1923 Compton advanced that only the light quanta could explain the experimental observations of the X-rays scattering on free electrons [11]. Hence, the photoelectric effect and Compton X-rays scattering have been always considered as the strongest arguments in favor of the particle nature of light and Quantum Electrodynamics (QED) theory was developed during the years of thirties to sixties based on the point photon model.

However, it is scarcely quoted in the literature that some important studies in favor of the electromagnetic wave theory have been totally disregarded and fell into oblivion. In fact, Wentzel in 1926 [12] and Beck in 1927 [13], as well as Lamb and Scully in the 1960s [14], demonstrated that the photoelectric effect can be interpreted directly by considering only the electromagnetic wave nature of light and without referring to photons at all. Furthermore, Klein and Nishina in 1929 [15] interpreted fully Compton's scattering by also considering the wave nature of light and without invoking the photon concept. So, still after the second world war, although the majority of scientists joined Bohr's *complementarity principle* according to which *light exhibits both wave and particle natures appearing mutually exclusively in each experiment*, a significant part of the scientific community still had the conviction that the electromagnetic wave representation was sufficient to understand light's nature.

Things radically changed in the years of fifties and sixties. Robinson in 1953 [16] and Hadlock in 1958 [17] carried out systematic experiments using microwaves crossing rectangular or circular apertures with variable dimensions and deduced that no energy is transmitted when the apertures dimensions are smaller than approximately the quarter of the wavelength ($\sim \lambda_c/4$). This was also confirmed later by Hunter and Wadlinger [18] [19] using X-band microwaves. On the other hand, Mandel's experiments in the sixties [20] [21], employing the recently discovered laser technics, concluded that a single photon has circular polarization and cannot be localized in a length shorter than its wavelength λ_c [22] and more generally it cannot be better localized than within a volume of the order of the cube of its wavelength (λ_c^3) [23] [24]. Consequently, the experimental evidence conflicts with both the point photon model, upon which QED has been developed, and the classical continuous electromagnetic wave theory, showing that light must be composed of localized "wave packets" traveling in vacuum at the universal velocity c [25] carrying energy $h\nu$ and momentum $h\nu/c$. In reality, the point photon concept has permitted to establish an extremely efficient mathematical approach in QED for describing states before and after an interaction processes [26] [27] [28]. However, it is obvious that it is inappropriate for the description of the real nature of a single photon state.

The development of the revolutionary parametric down converters techniques [29] [30] in the seventies permitted to realize conditions in which, with an excellent statistical confidence, only a single photon is present in the experimental device. Employing these techniques, the double prism experiment [31] carried out in the nineties contradicted Bohr's *mutual exclusiveness principle* showing that a *single photon exhibits both the wave and particle natures in the same experiment*. Furthermore, Grangier *et al.* demonstrated experimentally the indivisibility of photons [28] [32] while the entangled states experiments, first observed by Kosher and Commins [33] and further investigated in the eighties [34] [35], have shown that the photon should be locally an integral entity during the detection procedure but with a real non-local wave function.

Finally, a synthesis of the experimental studies shows that a single k -mode photon is a local, indivisible segment of the electromagnetic field with circular left or right polarization corresponding to spin $\pm h/2\pi$ respectively. It carries a quantum of energy $h\nu$ and has a momentum $h\nu/c$. Its intrinsic spatial length extends over a wavelength λ_k and can only be detected within a volume $\sim \lambda_k^3$, yielding that its radial expansion should be proportional to a fraction of its wavelength. Consequently, it seems to be a local "wave-corpuscle" absorbed and emitted as a whole and guided by a non-local wave function.

All these physical characteristics yield particular difficulties for the description of a real single photon state and consequently for the definition of an appropriate wave function within the quantum mechanics concepts.

In what follows we take into account all the above experimental and theoretical facts in order to obtain a physical picture of the single photon and establish

its wave function. We first give a brief presentation of the link between the classical and quantum theory through the vector potential and then we analyze the second quantization process resulting to a quantum description of the electromagnetic field. Then, we consider the spatial properties of a single photon state and enhance the quantization of the vector potential amplitude to a single photon level getting a physical representation conform to the experimental evidence. Next, we advance a photon wave function based on the vector potential quantization satisfying the wave propagation equation, Schrodinger's equation with the relativistic massless particle Hamiltonian and an equivalent equation for the vector potential. Finally, we discuss the characteristic properties as well as the normalization of the established photon wave-function.

2. The Electromagnetic Field Vector Potential in Classical and Quantum Theories

2.1. The Vector Potential: Classical to Quantum Link

Experiments have shown that the vector potential is not a mathematical artefact but a real physical field exerting a direct influence on charges [36] [37] [38] [39]. As that, it represents the fundamental link between the classical electromagnetic wave theory issued from Maxwell's equations and QED [8] [22] [28] [40] [41]. In the classical theory [3] [4] [5] [6] the energy density of a mode k of the electromagnetic wave depends on the square of the modulus of the electric field $\vec{E}_k(\vec{r}, t)$ and the magnetic induction $\vec{B}_k(\vec{r}, t)$ and writes

$$W_k(\vec{r}, t) = \frac{1}{2} \left(\varepsilon_0 |\vec{E}_k(\vec{r}, t)|^2 + \frac{1}{\mu_0} |\vec{B}_k(\vec{r}, t)|^2 \right) \quad (1)$$

with ε_0 and μ_0 being the electric permittivity and magnetic permeability of the vacuum respectively satisfying the relation $\varepsilon_0 \mu_0 c^2 = 1$, where c is the speed of light in vacuum.

For a monochromatic plane wave mode k with angular frequency ω_k and vector potential amplitude $a_{0k}(\omega_k)$ we have

$$\vec{E}_k(\vec{r}, t) = -2\omega_k a_{0k}(\omega_k) \hat{\varepsilon} \sin(\vec{k} \cdot \vec{r} - \omega_k t) \quad (2)$$

$$\vec{B}_k(\vec{r}, t) = -\frac{1}{c} 2\omega_k a_{0k}(\omega_k) (\hat{k} \times \hat{\varepsilon}) \sin(\vec{k} \cdot \vec{r} - \omega_k t) \quad (3)$$

with $\hat{\varepsilon}$ the unit vector perpendicular to the propagation axis, $|\vec{k}| = 2\pi/\lambda_k$ the wave-vector along the propagation axis and λ_k the wavelength of the mode k .

Using (2) and (3) in (1), the energy density writes as a function depending uniquely on the square of the vector potential amplitude and on the angular frequency

$$W_k(\vec{r}, t) = 4\varepsilon_0 \omega_k^2 |a_{0k}(\omega_k)|^2 \sin^2(\vec{k} \cdot \vec{r} - \omega_k t) \quad (4)$$

The mean value of the last expression over a period, that is over a wavelength, is

$$\langle W_k \rangle = 2\varepsilon_0 \omega_k^2 |a_{0k}(\omega_k)|^2 \quad (5)$$

It is important noting that the mean energy density $\langle W_k \rangle$ is independent on any external volume parameter. Hence, in the classical theory a free of cavity electromagnetic radiation mode occupies naturally a minimum volume while in a resonant cavity this volume corresponds roughly to that imposed by the boundary conditions and the cut-off wave vectors for the mode [4] [42].

Now, in the quantum description the energy density in a given volume V for a number $n(\omega_k)$ of k -mode photons, each with angular frequency ω_k and energy $\hbar\omega_k$ is simply

$$W_k = \frac{n(\omega_k) \hbar \omega_k}{V} \quad (6)$$

The link between the classical and the quantum theories is established by imposing the classical mean energy density over a period (5) to be equal to that of the quantum description (6). Thus, considering $n(\omega_k)=1$ we get the vector potential amplitude for a single k -mode photon

$$a_{0k}(\omega_k) = \sqrt{\frac{\hbar}{2\varepsilon_0 \omega_k V}} \quad (7)$$

The last relation constitutes the fundamental link between the classical and quantum theories of light and it is applied in QED in order to define the vector potential amplitude operators for a single photon [23] [27] [28]

$$\tilde{a}_{k\lambda} = \sqrt{\frac{\hbar}{2\varepsilon_0 \omega_k V}} a_{k\lambda}, \quad \tilde{a}_{k\lambda}^* = \sqrt{\frac{\hbar}{2\varepsilon_0 \omega_k V}} a_{k\lambda}^+ \quad (8)$$

where $a_{k\lambda}$ and $a_{k\lambda}^+$ are respectively the annihilation and creation non-Hermitian operators for a single k -mode and λ -polarization photon.

Obviously, this procedure introduces an external volume parameter V in the definitions of the single photon vector potential amplitude operators. Consequently, one could draw out that single photons in free space, in other words in a cavity with infinite dimensions, should have zero vector potential amplitudes and thus zero energy. This ambiguity, which is scarcely quoted in the literature, automatically implies that for a single photon the volume V in (8) cannot be an arbitrary external parameter but corresponds roughly to that defined by the boundary conditions in a cavity for the single radiation mode k .

2.2. The Electromagnetic Field Vector Potential in QED

In quantum field theory [23] [28] [40] the radiation vector potential writes

$$\vec{A}(\vec{r}, t) = \int \frac{d^3k}{(2\pi)^{3/2}} \sqrt{\frac{\hbar}{2\varepsilon_0 \omega_k}} \sum_{\lambda=1}^2 \left[\alpha_{k\lambda}(t) e^{i\vec{k} \cdot \vec{r}} \hat{\varepsilon}_{k\lambda} + \alpha_{k\lambda}^+(t) e^{-i\vec{k} \cdot \vec{r}} \hat{\varepsilon}_{k\lambda}^* \right] \quad (9)$$

where the discrete summation runs over two polarizations λ , circular left and right, and $\hat{\varepsilon}_{k\lambda}$ is the polarization complex unit vector.

According to the density of states theory the continuous summation over the modes k in (9) can be transformed to a discrete one over the k -mode photons by considering the electromagnetic field in a cavity of volume V so that we can put

$$\int \frac{d^3k}{(2\pi)^{3/2}} \rightarrow \sqrt{\frac{1}{V}} \sum_k \quad (10)$$

Therein, it is extremely important to note that the last relation is only valid for an ensemble of k -mode photons whose wavelengths λ_k are much shorter than the characteristic dimensions of the volume V

$$\lambda_k \ll V^{1/3} \quad (\forall k) \quad (11)$$

Adopting now Heisenberg's representation,

$$\alpha_{k\lambda}(t) = \alpha_{k\lambda} e^{-i\omega_k t}; \quad \alpha_{k\lambda}^+(t) = \alpha_{k\lambda}^+ e^{i\omega_k t} \quad (12)$$

and using the transformation relation (10) the vector potential of the electromagnetic field writes in QED [23] [28]

$$\vec{A}(\vec{r}, t) = \sum_k \sqrt{\frac{\hbar}{2\epsilon_0 \omega_k V}} \sum_{\lambda=1}^2 \left[\alpha_{k\lambda} e^{i(\vec{k} \cdot \vec{r} - \omega_k t)} \hat{\epsilon}_{k\lambda} + \alpha_{k\lambda}^+ e^{-i(\vec{k} \cdot \vec{r} - \omega_k t)} \hat{\epsilon}_{k\lambda}^* \right] \quad (13)$$

where we note that the vector potential amplitude for each mode k is defined exactly as in (7).

Considering the scalar potential of the electromagnetic field to be constant in space then the electric field writes

$$\vec{E}(\vec{r}, t) = -\partial \vec{A}(\vec{r}, t) / \partial t = i \sum_k \sqrt{\frac{\hbar \omega_k}{2\epsilon_0 V}} \sum_{\lambda=1}^2 \left[\alpha_{k\lambda} e^{i(\vec{k} \cdot \vec{r} - \omega_k t)} \hat{\epsilon}_{k\lambda} - \alpha_{k\lambda}^+ e^{-i(\vec{k} \cdot \vec{r} - \omega_k t)} \hat{\epsilon}_{k\lambda}^* \right] \quad (14)$$

The amplitudes in (13) and (14) have been obtained using the density of states theory and are valid only on the condition (11).

Now, the boundary conditions in cavities and waveguides impose the wave-vectors k of the modes to be higher than a characteristic cut-off value $k > k_{\text{cut-off}}$ ($\lambda_k < \lambda_{\text{cut-off}}$) depending on the dimensions as well as on the shape of the volume V containing the radiation field [42]. Consequently, for a volume V with finite dimensions the summation in (13) and (14) runs only over the modes k with wave-vectors higher than the minimum cut-off value $k_{\text{cut-off}}(V)$ corresponding to the shape and dimensions of V so that we can write more precisely

$$\vec{A}(\vec{r}, t) = \sum_{k > k_{\text{cut-off}}(V)} \sqrt{\frac{\hbar}{2\epsilon_0 \omega_k V}} \sum_{\lambda=1}^2 \left[\alpha_{k\lambda} e^{i(\vec{k} \cdot \vec{r} - \omega_k t)} \hat{\epsilon}_{k\lambda} + \alpha_{k\lambda}^+ e^{-i(\vec{k} \cdot \vec{r} - \omega_k t)} \hat{\epsilon}_{k\lambda}^* \right] \quad (15)$$

$$\vec{E}(\vec{r}, t) = i \sum_{k > k_{\text{cut-off}}(V)} \sqrt{\frac{\hbar \omega_k}{2\epsilon_0 V}} \sum_{\lambda=1}^2 \left[\alpha_{k\lambda} e^{i(\vec{k} \cdot \vec{r} - \omega_k t)} \hat{\epsilon}_{k\lambda} - \alpha_{k\lambda}^+ e^{-i(\vec{k} \cdot \vec{r} - \omega_k t)} \hat{\epsilon}_{k\lambda}^* \right] \quad (16)$$

The last equations represent the vector potential and the electric field of a large number of k -mode photons in a finite volume V with $\lambda_k \ll V^{1/3} (\forall k)$.

Following the last relations, we can now draw the amplitude of the electric field $|\vec{E}_k|$ and magnetic field $|\vec{B}_k|$ of a single k -mode photon

$$|\vec{E}_k| = \sqrt{\frac{\hbar \omega_k}{2\epsilon_0 V}}; \quad |\vec{B}_k| = \sqrt{\frac{\mu_0 \hbar \omega_k}{2V}} \quad (17)$$

These expressions are often used in the literature for the free single photon electric and magnetic fields amplitudes. However, as mentioned above, the expres-

sions (15) and (16) are only valid for a large number of photons in a cavity V under the condition (11) and consequently the attribution of the amplitudes (17) to free of cavity single photons has to be considered with caution.

3. Photon Wave Function from the Electric and Magnetic Fields

In quantum theory, a wave function $\Psi(\vec{r}, t)$, depending on space and time, characterizes precisely the quantum state of a given particle. It satisfies Schrödinger's equation with eigenvalue the corresponding energy of the quantum state. However, in the case of the single photon, which is a relativistic massless "particle", the wave functions that have been proposed until now are not in general satisfactory. Nevertheless, some very fruitful studies have already been carried out in order to establish a photon wave function and they are given here in the bibliography [43]-[48] and commented extensively [49] [50] [51] [52] [53]. The most pertinent theoretical developments are based on the complex vector function $\vec{F}(\vec{r}, t)$ initially introduced by Riemann

$$\vec{F}(\vec{r}, t) = \frac{1}{2\sqrt{\epsilon_0}} \vec{D}(\vec{r}, t) + i \frac{1}{2\sqrt{\mu_0}} \vec{B}(\vec{r}, t) \quad (18)$$

where the electric displacement flux density $\vec{D}(\vec{r}, t)$ and the magnetic field flux density $\vec{B}(\vec{r}, t)$ are defined respectively through the electric and magnetic fields intensities $\vec{E}(\vec{r}, t)$ and $\vec{H}(\vec{r}, t)$ following the expressions

$$\vec{D}_{\alpha\beta}(t) = \hat{\epsilon}_{\alpha\beta} \cdot \vec{E}_{\alpha\beta}(t); \quad \vec{B}_{\alpha\beta}(t) = \hat{\mu}_{\alpha\beta} \cdot \vec{H}_{\alpha\beta}(t) \quad (19)$$

with the electric permittivity $\hat{\epsilon}_{\alpha\beta}$ and magnetic permeability $\hat{\mu}_{\alpha\beta}$ being tensors ($\alpha = x, y, z$ and $\beta = x, y, z$) characterizing the intrinsic electric and magnetic nature of the medium in which the electric and magnetic fields propagate.

At that point it is important to mention that the electric field $\vec{E}(\vec{r}, t)$ and the magnetic induction $\vec{B}(\vec{r}, t)$ are fundamental fields while $\vec{D}(\vec{r}, t)$ and $\vec{H}(\vec{r}, t)$ are fields that include macroscopically the response of the medium. In the case of an isotropic medium, in which the electric and magnetic properties are the same in all directions, the electric field is parallel to the electric displacement while the magnetic field is parallel to the magnetic field flux intensity

$$\vec{D}(\vec{r}, t) = \epsilon \vec{E}(\vec{r}, t); \quad \vec{H}(\vec{r}, t) = \frac{1}{\mu} \vec{B}(\vec{r}, t) \quad (20)$$

Using the complex function $\vec{F}(\vec{r}, t)$ Maxwell's equations in an homogeneous medium become

$$i \frac{\partial}{\partial t} \vec{F}(\vec{r}, t) = c \vec{\nabla} \times \vec{F}(\vec{r}, t) \quad (21)$$

$$\vec{\nabla} \cdot \vec{F}(\vec{r}, t) = 0 \quad (22)$$

In this way, a suggested photon wave function was defined as

$$\Psi(\vec{r}, t) = \begin{pmatrix} \frac{1}{2\sqrt{\epsilon_0}} \vec{D}(\vec{r}, t) + i \frac{1}{2\sqrt{\mu_0}} \vec{B}(\vec{r}, t) \\ \frac{1}{2\sqrt{\epsilon_0}} \vec{D}(\vec{r}, t) - i \frac{1}{2\sqrt{\mu_0}} \vec{B}(\vec{r}, t) \end{pmatrix} \quad (23)$$

which is a six component vector satisfying Schrödinger's equation

$$i\hbar \frac{\partial}{\partial t} \Psi(\vec{r}, t) = c \begin{pmatrix} -i\hbar \vec{\nabla} \cdot \vec{S} & 0 \\ 0 & i\hbar \vec{\nabla} \cdot \vec{S} \end{pmatrix} \Psi(\vec{r}, t) \quad (24)$$

where $\vec{S}$ is the spin matrix for a spin 1 particle with the Cartesian components

$$S_x = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -i \\ 0 & i & 0 \end{pmatrix}, \quad S_y = \begin{pmatrix} 0 & 0 & i \\ 0 & 0 & 0 \\ -i & 0 & 0 \end{pmatrix}, \quad S_z = \begin{pmatrix} 0 & -i & 0 \\ i & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad (25)$$

The square modulus of the above defined photon wave function $\Psi(\vec{r}, t)$ provides the classical electromagnetic field energy density (1) at an instant t and at a given coordinate $\vec{r}$ on the propagation axis

$$|\Psi(\vec{r}, t)|^2 = W(\vec{r}, t) = \frac{1}{2} \left(\epsilon_0 |\vec{E}(\vec{r}, t)|^2 + \frac{1}{\mu_0} |\vec{B}(\vec{r}, t)|^2 \right) \quad (26)$$

The energy is obtained directly by integrating in space the bilinear form of the complex vector function $\vec{F}(\vec{r}, t)$

$$E = \int \vec{F}^*(\vec{r}, t) \cdot \vec{F}(\vec{r}, t) d^3r \quad (27)$$

However, the relations (26) and (27) give respectively the energy density and the energy of the electromagnetic field but not precisely those of a single photon state. Furthermore, in quantum mechanics, the integration in space of the square modulus of the normalized photon wave function, given by (26), should correspond to the probability of localizing a photon and not to the energy of the classical electromagnetic field representation. In a characteristic single photon quantization volume on the propagation axis this quantity should be equal to unity and this is not really the case here for $\Psi(\vec{r}, t)$.

Finally, some different versions from the one described above have been advanced [43]-[48] for the photon wave function but without introducing a real progress in the whole problematics.

4. Spatial Properties of a Single Photon and the Vector Potential Quantization

4.1. Lateral Expansion of a Single Photon, Theoretical and Experimental Facts

The moment of inertia I_k that may be attributed to a single k -mode photon with angular frequency ω_k and wavelength λ_k writes

$$I_k = \frac{\hbar}{\omega_k} \quad (28)$$

From the mass-energy equivalence the mass m_k , which is not a rest mass, that

may be attributed to a k -mode photon is

$$m_k = \frac{\hbar \omega_k}{c^2} \quad (29)$$

In order to estimate the radial expansion of the photon we may assume here a homogeneous distribution of the field in the single photon state quantization volume and consider the mechanical analogue in two simple cases, that of a cylinder whose axis is along the propagation axis and that of a sphere.

For a cylinder we have

$$I_{\text{cylinder}} = \frac{1}{2} m_k r_{\text{cylinder}}^2 = \frac{1}{2} \frac{\hbar \omega_k}{c^2} r_{\text{cylinder}}^2 = \frac{\hbar}{\omega_k} \quad (30)$$

and consequently

$$r_{\text{cylinder}} = \left(\frac{2c^2}{\omega_k^2} \right)^{1/2} = \frac{\lambda_k}{\pi\sqrt{2}} \sim \frac{\lambda_k}{4.4} \quad (31)$$

In the case of a spherical distribution of the field inside the photon quantization volume we have

$$I_{\text{sphere}} = \frac{2}{5} m_k r_{\text{sphere}}^2 = \frac{2}{5} \frac{\hbar \omega_k}{c^2} r_{\text{sphere}}^2 = \frac{\hbar}{\omega_k} \quad (32)$$

so that the sphere radius is

$$r_{\text{sphere}} = \left(\frac{5}{2} \right)^{1/2} \frac{c}{\omega_k} = \frac{\lambda_k}{\pi\sqrt{8/5}} \sim \frac{\lambda_k}{3.9} \quad (33)$$

Consequently, from the above simple theoretical calculations we may presume that the radius of a single k -mode photon should be a fraction of its wavelength, roughly $\sim \lambda_k/4$.

From the experimental point of view, the fact that a single mode plane electromagnetic wave should have a minimum radius was already suspected since the very first applications of the Faraday grid shielding. As mentioned in the introduction, various experiments carried out by Robinson, Hadlock, Hunter and Wadlinger [16] [17] [18] [19] using microwaves and measuring the transmitted power through rectangular or circular apertures concluded that the photons lateral expansion should be indeed close to the quarter of its wavelength $\sim \lambda_k/4$, confirming the above simple calculations based on the moment of inertia.

On the other hand, it is well established [20] [21] [22] that a single photon cannot be conceived in a length shorter than its wavelength and it cannot be localized within a volume less than $\sim \lambda_k^3$ [21] [23] entailing that the single photon quantization volume should be proportional to λ_k^3 , that is proportional to ω_k^{-3} .

4.2. Single Photon Quantization Volume

4.2.1. The Single Photon Quantization Volume from the Density of States Theory

According to the density of states theory, taking into account the two spin values

$\pm\hbar$, corresponding to circular left and right polarization, as well as the two possible directions $\pm z$ along the same propagation axis z , the number of k -mode photons $dn(\omega_k)$ in the quantization volume V in the frequency interval between ω_k and $\omega_k + d\omega_k$ writes [23] [27]

$$dn(\omega_k) = 4V \frac{\omega_k^2}{\pi^2 c^3} d\omega_k \quad (34)$$

where we have also considered that all the possible states in the volume V are occupied.

Thus,

$$n(\omega_k) = 4V \frac{\omega_k^3}{3\pi^2 c^3} \quad (35)$$

Considering now $n(\omega_k) = 1$ the corresponding quantization volume that could be attributed to a single photon is

$$V_k = \left(\frac{3}{4} \pi^2 c^3 \right) \omega_k^{-3} \quad (36)$$

The last expression is in agreement with the experimental evidence that the single photon quantization volume is proportional to ω_k^{-3} , getting

$$V_k = \frac{3}{32\pi} \lambda_k^3 \approx 0.029 \lambda_k^3 \quad (37)$$

Thus, the minimum quantization volume corresponding to a single photon with a wavelength λ_k is roughly 3% of the volume λ_k^3 .

4.2.2. The Single Photon Quantization Volume from the Energy Normalization

We can also obtain the single photon quantization volume in a different way by normalizing the electromagnetic energy of a plane-wave mode to that of a single photon

$$E_k = \int \varepsilon_0 \omega_k^2 \alpha_{0k}^2(\omega_k) \left[\hat{\varepsilon}_{k\lambda} e^{i(\vec{k} \cdot \vec{r} - \omega_k t + \phi)} + \hat{\varepsilon}_{k\lambda}^* e^{-i(\vec{k} \cdot \vec{r} - \omega_k t + \phi)} \right]^2 d^3r = \hbar \omega_k \quad (38)$$

where $\alpha_{0k}(\omega_k)$ represents the single k -mode photon vector potential amplitude and $\hat{\varepsilon}_{k\lambda}$ the polarization complex unit vector.

The last equation holds at any instant t if the k -mode photon polarization unit vector $\hat{\varepsilon}_{k\lambda}$ has two orthogonal components $\hat{e}_1$ and $\hat{e}_2$ such as $\hat{e}_1 \cdot \hat{e}_2 = 0$ and $\hat{\varepsilon}_{k\lambda} = \sigma_{k1} \hat{e}_1 + \sigma_{k2} \hat{e}_2$ with $|\sigma_{k1}|^2 + |\sigma_{k2}|^2 = 1$. Obviously, these conditions are satisfied naturally by circular left (L) and right (R) polarization unit vectors

$$\hat{\varepsilon}_{L,R} = \frac{1}{\sqrt{2}} (\hat{e}_1 \pm i \hat{e}_2), \text{ so that we have [49]}$$

$$E_k = \int 2\varepsilon_0 \omega_k^2 \alpha_{0k}^2(\omega_k) d^3r = \hbar \omega_k \quad (39)$$

Now, the dimension analysis of the vector potential issued from the general solution of Maxwell's equations yields that it is proportional to an angular frequency [3] [4] [5] [6]

$$\bar{A}(\vec{r}, t) = \frac{\mu}{4\pi} \int \frac{\vec{J}(\vec{r}', t - |\vec{r} - \vec{r}'|/c)}{|\vec{r} - \vec{r}'|} d^3r' \propto \omega \quad (40)$$

where μ is the magnetic permeability and J is the current density ($\text{C}\cdot\text{m}^{-2}\cdot\text{s}^{-1}$).

Thus, considering in (5) that the vector potential amplitude is proportional to ω then $\langle W_k \rangle$ is proportional to ω^4 which is confirmed experimentally for the energy density radiated by a dipole [5] [6]. The concluded property for the vector potential amplitude is related to its natural units and consequently is gauge independent.

Hence, we may consider that the vector potential amplitude $\alpha_{0k}(\omega_k)$ of a single k -mode photon can be expressed as follows [49] [50] [51]

$$\alpha_{0k}(\omega_k) = \xi \omega_k \quad (41)$$

where ξ is a constant to be determined through the normalization procedure.

Obviously, when introducing (41) in (39) an appropriate volume V_k has to be considered for the single k -mode photon for the equation to hold

$$E_k = 2\varepsilon_0 \xi^2 \omega_k^4 V_k = \hbar \omega_k \quad (42)$$

From the last relation we draw that the energy density of a single photon state is

$$W_k = 2\varepsilon_0 \xi^2 \omega_k^4 \quad (43)$$

It is worth noticing that the last expression depends on the fourth power of the angular frequency as in the case of the energy density radiated by a dipole [5] [6].

We now deduce from (42) that the characteristic volume of a cavity-free single photon writes [51] [53] [54]

$$V_k = \left(\frac{\hbar}{2\varepsilon_0 \xi^2} \right) \omega_k^{-3} \quad (44)$$

which is in agreement with the experimental evidence regarding the frequency dependence.

Hence, in order to estimate ξ we can integrate (39) by using some experimental facts. Considering the propagation along the z axis, with ρ the lateral coordinate, and respecting the experimental results by imposing the limits $0 \leq z \leq \lambda_k$ and $0 \leq \rho \leq \eta \lambda_k$ [49] where η characterizes the radial extension, (39) gives

$$\varepsilon_0 \xi^2 \eta^2 \omega_k^4 \lambda_k^3 = (2\pi c)^3 \varepsilon_0 \xi^2 \eta^2 \omega_k = \hbar \omega_k \quad (45)$$

Hence,

$$\eta \xi = \pm \frac{1}{(2\pi)^{3/2}} \sqrt{\frac{\hbar}{\varepsilon_0 c^3}} \quad (46)$$

Instead of using the approximate experimental value of $\eta \sim 1/4$ in (46) in order to fix ξ we can try to get a more precise result. In fact, Equation (45) writes with a slight rearrangement

$$4\pi c \left(\frac{1}{4\pi c} \varepsilon_0 \omega_k^2 \eta^2 \alpha_{0k}(\omega_k) \lambda_k^3 \right) \alpha_{0k}(\omega_k) = 4\pi c Q \alpha_{0k}(\omega_k) = \hbar \omega_k \quad (47)$$

where Q has charge units. Using (41) and (46) we get

$$Q = \frac{1}{4\pi c} \varepsilon_0 \omega_k^2 \eta^2 \alpha_0 (\omega_k) \lambda_k^3 = \pm \frac{\eta}{2} \sqrt{\varepsilon_0 \hbar c} \quad (48)$$

with $\hbar$ being Planck's constant.

It is of high importance to underline here that when introducing in (48) the approximate experimental value for $\eta \sim 1/4$ we get

$$Q \simeq \pm 1.6 \times 10^{-19} \text{ C} \quad (49)$$

which is the electron-positron charge, a physical constant appearing naturally [51] in the normalization process.

We draw that the physical origin of the elementary charge is strongly related to the photon vector potential [51] [53] [54]. Now, we can define more precisely η and ξ by replacing Q in (48) by the elementary charge e and using the fine structure constant $\alpha = e^2/4\pi\varepsilon_0\hbar c = 1/137$, we get

$$\eta = \sqrt{8\alpha} \quad (50)$$

Introducing (50) in (46) we obtain the vector potential amplitude normalization constant [50] [51] [53] [54]

$$\xi = \pm \frac{1}{(2\pi)^{3/2}} \sqrt{\frac{\hbar}{8\alpha\varepsilon_0 c^3}} = \pm \frac{\hbar}{4\pi|e|c} = \pm 1.747 \times 10^{-25} \text{ Volt} \cdot \text{m}^{-1} \cdot \text{s}^2 \quad (51)$$

Using (51) in (44) the single k -mode photon quantization volume writes now

$$V_k = \left(\frac{\hbar}{2\varepsilon_0 \xi^2} \right) \omega_k^{-3} = \frac{\eta^2}{2} \lambda_k^3 = 4\alpha \lambda_k^3 \simeq 0.029 \lambda_k^3 \quad (52)$$

which is equivalent to (37).

Consequently, the helicoidally distributed vector potential field of a single photon state over a wavelength occupies roughly 3% of the λ_k^3 volume, as we have seen before, and has a wave front section [54]

$$S_k = 8\pi\alpha\lambda_k^2 \simeq 0.18\lambda_k^2 \quad (53)$$

The last relations are useful for technological applications like wave-guides and fiber optics transmissions or Faraday grid shielding...etc.

Thus, the fundamental properties of the photon, energy E_k , momentum $\vec{p}_k$ and wave-vector $\vec{k}$, are complemented by the vector potential amplitude α_{0k} expressing its intrinsic electromagnetic nature

$$E_k/\hbar = |\vec{p}_k|c/\hbar = |\vec{k}|c = \alpha_{0k}/|\xi| = \omega_k \quad (54)$$

From the relation (54) and Heisenberg's *energy-time* uncertainty a *vector potential-time* uncertainty is drawn

$$\delta E_k \delta t \geq \hbar \rightarrow \delta \alpha_{0k} \delta t \geq |\xi| \quad (55)$$

showing that the particle (energy) and the wave (vector potential) are intrinsic properties of the single photon nature.

5. Single Photon Wave-Particle (Classical-Quantum) Physical Properties

Following the above considerations, the vector potential for a free single k -mode photon with λ -polarization (circular left or right) writes in both classical and quantum formalism respectively [53]

$$\begin{aligned}\tilde{\alpha}_{k\lambda}(\vec{r}, t) &= \xi \omega_k \left[\hat{\epsilon}_{k\lambda} \alpha_{k\lambda} e^{i(\vec{k} \cdot \vec{r} - \omega_k t)} + \hat{\epsilon}_{k\lambda}^* \alpha_{k\lambda}^+ e^{-i(\vec{k} \cdot \vec{r} - \omega_k t)} \right] \\ \tilde{\alpha}_{k\lambda}(\vec{r}, t) &= \xi \omega_k \left[\hat{\epsilon}_{k\lambda} e^{i(\vec{k} \cdot \vec{r} - \omega_k t)} + \hat{\epsilon}_{k\lambda}^* e^{-i(\vec{k} \cdot \vec{r} - \omega_k t)} \right]\end{aligned}\quad (56)$$

For a free k -mode photon, the volume V_k expressed by (37) and (52) corresponds to the space in which the quantized vector potential rotates around the propagation axis at the angular frequency ω_k over a period (that is over a wavelength λ_k). This precession generates orthogonal electric and magnetic fields whose amplitudes are

$$|\vec{\mathcal{E}}_k| = |-\partial \tilde{\alpha}_{k\lambda}(\vec{r}, t)/\partial t| \propto |\xi| \omega_k^2; \quad |\vec{\mathcal{B}}_k| \propto \sqrt{\epsilon_0 \mu_0} |\xi| \omega_k^2 \quad (57)$$

It is important to note that last expressions are directly proportional to the square of the angular frequency and do not depend on any external volume parameter [53] [54] [55].

The quantum properties of a single photon, energy, momentum and spin result readily from the integration of the classical electromagnetic expressions over the quantization volume V_k , linking by this way the classical (wave) to the quantum (particle) representations [51] [53] [54]. Indeed, using (41) and (52) the energy writes

$$E_k = \int_{V_k} 2\epsilon_0 \alpha_{0k}^2 \omega_k^2 d^3r = \int_{V_k} 2\epsilon_0 \xi^2 \omega_k^4 d^3r = 2\epsilon_0 \xi^2 \omega_k^4 V_k = \hbar \omega_k \quad (58)$$

Considering circular polarizations for the electric and magnetic fields components and using (57) the momentum is obtained directly

$$\vec{p}_k = \int_{V_k} \epsilon_0 \vec{\mathcal{E}}_{k\lambda} \times \vec{\mathcal{B}}_{k\lambda} d^3r = \epsilon_0 \left(\sqrt{2} \omega_k \alpha_{0k} \right) \left(\sqrt{2} \omega_k \alpha_{0k} / c \right) V_k \vec{k} / |\vec{k}| = \hbar \vec{k} \quad (59)$$

According to the classical electromagnetic theory the spin writes through the electric and magnetic fields components, hence, using again the circular polarization we get

$$|\vec{S}| = \int_{V_k} \epsilon_0 \vec{r} \times (\vec{\mathcal{E}}_{k\lambda} \times \vec{\mathcal{B}}_{k\lambda}) d^3r = \pm \epsilon_0 (c/\omega_k) \left(\sqrt{2} \omega_k \alpha_{0k} \right) \left(\sqrt{2} \omega_k \alpha_{0k} / c \right) V_k = \pm \hbar \quad (60)$$

where we have considered the mean value of the distance [56] for a single photon state $\langle |\vec{r}| \rangle_k = c/\omega_k$

The last relations show that the quantum properties of the single photon can be obtained from the classical electromagnetic field considered in the quantization volume V_k entailing that the photon has naturally a three dimensional extension and consequently the term “wave-corpuscle” should be more appropriate.

Thus, a photon is not a point particle and Heisenberg’s uncertainty relation for the position and momentum is precisely due to its spatial properties and

more precisely to the quantization volume V_k . Indeed, replacing V in (8) by V_k we get the single photon vector potential amplitude operators expressed through the annihilation and creation operators [51] [53]

$$\tilde{\alpha}_{0k\lambda} = \xi\omega_k\alpha_{k\lambda}; \quad \tilde{\alpha}_{0k\lambda}^* = \xi\omega_k\alpha_{k\lambda}^+ \quad (61)$$

The corresponding position $\tilde{Q}_{k\lambda}$ and momentum $\tilde{P}_{k\lambda}$ Hermitian operators [23] [27] [28] write

$$\tilde{Q}_{k\lambda} = \sqrt{\varepsilon_0 V_k} (\tilde{\alpha}_{0k\lambda} + \tilde{\alpha}_{0k\lambda}^*); \quad \tilde{P}_{k\lambda} = -i\omega_k \sqrt{\varepsilon_0 V_k} (\tilde{\alpha}_{0k\lambda} - \tilde{\alpha}_{0k\lambda}^*) \quad (62)$$

Thus, introducing (61) in (62) and using (52) Heisenberg's commutation relation, a fundamental concept in quantum theory, results directly [51] [53] [54]

$$[\tilde{Q}_{k\lambda}, \tilde{P}_{k'\lambda'}] = -i\varepsilon_0\omega_k^2\omega_{k'}\sqrt{V_k V_{k'}} \left[(\xi a_{k\lambda} + \xi a_{k\lambda}^+), (\xi a_{k'\lambda'} - \xi a_{k'\lambda'}^+) \right] = i\hbar\delta_{kk'}\delta_{\lambda\lambda'} \quad (63)$$

The position-momentum uncertainty, as well as the energy and vector potential uncertainties with respect to time are intrinsic physical properties of the nature of the photon.

By this way, the general equation for the vector potential of the electromagnetic wave considered as a superposition of plane wave modes writes

$$\vec{A}(\vec{r}, t) = \sum_{k,\lambda} \xi\omega_k \left[\hat{\varepsilon}_{k\lambda} e^{i(\vec{k}\cdot\vec{r} - \omega_k t + \phi)} + \hat{\varepsilon}_{k\lambda}^* e^{-i(\vec{k}\cdot\vec{r} - \omega_k t + \phi)} \right] \quad (64)$$

and the vector potential operator for a large number of cavity free photons in quantum electrodynamics is

$$\vec{A} = \sum_{k,\lambda} \xi\omega_k \left[a_{k\lambda} \hat{\varepsilon}_{k\lambda} e^{i(\vec{k}\cdot\vec{r} - \omega_k t + \phi)} + a_{k\lambda}^+ \hat{\varepsilon}_{k\lambda}^* e^{-i(\vec{k}\cdot\vec{r} - \omega_k t + \phi)} \right] \quad (65)$$

In what follows, the photon wave-particle equation is introduced showing that the vector potential with the quantized amplitude can be naturally a wave function for the photon.

6. Wave-Particle Equation for the Photon and the Vector Potential Wave Function

It can be readily demonstrated that the vector potential $\vec{\alpha}_{k\lambda}(\vec{r}, t)$ with the quantized amplitude $\xi\omega_k$ expressed by (41) satisfies Maxwell's wave propagation

$$\vec{\nabla}_k^2 \vec{\alpha}_{k\lambda}(\vec{r}, t) - \frac{1}{c^2} \frac{\partial^2}{\partial t^2} \vec{\alpha}_{k\lambda}(\vec{r}, t) = 0 \quad (66)$$

and the photon *vector potential – energy (wave – particle)* equation [51] [53] [54]

$$i\left(\frac{\xi}{\hbar}\right) \frac{\partial}{\partial t} \vec{\alpha}_{k\lambda}(\vec{r}, t) = \left(\frac{\tilde{\alpha}_{0k}}{\tilde{H}_k}\right) \vec{\alpha}_{k\lambda}(\vec{r}, t) \quad (67)$$

where the operator $\vec{\nabla}_k$ acts upon the mode k .

Equation (67) expresses both the first (energy) and second (vector potential) quantization of the electromagnetic field.

Obviously, the relativistic Hamiltonian for a massless particle $\tilde{H}_k = -i\hbar c \vec{\nabla}_k$ as well as the vector potential amplitude operator $\tilde{\alpha}_{0k} = -i\xi c \vec{\nabla}_k$ have the eigenvalues $\hbar\omega_k$ and $\xi\omega_k$ respectively [51] [53] [54] [55] [57] [58] [59]. It is worth noting the symmetry between $\{E_k, \hbar\}$ and $\{\alpha_{0k}, \xi\}$ characterizing respectively the particle (energy) and electromagnetic wave (vector potential) intrinsic natures of a single photon.

Consider now the emission in vacuum of a single k -mode photon at the coordinate $\vec{r}$ at an instant t . The probability to be localized at the instant t' at the coordinate $\vec{r}' = \vec{r} + c(t' - t)\hat{r}$ can be obtained by the square of the modulus of the vector potential function $\tilde{\alpha}_{k\lambda}(\vec{r}, t)$

$$|\tilde{\alpha}_{k\lambda}(z, t)|^2 = \xi^2 \omega_k^2 \propto \lambda_k^{-2} \quad (68)$$

It is evident that the higher the frequency, the shorter the photon wavelength and the higher the localization probability. This is in agreement with Heisenberg's uncertainty principle as well as with the experimental evidence and consequently the vector potential function $\tilde{\alpha}_{k\lambda}(\vec{r}, t)$ can be considered as a real wave function for the photon that can be suitably normalized.

The photon momentum according to (59) is $\vec{p}_k = \hbar \vec{k} = (h/\lambda_k)\hat{k}$ and considering the propagation along the z -axis Heisenberg's position-momentum uncertainty, expressed by (63), writes now simply

$$\delta z \delta p_{k_z} \geq h \rightarrow \delta z \delta(1/\lambda_k) \geq 1 \quad (69)$$

The physical meaning of the last relation is that the localization uncertainty of a single photon on the propagation axis is of the order of the wavelength λ_k . Furthermore, as already mentioned, a single photon is localizable only within a volume proportional to λ_k^3 .

Now, taking into account the above considerations and weighting the vector potential function $\tilde{\alpha}_{k\lambda}(\vec{r}, t)$ by the factor $(\varepsilon_0 \omega_k / \hbar)^{1/2}$ a general wave function for the photon can be defined

$$\Phi_{k,(L,R)}(\vec{r}, t) = \left(\frac{\varepsilon_0 \omega_k}{\hbar} \right)^{1/2} \left[\alpha_{0k}(\omega_k) \left(\hat{\varepsilon}_{k,(L,R)} e^{i(\vec{k} \cdot \vec{r} - \omega_k t)} + \hat{\varepsilon}_{k,(L,R)}^* e^{-i(\vec{k} \cdot \vec{r} - \omega_k t)} \right) \right] \quad (70)$$

where the polarization can be either left (L) or right (R) circular.

It can be readily demonstrated that $\Phi_{k,(L,R)}(\vec{r}, t)$ satisfies the wave propagation equation (66) and the *vector potential – energy* equation (67).

In addition, whatever the circular polarization, the square of the modulus of the normalized general wave function (70) gives the inverse of the single photon quantization volume (52)

$$|\Phi_{k,(L,R)}(\vec{r}, t)|^2 = \frac{\varepsilon_0 \omega_k}{\hbar} \left[\xi^2 \omega_k^2 \left(|\hat{\varepsilon}_{k,(L,R)}|^2 + |\hat{\varepsilon}_{k,(L,R)}^*|^2 \right) \right] = \frac{2\varepsilon_0 \xi^2 \omega_k^3}{\hbar} = \frac{1}{V_k} \quad (71)$$

so that

$$\int_{V_k} |\Phi_{k,(L,R)}(\vec{r}, t)|^2 d^3 r = 1 \quad (72)$$

The probability to find the photon in the volume V_k around the coordinate $\vec{r}$

on the propagation axis equals one.

On the other hand, the mean value of the relativistic massless particle Hamiltonian in the $\Phi_{k,(L,R)}(\vec{r}, t)$ photon state is Planck's single photon energy

$$\begin{aligned} & \left| \left\langle \Phi_{k,(L,R)}(\vec{r}, t) \left| H_k \right| \Phi_{k,(L,R)}(\vec{r}, t) \right\rangle \right| = \left| \left\langle \Phi_{k,(L,R)}(\vec{r}, t) \left| -i\hbar c \vec{\nabla}_k \right| \Phi_{k,(L,R)}(\vec{r}, t) \right\rangle \right| \\ &= \left| \varepsilon_0 \xi^2 \omega_k^4 \int_{V_k} \left(e^{i(\vec{k} \cdot \vec{r} - \omega_k t)} \hat{\mathcal{E}}_{k,(L,R)} + e^{-i(\vec{k} \cdot \vec{r} - \omega_k t)} \hat{\mathcal{E}}_{k,(L,R)}^* \right) \left(e^{i(\vec{k} \cdot \vec{r} - \omega_k t)} \hat{\mathcal{E}}_{k,(L,R)} - e^{-i(\vec{k} \cdot \vec{r} - \omega_k t)} \hat{\mathcal{E}}_{k,(L,R)}^* \right) d^3 r \right| \quad (73) \\ &= \varepsilon_0 \xi^2 \omega_k^4 \int_{V_k} \left(\left| e^{i(\vec{k} \cdot \vec{r} - \omega_k t)} \hat{\mathcal{E}}_{k,(L,R)} \right|^2 + \left| e^{-i(\vec{k} \cdot \vec{r} - \omega_k t)} \hat{\mathcal{E}}_{k,(L,R)}^* \right|^2 \right) d^3 r \\ &= 2\varepsilon_0 \xi^2 \omega_k^4 V_k = \hbar \omega_k \end{aligned}$$

while the energy density is obtained directly by

$$W_k = \left| \Phi_{k,(L,R)}(\vec{r}, t) \right|^2 \left| \left\langle \Phi_{k,(L,R)}(\vec{r}, t) \left| H_k \right| \Phi_{k,(L,R)}(\vec{r}, t) \right\rangle \right| = \frac{\hbar \omega_k}{V_k} = 2\varepsilon_0 \xi^2 \omega_k^4 \quad (74)$$

By the same token the mean value of the momentum operator is

$$\left| \left\langle \Phi_{k,(L,R)}(\vec{r}, t) \left| \vec{p}_k \right| \Phi_{k,(L,R)}(\vec{r}, t) \right\rangle \right| = \left| \left\langle \Phi_{k,(L,R)}(\vec{r}, t) \left| -i\hbar \vec{\nabla}_k \right| \Phi_{k,(L,R)}(\vec{r}, t) \right\rangle \right| = \frac{\hbar \omega_k}{c} \quad (75)$$

Consequently, the photon wave function defined in (70) obtained from the vector potential quantization at a single photon level, satisfies the propagation equation (66), the vector potential – energy equation (67), the normalization probability condition (72) and represents a real eigenstate of the relativistic massless particle Hamiltonian and the corresponding momentum operator with eigenvalues $\hbar \omega_k$ and $\hbar |\vec{k}|$ respectively.

7. Epilogue

Based on the experimental confirmation that the vector potential is a real physical entity we advanced theoretical developments complementing the standard formalism for the description of a single photon and the definition of a wave function.

The quantization of the vector potential amplitude of a cavity free k -mode photon with angular frequency ω_k is $\alpha_{0k} = \xi \omega_k$, where $\xi = \hbar/4\pi e c$, and yields naturally a *vector potential-energy* (electromagnetic wave-particle) equation expressing the simultaneous wave-particle nature of the photon. A single photon is a local indivisible part of the electromagnetic field extending over a wavelength λ_k and composed of the quantized vector potential rotating at the angular frequency ω_k with circular left or right polarization corresponding respectively to spin $\pm \hbar$. The precession of the vector potential gives birth to orthogonally oscillating electric and magnetic fields whose amplitudes are proportional to the square of the angular frequency $|\xi| \omega_k^2$. The lateral expansion of the photon has been estimated experimentally yielding a photon quantization volume V_k proportional to λ_k^3 . The energy and momentum of the photon as well as the spin are obtained from the classical electromagnetic expressions integrated over the volume V_k showing that the photon is not a point particle but

rather a “wave-corpuscule”. Indeed, it is readily demonstrated that the origin of Heisenberg’s uncertainty lays precisely on the photon spatial properties.

The quantized vector potential $\Phi_{k,(L,R)}(\vec{r},t)$, with circular left (L) or right (R) polarization, defined in Equation (70) satisfies both the propagation equation and the *vector potential-energy* equation behaving as a natural wave function for the photon. In fact, the square modulus of its amplitude is proportional to the square of the angular frequency and consequently can be employed for defining a localization probability yielding that the higher the frequency the better the localization in agreement with the experiments. Furthermore, the normalization factor $(\varepsilon_0\omega_k/\hbar)^{1/2}$ ensures that the summation of $|\Phi_{k,(L,R)}(\vec{r},t)|^2$ over the quantization volume V_k equals unity. Also, the mean value of the relativistic massless particle Hamiltonian $\left\langle \Phi_{k,(L,R)}(\vec{r},t) \left| -i\hbar c \vec{\nabla}_k \right| \Phi_{k,(L,R)}(\vec{r},t) \right\rangle$ equals the single photon energy $\hbar\omega_k$ while that of the momentum operator $\left\langle \Phi_{k,(L,R)}(\vec{r},t) \left| -i\hbar \vec{\nabla}_k \right| \Phi_{k,(L,R)}(\vec{r},t) \right\rangle$ equals the single photon momentum $\hbar|\vec{k}|$.

Thus, a photon appears to be a local three-dimensional electromagnetic “wave-corpuscule” carrying the energy quantum $\hbar\omega_k$ and the momentum $\hbar|\vec{k}|$, guided by the non-local vector potential function $\Phi_{k,(L,R)}(\vec{r},t)$ with circular polarization corresponding to spin $\pm\hbar$.

Conflicts of Interest

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

References

- [1] Newton, I. (1952) Opticks. Dover, New York.
- [2] Ronchi, V. (1970) The Nature of Light: An Historical Survey. Harvard University Press, Cambridge.
- [3] Born, M. and Wolf, E. (1999) Principles of Optics. Harvard University Press, Cambridge.
- [4] Saleh, B.E.A. and Teich, M.C. (2007) Fundamentals of Photonics. John Wiley & Sons, Hoboken.
- [5] Read, F.H. (1980) Electromagnetic Radiation. John Wiley & Sons, Hoboken.
- [6] Jackson, J.D. (1998) Classical Electrodynamics. Wiley, Hoboken.
- [7] Planck, M. (1959) The Theory of Heat Radiation. Dover, New York.
- [8] Bransden, B.H. and Joachain, C.J. (2003) Physics of Atoms and Molecules. 2nd Edition, Longman, London.
- [9] Hertz, H. (1887) *Annalen der Physik*, **33**, 983.
<https://doi.org/10.1002/andp.18872670827>
- [10] Einstei, A. (1987) The Collected Papers of Albert Einstein. Princeton University Press, Princeton.
- [11] Compton, A.H. (1923) *Physical Review*, **22**, 409.
<https://doi.org/10.1103/PhysRev.22.409>
- [12] Wentzel, G. (1926) *Zeitschrift für Physik*, **40**, 574.
<https://doi.org/10.1007/BF01390456>

- [13] Beck, G. (1927) *Zeitschrift für Physik*, **41**, 443. <https://doi.org/10.1007/BF01400206>
- [14] Lamb, W.E. and Scully, M.O. (1969) In Polarization, matière et rayonnement, volume jubilaire en l'honneur d'Alfred Kastler. French Physical Society, Press Universitaires de France, Paris.
- [15] Klein, O. and Nishina, Y. (1929) *Zeitschrift für Physik* **52**, 853. <https://doi.org/10.1007/BF01366453>
- [16] Robinson, H.L. (1953) *Journal of Applied Physics*, **24**, 35. <https://doi.org/10.1063/1.1721127>
- [17] Hadlock, R.K. (1958) *Journal of Applied Physics*, **29**, 918. <https://doi.org/10.1063/1.1723330>
- [18] Hunter, G. and Wadlinger, R.L.P. (1989) *Physics Essays*, **2**, 2. <https://doi.org/10.4006/1.3035861>
- [19] Hunter, G. (1986) Physical Photons: Theory, Experiment and Implications. In: Honig, W.M., Kraft, D.W. and Panarella, E., Eds., *Quantum Uncertainties Recent and Future Experiments and Interpretations*, Springer, Berlin, 331-343. https://doi.org/10.1007/978-1-4684-5386-7_18
- [20] Mandel, L. (1966) *Physical Review*, **144**, 1071. <https://doi.org/10.1103/PhysRev.144.1071>
- [21] Pfleegor, R.L. and Mandel, L. (1967) *Physical Review*, **159**, 1084. <https://doi.org/10.1103/PhysRev.159.1084>
- [22] Akhiezer, A.I. and Berestetskii, B.V. (1965) *Quantum Electrodynamics*. Interscience, New York. <https://doi.org/10.1119/1.1971111>
- [23] Grynberg, G., Aspect, A. and Fabre, C. (2010) *Quantum Optics*. Cambridge University Press, Cambridge.
- [24] Auletta, G. (2001) *Foundations and Interpretation of Quantum Mechanics*. World Scientific, Singapore. https://doi.org/10.1142/9789812793560_0002
- [25] Mickelson, A.A., *et al.* (1928) *Astrophysical Journal*, **68**, 341.
- [26] Mittleman, M.H. (1982) *Introduction to the Theory of Laser-Atom Interactions*. Plenum Press, New York. <https://doi.org/10.1007/978-1-4684-1143-0>
- [27] Weissbluth, M. (1988) *Photon-Atom Interactions*. Academic Press, Inc., Cambridge.
- [28] Garrison, J.C. and Chiao, R.Y. (2008) *Quantum Optics*. Oxford University Press, Oxford. <https://doi.org/10.1093/acprof:oso/9780198508861.001.0001>
- [29] Klyshko, D.N., Penin, A.N. and Polkovnikov, B.F. (1970) *JETP Letters*, **11**, 5.
- [30] Burnham, D.C. and Weinberg, D.L. (1970) *Physical Review Letters*, **25**, 84. <https://doi.org/10.1103/PhysRevLett.25.84>
- [31] Ghose, P. and Home, D. (1996) *Foundations of Physics*, **26**, 7. <https://doi.org/10.1007/BF02148836>
- [32] Grangier, P., Roger, G. and Aspect, A. (1986) *Europhysics Letters*, **1**, 173. <https://doi.org/10.1209/0295-5075/1/4/004>
- [33] Kocher, C.A. and Commins, E.D. (1967) *Physical Review Letters*, **18**, 575-577. <https://doi.org/10.1103/PhysRevLett.18.575>
- [34] Freedman, S. and Clauser, J. (1972) *Physical Review Letters*, **28**, 938. <https://doi.org/10.1103/PhysRevLett.28.938>
- [35] Aspect, A., *et al.* (1981) *Physical Review Letters*, **47**, 460. <https://doi.org/10.1103/PhysRevLett.47.460>
- [36] Ehrenberg, W. and Siday, R.E. (1949) *Proceedings of the Physical Society. Section B*,

- 62, 8. <https://doi.org/10.1088/0370-1301/62/1/303>
- [37] Chambers, R.G. (1960) *Physical Review Letters*, **5**, 3-5.
<https://doi.org/10.1103/PhysRevLett.5.3>
- [38] Tonomura, A., *et al.* (1982) *Physical Review Letters*, **48**, 1443.
<https://doi.org/10.1103/PhysRevLett.48.1443>
- [39] Osakabe, N., *et al.* (1986) *Physical Review A*, **34**, 815.
<https://doi.org/10.1103/PhysRevA.34.815>
- [40] Ryder, L.H. (1987) *Quantum Field Theory*. Cambridge University Press, Cambridge.
- [41] Milonni, P.W. (1994) *The Quantum Vacuum*. Academic Press Inc., London.
- [42] Pozar, M.D. (2005) *Microwave Engineering*. 3rd Edition, John Wiley & Sons, Inc., Hoboken.
- [43] Sipe, J.E. (1995) *Physical Review A*, **52**, 1875.
<https://doi.org/10.1103/PhysRevA.52.1875>
- [44] Bialynicki-Birula, I. (1996) Photon Wave Function. In: Wolf, E., Ed., *Progress in Optics XXXVI*, Elsevier, Amsterdam.
- [45] Smith, B.J. and Raymer, M.G. (2007) *New Journal of Physics*, **9**, 414.
<https://doi.org/10.1088/1367-2630/9/11/414>
- [46] Khokhlov, D.L. (2010) *Applied Physics Research*, **2**, 49-54.
- [47] Cugnon, J. (2011) *Open Journal of Microphysics*, **1**, 41-45.
<https://doi.org/10.4236/ojm.2011.13008>
- [48] Chandrasekar, N. (2012) *Advanced Studies in Theoretical Physics*, **6**, 391-397.
- [49] Meis, C. (1997) *Foundations of Physics*, **27**, 865.
<https://doi.org/10.1007/BF02550345>
- [50] Meis, C. (2014) *Physics Research International*, **2014**, Article ID: 187432.
- [51] Meis, C. (2017) *Light and Vacuum*. 2nd Edition, World Scientific, Singapore.
<https://doi.org/10.1142/10450>
- [52] Meis, C. and Dahoo, P.R. (2018) *AIP Conference Proceedings*, **2040**, Article ID: 020011.
- [53] Meis, C. (2019) Quantized Field of Single Photons. In: Xia, K.Y., Ed., *Single Photon Manipulation*, IntechOpen, London, 15-36.
<https://doi.org/10.5772/intechopen.88378>
<https://www.intechopen.com/online-first/quantized-field-of-single-photons>
- [54] Meis, C. (2022) *International Journal of Applied Electromagnetics and Mechanics*, **Pre-press**, S1-S7. <https://doi.org/10.3233/JAE-220111>
- [55] Meis, C. and Dahoo, P.R. (2017) *The International Journal of Quantum Information*, **15**, Article ID: 1740003. <https://doi.org/10.1142/S0219749917400032>
- [56] Rybakov, Yu.P. and Saha, B. (1995) *Foundations of Physics*, **25**, 1723-1731.
<https://doi.org/10.1007/BF02057884>
- [57] Meis, C. and Dahoo, P.R. (2017) *Journal of Physics: Conference Series*, **936**, Article ID: 012004. <https://doi.org/10.1088/1742-6596/936/1/012004>
- [58] Meis, C. (2020) *International Journal of Modern Physics A*, **35**, Article ID: 2050153.
<https://doi.org/10.1142/S0217751X20501535>
- [59] Meis, C. (2018) *Journal of Physics: Conference Series*, **1141**, Article ID: 012072.
<https://doi.org/10.1088/1742-6596/1141/1/012072>

Inadequacies in the Current Definition of the Meter and Ramifications Affecting Einstein's Second Postulate of Relativity

Steven D. Deines

Donatech Corporation, Inc., Fairfield, IA, USA

Email: physicspublishing@gmail.com

How to cite this paper: Deines, S.D. (2023) Inadequacies in the Current Definition of the Meter and Ramifications Affecting Einstein's Second Postulate of Relativity. *Journal of Modern Physics*, 14, 330-360.

<https://doi.org/10.4236/jmp.2023.143021>

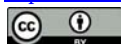
Received: December 8, 2022

Accepted: February 25, 2023

Published: February 28, 2023

Copyright © 2023 by author(s) and Scientific Research Publishing Inc. This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

The current definition of the meter as based on the time of light transmission and the postulated universal constant light speed is ill-defined and inadequate. The definition fails to identify which second is required, whether to use coordinate or proper time, or which method to construct an exact meter, besides ignoring gravity's effect. In Einstein's 1905 paper that defined special relativity, Einstein stipulated correctly that light traversing the ends of a resting rod takes equal time transmissions in either direction. If that rod is oriented parallel to a constant velocity, a photon from one end of the moving rod takes a longer time span with a universal constant light speed to overtake the receding end and takes a shorter time span to intercept the approaching end of the rod when transmitted in the opposite direction, resulting in a longer roundtrip distance of photons traversing the moving rod versus the resting rod. Length contraction undercompensates this difference. Einstein did not address this issue. However, Einstein claimed the unequal time intervals over the moving rod versus equal intervals over the resting rod are because simultaneous states for the resting observer and resting rod are nonsimultaneous for the constant moving observer. This contradicts his first postulate of relativity: any state of a physical system (*e.g.*, equal timed traverses of photons moving over a rod) is unaffected by a constant translational velocity between inertial reference frames. An in-depth analysis examines Einstein's thought experiment for an adequate redefinition. The analysis reveals one-way photon velocities obey vector velocity addition involving moving photon sources, but it proves by induction that roundtrip photon traverses have an average speed that is identical to the standard light speed c . Thus, Einstein's second postulate of relativity is not general, but is valid for roundtrip traverses of photon transmissions. This may change many physical concepts, since one-way velocities for photons and particles are not limited by the second postulate. A suggested redefinition of the meter is submitted.

Keywords

Light Speed, Time Dilation, Length Contraction, Special Relativity, Gravitation

1. History of the Definition of the Meter

The first definition of the meter was $1\text{E}-7$ of the distance from the equator to the north pole along a meridian at sea level, which was assumed to go through Paris, as the French metric system was adopted in 1791 using a decimal system [1] [2]. Unfortunately, the Earth is not a perfect oblate spheroid, and the flattening factor of the radii to the equator and the poles was not known then. The meter also varies depending on the meridian of longitude due to the lumpiness of Earth. Over decades of international research, the meter's uncertainty in surveyed length was reduced by improvements in geodesy.

This led to the Meter Convention (Convention du Mètre) of 1875, which mandated the permanent founding of the International Bureau of Weights and Measures (BIPM: Bureau International des Poids et Mesures) located in Sèvres, France. This organization was chartered to construct and preserve a meter bar as a standard, distribute duplicate bars to other nations, and compare those national standards periodically against the BIPM meter standard. Such bars were distributed in 1889 at the first General Conference on Weights and Measures (CGPM: Conférence Générale des Poids et Mesures) to participating nations to establish the international meter.

James Clerk Maxwell proposed that an emitted light ray could be used to standardize both the meter and the second [3]. This proposal was implemented in 1893 by Albert A. Michelson, who invented the interferometer and measured the meter using light from heated cadmium [4]. BIPM regularly used interferometry with the meter standard until 1960 when the 11th CGPM redefined the meter in the new International System of Units (SI). That definition of the meter was 1650763.73 wavelengths of the orange-red emission of krypton-86 in a vacuum [5].

In 1983, the 17th CGPM redefined the meter in the current terms of the second and the postulated constant speed of light in a vacuum. The current definition of the meter is [6]:

The metre is the length of the path travelled by light in vacuum during a time interval of $1/299,792,458$ of a second.

CGPM recommended a laboratory utilize helium-neon lasers to measure the meter. If conducting the test with air, the refractive index n is used to convert the unit of wavelength λ to meters using c , the fixed speed of light in a vacuum in m/s and f as the measured frequency of the source utilizing the equation, $\lambda = c/nf$ [7].

The BIPM issued a clarification of the meter [8] in 2002:

Its definition, therefore, applies only within a spatial extent sufficiently small that the effects of the nonuniformity of the gravitational field can be ignored (note that, at the surface of the Earth, this effect in the vertical direction is about 1 part in $1E16$ per meter). In this case, the effects to be taken into account are those of special relativity only.

Initially, this would appear to be possible by calibrating two clocks together by synchronizing to a master time standard to replicate the second and then move one clock about one meter away to construct an exact meter in the laboratory. However, special relativity predicts time dilation with the moved clock so that the two clocks are no longer synchronized to define a meter with a one-way transmission of the emitted photons. So, a roundtrip using photon transmissions is performed so that only one clock is used to time the emission and reception after the reflected photons arrive at the source. The meter's definition lacks this restrictive condition.

2. Unstated Issues Affecting the Definition of the Meter

According to the BIPM clarification, only special relativity effects will affect the determination of the meter if measured at Earth's surface. Special relativity has two types of time—coordinate time and proper time. In Einstein's 1905 paper [9] [10], coordinate time is constructed by having stationary perfect clocks calibrated to a master clock by recording the roundtrip transmission between the master clock and each remote clock and halving that transmission interval to be added to the next time record broadcasted from the master clock to the remote clock. This is Einstein's synchronization method. Any event in the neighborhood of a fixed coordinate location of a calibrated remote clock has the coordinate time at that instant. Effectively, coordinate time is time transmitted by the master clock anywhere in the region of interest provided light speed was infinite instead of finite.

Proper time is the special relativity effect that a perfect clock exhibits due to time dilation caused by the velocity of that clock relative to the resting frame of reference. The problem is a proper second of a moving clock will be slightly longer than the second of the coordinate clock. Unfortunately, the definition of the meter makes no mention of the preferred inertial reference frame, so that the time dilation of the laboratory's proper clock compared to the coordinate clock of the inertial resting frame is nebulous. Many inertial frames of reference are possible: Earth's nonrotational center of mass frame, the nonrotational Earth-Moon barycenter frame, the solar nonrotational barycenter frame, etc., which are needed to determine the velocity of the laboratory relative to that frame of choice to compensate for proper time.

The definition of the meter does not stipulate which second from any time-scale should be used. The original second, which is $1/86,400$ of a mean solar day, is based on Earth's rotation relative to the Sun, which its elliptical orbit changes the length between contiguous solar days. A tropical year is defined by the mean

motion of the Sun's longitude relative to the dynamical equinox set by one complete revolution in Earth's orbit. The calendar year closely follows the tropical year. The original timescale was often called mean solar time or Greenwich Mean Time (GMT) based on meridian transits of stars as the Earth rotated. The second timescale based on the tropical year was Ephemeris Time (ET) and has now been renamed Universal Time (UT1). The International Atomic Time (TAI) is based on the Système International (SI) second, which is defined by the calibration between the US Naval Observatory (USNO) and the National Physical Laboratory (NPL) performed between 1955-1958 [11] by comparing ET from the lunar ephemeris to an atomic clock. The UT1 second is slightly longer than the SI (or atomic) second as evidenced by the numerous positive leap seconds inserted into the Coordinated Universal Time (UTC) timescale, which uses the SI second of TAI, so that UTC is retarded to remain within 0.7 seconds of UT1. Atomic clocks are frequency generators which count cycles that are converted into equivalent SI seconds to function as a clock.

Getting a precise SI second via synchronization of a tester's atomic clock with an ultraprecise timing institute is difficult as no master TAI standard exists. BIPM obtains a theoretical EAL (Échelle Atomique Libre or free atomic time scale) from weighting the times of all reported atomic clocks in the participating timing institutes, which each periodically sends its UTC to BIPM and which, in turn, BIPM creates a weighted average timescale. With frequency adjustments by monitoring the past weighted performance and uncertainty of individual clocks, TAI is obtained from EAL. BIPM then reports periodically the time offset and the second's rate of change for each laboratory's atomic clocks compared to the theoretical TAI timescale, which each timing facility is free to steer its clocks accordingly [12] [13].

Gravitational effects on time are missing within the current definition of the meter. In geodesy, the reference surface to measure variations in gravity is the geoid (*e.g.*, mean sea level). The Pound-Rebka experiment [14] measured the change in frequency of photons moving through gravity over a 22.5 m height. Gamma ray photons changed energy, δE , obtaining $\delta E/E = 2.5 \times 10^{-15}$, which directly affects frequency f accordingly as photon energy $E = hf$ where h is Boltzmann's constant. For every meter change in increasing altitude, the frequency change is 1.1×10^{-16} that makes the same change in a shorter second of an atomic clock. Some ultraprecise time institutes could detect a 1 m altitude change. When recalibrating the meter by the definition in a laboratory, elevating the test apparatus by a meter can shorten the new meter by 1.1×10^{-7} nm when ignoring the effects of gravity. Elsewhere, it is necessary to choose a reference, such as the geoid where a meter is defined, so that the time interval can be adjusted for very significant gravity changes such as other planets or outer space.

Lastly, when technology has been improved significantly since the adoption of this definition in 1983, some light speed experiments should obtain more precision with more significant figures. Unfortunately, any possible test would be

doomed if it obtained a different result than the “exact” value, because critics could claim something was uncompensated in the test results, making the test invalid. Sometimes, even “exact” concepts or values in definitions are wrong. For example, the nautical mile is now defined to be exactly 1852 meters. However, the nautical mile was originally defined as 1 arcminute along the equatorial circle. The Global Positioning System (GPS) in its WGS-84 model and the International Union of Geodesy and Geophysics (IUGG) in its Geodetic Reference System 1980 have the same value of 6,378,137 m for the equatorial radius of Earth [15] [16]. To seven significant figures, the nautical mile is 1855.324 m, which means the exact value in the CODATA standard is off by 3.324 m. Even “exact” definitions need regular scrutiny for accuracy.

3. Timed Measurements Differing from Fixed Distances between Source and Receptor

The current definition of the meter using timed transmissions of photons does not exclude any setup between the emission source and the light receptor. There are many classes of test setups using photon transmissions that can affect the measurements. For example, the gravity field could fluctuate in the laboratory, the accelerated laboratory could gyrate with various motions, the distance between the emitter and receptor could vary during photon transmission, etc. One simple class is a turntable with a constant angular rotation oriented perpendicular to the local gravity. Even within this class, there are infinite possible choices for parameters. This analysis will assume the radius of the turntable is 0.5 meters where both the photon emitter and absorber are fixed on the perimeter of the turntable. Before any measurements are taken in this theoretical experiment, the following definitions are given.

For all points within an acceptably small neighborhood, the coordinate time is given by a perfect clock in that neighborhood, which has been synchronized to a master time standard or timescale where the distance between these clocks remains fixed. Synchronization could be accomplished by Einstein’s synchronization method by timing the roundtrip transmission between the clock and the master time standard, divide the transmission interval by two, and add that result to the time tag of any later transmission from the master time standard. Essentially, coordinate time would be identical to the master time standard’s broadcast if electromagnetic transmissions had an infinite speed rather than the finite (although very high) speed for the standard speed of light, c , in a vacuum.

Two or more phenomena are simultaneous if they overlap, merge, collide, intersect, etc., or divide, separate, split, etc., at one point at one instant of time. No time dilation exists for an instant of time, and no length contraction is possible for a single point. Nonsimultaneous phenomena occur when they arrive at or emit from the same point at different coordinate times. Simultaneous events affect all phenomena at that location at that instant of time.

Two or more phenomena are synchronized when they occur at separate coordinate locations at the same coordinate time. Synchronization is not to be confused with simultaneity. Often synchronized phenomena are characterized with duplicate actions. For example, synchronized competitions have multiple athletes performing the same actions at separate locations, but any athlete can stop without affecting the other athletes maintaining the synchronization in the competition.

Often, simultaneous is incorrectly stated when events or actions occur at the same coordinate time at different locations. Simultaneity can not be applied on two clocks to maintain the same timescale of coordinate time. Synchronization is used to adjust the two clocks to keep the same coordinate time. It is also obvious such clocks occupy separate locations for synchronization. Two clocks indicating the same time are operating in synchronization. One never states two clocks are simultaneous. Even in common English, simultaneity and simultaneous actions occur at one instant of coordinate time and location. For example, two cars collide in an intersection, which is the result of two cars entering the intersection (same location and same time) simultaneously. Car collisions are not synchronized events. It is incorrect to state two separated events occurring at the same time are simultaneous. When an observer detects two inputs or events arriving at the observer's location at the same time, that constitutes simultaneous detection through sight, sound or touch.

Consider the turntable in **Figure 1** to be tested in a laboratory with the photon emitter located at B and the photon absorber at A . The plane of the turntable is perpendicular to the local gravity in the laboratory. In the first test, the turntable is rotating counterclockwise. The emitter broadcasts photons omnidirectionally, but only the photons traveling along the chord BC are intercepted by the absorber, which travels the arc from A to C during the photon transmission. The movement of the emitter during transmission is extraneous and not included in **Figure 1**. A perpendicular bisector splits the chord AC in two, creating two identical right triangles, CMO and AMO , which make angles ϕ and θ complementary. The angle AOB is π radians, which makes the supplementary angle, $\Phi = 2\phi$. The arc AC equals $2\theta R = R\omega\Delta t$, where ω is the angular velocity and Δt is the transmission interval from B to C . L is the length of the chord BC , so $\Delta t = L/c$. For simplicity, assume the laboratory houses the platform in a vacuum, making the index of refraction, n , equal to one. Thus, $\theta = \omega L/2c$, and $R \cos(\theta) = L/2 = R \sin(\phi)$. This example shows that the distance between the emitter and absorber is always one meter, but the measurement based on the photon transmission interval multiplied by the photon speed of c is always less than a meter, depending on the value of the angular velocity, ω .

In the second test, the absorber is located at C , and the turntable rotates clockwise in **Figure 1**. In this case, ω is chosen so that the absorber, initially at C , acquires the photons at A , which is directly across point B , where the photon emitter initially had broadcasted photons to the absorber. Thus, $\omega = 2\theta c/L$ with

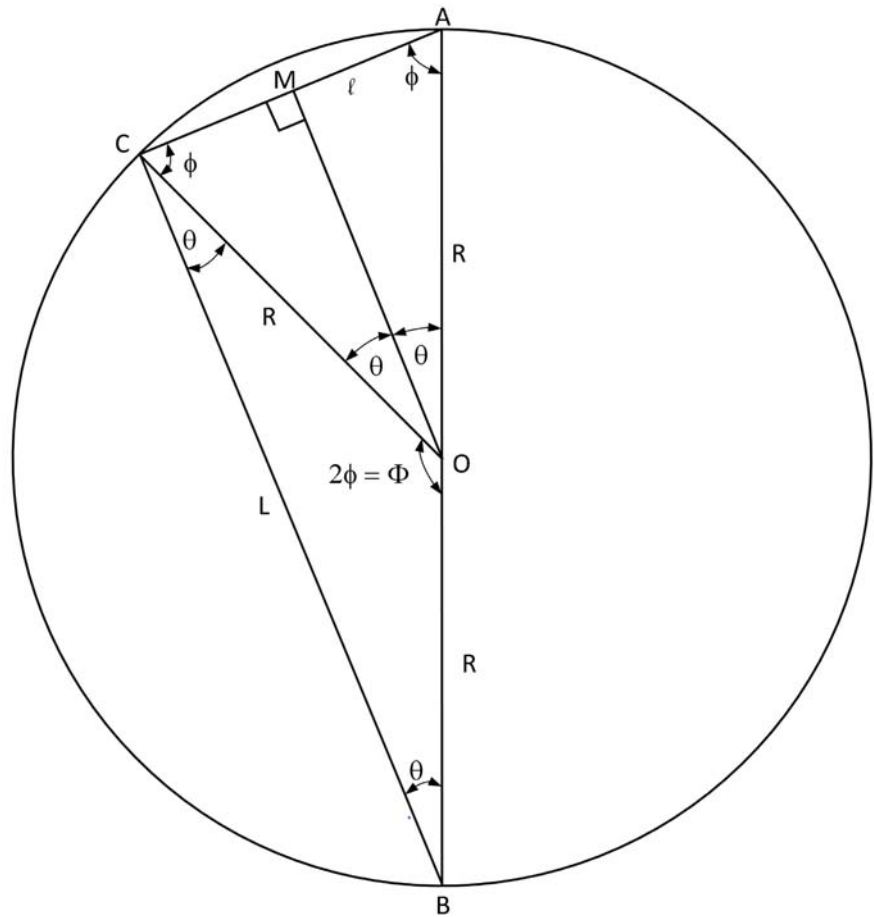


Figure 1. Turntable geometry using photons.

the added requirement that $\cos(\theta) = L/2R$. In this case, the fixed distance between the emitter and absorber is L , but the measurement from photon transmission is the longer distance of $2R$. These two examples illustrate that the current definition of the meter using timed intervals of photon transmission will not replicate the actual fixed distance between the photon emitter and absorber.

This analysis of the turntable only addressed the distances between the two devices. It does allow a comparison between coordinate time and proper time to measure a timed interval. Embellish the experiment with two recorders with antennas serving as observers. One recorder is stationary in the laboratory, and the second recorder is directly above point O . To be technically correct, the second recorder could be rotating to be fixed relative to the turntable, but an omnidirectional antenna rotating on the axis of rotation is equivalent to being fixed in the laboratory. The laboratory has a precise master time standard that broadcasts coordinate time from an antenna located at point O on the turntable. Many designs are possible. Equivalently, the broadcast must consist of a changing signal (*e.g.*, a step function whose vertical slopes denote one ns advances of the master clock) and an accompanying message that denotes the coordinate time tag for each change in the timing signal (*e.g.*, a complete set of messages that denote all

the significant figures of the time associated to each ns in the signal). With the foreknowledge that the antenna is 0.5 m away, both the photon emitter and absorber can adjust the time tag at reception by adding a time span of $0.5/c$ to the received coordinate time message so that coordinate time at reception is identical with the master clock.

An atomic clock is integrated with the emitter, and another with the absorber. Each atomic clock is synchronized to the master time standard using Einstein's synchronization method in inertial conditions before the turntable begins rotation. The coordinate time and proper time for each device is identical while the turntable is stationary relative to the laboratory. This ensures the embedded clocks are operationally synchronized.

The turntable can be modified so that the emitter releases photons only once per rotation of the turntable undergoing a constant angular velocity. This can be done, either electronically or mechanically, at some contact point when the emitter reaches position B to release photons omnidirectionally. Simultaneously, when the emitter detects the contact point, it generates a pulse of photons and a message in the coordinate time (as interpolated between the ns pulses with the adjusted coordinate messages from point O) and proper time of the embedded clock. When the absorber receives photons later from the emitter, it generates a message of reception in coordinate time (also interpolated between the ns pulses with adjusted coordinate messages) and its proper time at reception.

When duplicate observers concurrently record data from an event or phenomenon, the results must be identical, especially if all data are transformed to a common reference frame. If theory predicts different results for each observer concurrently witnessing an event, then the theory must be revised, either due to an erroneous assumption, an error in the derivation, or a misinterpretation of the event or phenomenon. In all cases, the theory must be self-consistent.

According to special relativity, the clocks on the moving turntable undergo a constant tangential speed of $\omega R = v$, but both moving clocks remain synchronized as they both move at the identical tangential speed v relative to the laboratory and its master clock. According to special relativity, the proper second of the moving clocks is dilated by the factor $\gamma = (1 - v^2/c^2)^{-1/2}$ compared to the master standard second. The broadcast coordinate time interval, Δt , between emission and absorption is larger numerically than the broadcasted proper time interval, or $\Delta t/\gamma$ for counted units.

For example, suppose the reader's wristwatch is slower by half than the standard clock. If the actual time elapsed was two seconds, the slow wristwatch counted only one second. If the reader moves at 1 m/s, the calculation predicts a distance of 1 meter traversed using the wristwatch, but actually two standard seconds in coordinate time elapsed for a traveled distance of 2 meters.

The emitter transmits one message at the instant of the photon emission to both recorders in the test setup. The absorber does the same upon photon absorption. The contents of each message do not change, no matter how fast or slow the messages arrived at the recorders, much like receiving letters in one's

mail.

Suppose the turntable rotated faster to approach the limit that the arc AC was as long as the chord BC (*i.e.*, the tangential speed approached the limit of c). In the limit, $L = R\omega t = R\omega L/c$, so that $\omega = c/R$. The elapsed coordinate time between emitter and receiver would be the interval, L/c . There would be no elapsed proper time between emitter and absorber as the interval would be infinitely long as $v \rightarrow c$ and $\gamma \rightarrow \infty$. In this case, the emitter and absorber would have identical proper time tags, implying that the absorber received photons at B , but photons were transmitted from B . The theory is not self-consistent for distances via proper time.

This theoretical test shows time measurements may not obtain the actual fixed distance between an emitting source and absorbing target. In the first example, the distance between the devices is set at one meter (*i.e.*, $2R$ in [Figure 1](#)), but the measured distance is less than a meter (L in [Figure 1](#)) as the absorber rotated slightly during photon transmission. In the second case, the distance between the devices is fixed to be less than a meter, but the turntable rotates at the needed angular velocity for the measured time interval to predict exactly one meter distance. It was also discussed that moving clocks on the turntable will have slower proper time tags that will underpredict what the coordinate time tags would predict for distance between the devices. These examples demonstrate that the length between the source and detector was maintained during the measurement, but the velocities of each changed relative to each other, so that the measurement did not replicate the actual distance. The current meter definition allows any setup of precise measurements that produces the wrong meter.

Some may advocate that the current definition of the meter be limited to inertial cases, so that the examples given in this section would be excluded. The problem still remains that the theory does not produce the same results when using coordinate time versus proper time.

4. Einstein's Rod with Traversing Light Beams

In 1873, Maxwell advocated that the transmission of light would serve as a definition of length [3]. Einstein was aware that the electromagnetic equations do not accommodate vector velocity addition, which is a key property in Newtonian mechanics. He discussed this in his 1905 paper "On the Electrodynamics of Moving Bodies" [9] [10], which he established his two postulates for developing his special relativity theory.

To avoid lengthy translations from Einstein's German into English, the author chose the translation of Arthur I. Miller, who was a professor of physics at Harvard and Lowell Universities and who later transitioned as a historian of science due to his keen interest in Einstein's special relativity paper. In his book [Albert Einstein's Special Theory of Relativity](#), he collected and translated many letters from Einstein and his contemporaries and translated Einstein's 1905 paper. The

book's appendix contains his translation of that 1905 paper, which differs from previous (and, in places, unacceptable) English translations. Typographical errors in the original *Annalen* version are flagged, which went into the Teubner edition, and additional errors that appear in the Dover reprinted volume The Principle of Relativity. Einstein's and Sommerfeld's footnotes are annotated. This appendix will be the English translation with demarcations for sections, §, and line numbers within each section (a new section resets line numbering to 1).

Einstein's 1905 relativity paper does not cite any reference to other papers. Instead, it gives a logical discourse based on experiences, which could mislead. He considered simultaneous events occur with a time associated at a single location. Einstein's example was a train arrived at 7 o'clock, which he meant the small hand of his watch pointed at 7 as a train pulled in ([10], §1, lines 13-17). Einstein footnoted there is an inexactitude of two events occurring at (approximately) the same place. Note he did not define if the watch was a master time standard, or how horologists do synchronize mechanical time pieces to a time standard and monitor how that master time standard defines time for Earth. He admitted that an observer with a clock will not determine time independently of a remote event communicated to the observer by light ([10], §1, lines 24-32).

Einstein considered a clock A at position A and an identical clock B at location B , but there was no way to connect " A time" to " B time" with a common time for locations A and B ([10], §1, lines 33-41). He stated, "The latter time can now be defined by requiring that *by definition* the 'time' necessary for light to travel from A to B be identical to the 'time' necessary to travel from B to A . Let a ray of light start at the ' A time' t_A from A toward B , let it at the ' B time' t_B be reflected at B in the direction of A , and arrive again at A at the ' A time' t'_A ." ([10], §1, lines 41-45) He concluded that the two clocks run in synchronization if ([10], §1, Equation (1.1))

$$t_B - t_A = t'_A - t_B$$

Note this involves one-way light transmissions. Einstein assumed this definition of synchronization was free of any contradictions. He further claimed that (1) if a clock at B synchronizes with the clock at A , the clock at A synchronizes with the clock at B , and (2) if the clock at A synchronizes with the clock at B and also the clock at C , then the clocks at B and C synchronize with each other ([10], §1, lines 48-60). He added the requirement ([10], §1, Equation (1.2)) that for length AB

$$\frac{2AB}{t'_A - t_A} = c$$

where c is the universal constant for light speed. Equation (1.2) is Einstein's synchronization procedure between two clocks. Note this time interval is a roundtrip transmission, and one divides the roundtrip interval by 2 and then adds this time span to the next broadcast of a master clock's time to set the remote clock's time tag for synchronization. He stated, "The 'time' of an event is

the reading simultaneous with the event of a clock at rest and located at the position of the event, this clock being synchronous, and indeed synchronous for all time determinations, with a specified clock at rest.” ([10], §1, lines 57-60) Although he did not state it, other physicists considered this to be coordinate time throughout the reference frame. In any case, the time t is the time recorded by stationary synchronized clocks throughout the resting system.

In Section 2, Einstein defined his two postulates of special relativity ([10], §2, lines 1-12). The translation lists

1) “The laws of which the states of physical systems undergo changes are independent of whether these changes of state are referred to one or the other of two coordinate systems moving relatively to each other in uniform translational motion.” and,

2) “Any ray of light moves in the ‘resting’ coordinate system with the definite velocity c , which is independent of whether the ray was emitted by a resting or by a moving body.”

He added, “Consequently,

$$\text{velocity} = \frac{\text{light path}}{\text{time interval}}$$

where time interval is to be understood in the sense of the definition in §1”. Technically, postulate (1) is not universal for all equations of force. Newtonian forces are the derivatives of momentum, which a constant velocity $\mathbf{v}$ results in zero force for any inertial reference frame (those frames that translate linearly by a constant velocity). The Newtonian force equations remain the same for all inertial frames. This is not true for electromagnetic forces. In particular, the general Lorentz force $\mathbf{F}$ is

$$\mathbf{F} = q\mathbf{E} + q\mathbf{v} \times \mathbf{B}$$

where $\mathbf{E}$ is the electric field intensity, $\mathbf{B}$ is the magnetic field induction, q is the charge of the particle, and $\mathbf{v}$ is the velocity of the charged particle in the reference frame. For an observer fixed relative to the charges, only the electric term is detected, especially in electrostatic conditions where no magnetic field appears. If that observer is fixed to the reference frame, both electric and magnetic contributions are detected that result in a different observed Lorentz force. If a second observer is moving at a constant velocity $\mathbf{V}$ relative to the reference frame, that observer records a dissimilar Lorentz force that is $q\mathbf{E} + q(\mathbf{v} + \mathbf{V}) \times \mathbf{B}$. In this case, the observed state of the charged particle does depend on the choice of the reference frame, even if it has a constant translational velocity relative to a “resting” frame. Lorentz published the complete derivation for the Lorentz force in 1895 [17] by adding the electric force to Heaviside’s earlier identification of the magnetic force to a charged particle [18]. In theory, one could use the Lorentz force to find the relative velocity between two inertial frames.

Einstein considered that a rigid rod at rest of length L is measured by a measuring rod that is also at rest. The observer would conclude the rigid rod was of

length L . He then considered moving this rigid rod at a velocity $\mathbf{v}$ and orienting it parallel to its velocity. The observer of the moving rod would ascertain at time t using the stationary, synchronized clocks of the resting system where the ends of the rod would be located and then measure the distance between the predicted two locations. He predicted this distance would not be L . ([10], §2, lines 13-35) He then added two clocks to the two ends, A and B of the rod, that were synchronized with the clocks of the resting system. He further added two moving observers, each fixed to each moving clock. He stated, “Let a ray of light depart from A at the time t_A , let it be reflected at B at the time t_B , and reach A again at the time t'_A . Taking into consideration the principle of the constancy of the velocity of light we find that

$$t_B - t_A = \frac{r_{AB}}{c - v} \text{ and } t'_A - t_B = \frac{r_{AB}}{c + v}$$

where r_{AB} denotes the length of the moving rod—measured in the resting system. Observers moving with the moving rod would thus find that the two clocks were not synchronous, while observers in the resting system would declare the clocks to be synchronous.” ([10], §2, lines 47-54)

Einstein gave no derivation for the above equations. It is obvious his result does not support his Equation (1.1) that requires equal time intervals for both light transmissions between endpoints A and B . His claim is contradictory that stationary observers agree the clocks are synchronized versus moving observers arguing the clocks A and B are not synchronized. All resting clocks were synchronized initially. The two clocks at the rod's ends at A and B were synchronized to the resting clocks. There is no need for the moving observers to resynchronize the clocks attached to A and B , which Einstein stated the moving observers performed again ([10], §2, lines 45-47). According to Einstein's lemma (2), clock A is synchronized to the resting clocks and clock B is synchronized to the same resting clocks so that clock A is synchronized with clock B . When clocks A and B move, time dilation is the same for both clocks, so these two clocks still remain synchronized to each other even in the moving frame. The time t is from the resting clock time, which is the coordinate time of the resting system, and t is the time listed in the above equations. At time t , endpoint A is at the location of one resting clock, which broadcasts its location with its time. Likewise, endpoint B is at its location of another resting clock, which broadcasts its location with its time t , which is also the same t as the first resting clock where endpoint A happens to be at this instant. The two messages from endpoints A and B are the same for both stationary and moving observers (*i.e.*, a single coordinate point at instantaneous time t), so all observers have the two resting locations at the resting time t , which must be identical between observers. An inconsistent theory reveals some error exists, whether it is an error in a derivation, a false assumption, or misinterpretation of the experimental results. The next section will give an in-depth examination of the photons' path in Einstein's thought experiment.

5. Light Paths between Ends of Stationary and Moving Rods

Einstein's equation that follows his second postulate is equivalent to

$$\text{time interval} = \frac{\text{light path}}{\text{velocity}}.$$

In quantum mechanics, photons emitted from endpoint A combine with the reflective atoms at endpoint B to create excited atoms. New photons are emitted from those excited atoms at B to endpoint A , so that virtually all photons from B follow Snell's law to A . In the roundtrip test, one beam traverses the distance $A \rightarrow B$ and the other beam $B \rightarrow A$.

As Einstein stated, orient a uniformly moving rod with its length parallel to its velocity, v , relative to the resting frame. When the rod is resting or stationary, the photons traverse a distance of L , the length of the rod. The roundtrip distance for the photons over the resting rod is $2L$, and each time span across is L/c .

However, photons traveling with or opposite to the moving rod experience different distances. For example, runners know that a footrace covers more (or less) ground if the finish line is moved away from (or toward) the runners during the race. This happens also in Zeno's paradox of Achilles and the tortoise. Zeno's paradox relates how the tortoise challenged Achilles to a race with a head start of L , claiming the tortoise would win. Achilles conceded the race without running due to the tortoise's logic that Achilles was always behind the tortoise for an infinite number of time intervals, despite that Achilles was faster than the tortoise. Zeno did not consider that a finite length can be divided into an infinite number of lengths, but the sum of those infinite spans would still be a finite length. Algebra can directly solve the distances that photons traverse to overtake the opposite end of the moving rod. In general, the photon's speed from endpoint A to endpoint B is c_{AB} , and the photon's speed from endpoint B to endpoint A is c_{BA} . In the resting frame, the rod moves with velocity v so that the endpoint B is at B' where the photons from A overtake the receding endpoint at B' in the time span of Δt_{AB} as shown in **Figure 2**. Also, the newly emitted photons from endpoint B intercept the approaching endpoint A at the position A' in the time span of Δt_{BA} as shown in **Figure 3**.

Solve for D in $D/v = (L + D)/c_{AB}$ and replace D in $L_{AB} = L + D$.

Solve for d in $d/v = (L - d)/c_{BA}$ and replace d in $L_{BA} = L - d$. The resulting distances in the resting frame are:

$$L_{AB} = \frac{Lc_{AB}}{c_{AB} - v}, \text{ and} \quad (1)$$

$$L_{BA} = \frac{Lc_{BA}}{c_{BA} + v} \quad (2)$$

It is immediately apparent that $L_{AB} > L$ if $c_{AB} > v$, and $L_{BA} < L$ if $v > 0$. Moreover, if the individual one-way speed of the photon is the standard c , the roundtrip distance is greater than $2L$. Simply add Equations (1) and (2) with the standard c for the roundtrip distance:

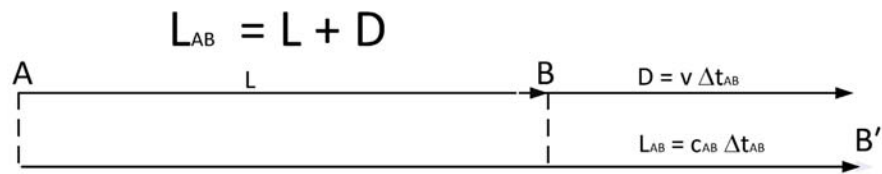


Figure 2. Photons overtaking receding endpoint.

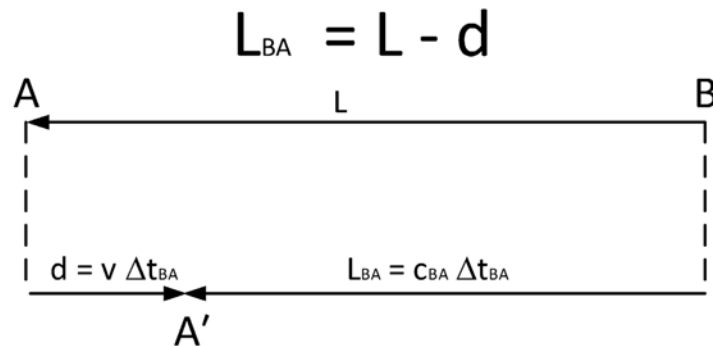


Figure 3. Photons intercepting approaching endpoint.

$$\frac{Lc}{c-v} + \frac{Lc}{c+v} = \frac{2Lc^2}{c^2-v^2} = \frac{2L}{1-\frac{v^2}{c^2}} = 2L\gamma^2 > 2L \quad (3)$$

With a universal c , the photons have a longer roundtrip distance in the resting frame for the uniformly moving rod than if that rod was stationary. Even length contraction from special relativity can not resolve this problem, as length contraction is γL . If one divides the one-way distances L_{AB} and L_{BA} by their respective photon speeds, unequal time spans to intercept the opposite ends of the moving rod occur for the stationary observer while the observer fixed with the moving rod records equal time spans. Einstein correctly required equal time spans, $\Delta t_{AB} = \Delta t_{BA}$, in his Equation (1.1). All observers must concurrently record equal time spans over opposite traverses as mandated by Einstein's first postulate when witnessing the same event. This contradiction implies the theory is not consistent.

Equations (1) and (2) give the actual distances in the resting frame that photons emitted from A or from B travel to the opposite end of the uniformly moving rod. The time to complete a one-way traverse is to divide the respective one-way photon speed into each distance. Since the resting observer and uniformly moving observer must record the same one-way time span, equate their Δt_{AB} time intervals and, next, their Δt_{BA} time spans.

$$\frac{L_{AB}}{c_{AB}} = \frac{L}{c_{AB} - v} = \frac{L}{c} \Rightarrow c_{AB} = c + v, \text{ and} \quad (4)$$

$$\frac{L_{BA}}{c_{BA}} = \frac{L}{c_{BA} + v} = \frac{L}{c} \Rightarrow c_{BA} = c - v \quad (5)$$

Equations (4) and (5) show that photon speeds obey vector velocity addition

for moving transmitters. The emitted photons from endpoint A obtain the additional $+v$ in the one-way velocity to overtake endpoint B . The second beam of ejected photons from endpoint B after absorption gets the opposite one-way velocity, $-v$, in the antiparallel direction of the uniformly moving rod.

Einstein's equation (1.2) applies to roundtrip transmissions of photons, which agrees with our experiences when the distance between A and B remain fixed during the roundtrip transmissions, yet Equation (1.1) was required for one-way photon transmissions over identical distances. Photon velocities obey vector velocity addition in one-way transmissions. Roundtrip transmissions have an average speed equal to the standard constant speed c as $([c+v] + [c-v])/2 = c$, as the roundtrip time interval always remains the same. When Einstein replaced any c_{AB} or c_{BA} with c , he essentially was assuming what he was proving before getting the result. He did uncover the mathematical contradiction, but he interpreted that as simultaneity versus nonsimultaneity between observers. However, the theory needs revision.

Einstein's second postulate is an excellent approximation due to the high speed of photons, but it is not exact in general. For example, his last section of his 1905 paper predicted that masses measured longitudinally versus transversely along the electron's velocity will be different ([10], §10, lines 1-81). He mathematically derived the dynamics of the slowly accelerated electron, which resulted in a mathematical difference in longitudinal and transverse mass measurements. No one has published any counter derivation showing any mathematical error in his paper. No national institute of standards has published any finding that a test mass has a diurnal variation, which is expected as Earth's rotation would displace the laboratory's local vertical for a hanging test mass with a component parallel or perpendicular to Earth's orbital velocity. With no known experimental verification of a different mass measurement depending on the direction of a moving mass and no error in the derivation, one must question if one or both of Einstein's postulates are not precisely true.

Arthur I. Miller attempted to show how Einstein could have derived his claim that events simultaneous in one inertial frame are not simultaneous in another inertial frame ([10], p. 264-265). Miller considered a rigid rod at rest in a resting frame that had length L_0 as congruently measured with a "ruler" by an observer in the resting frame. This is a nonsimultaneous activity with end A compared to the "ruler" at one time and another comparison of end B to the stationary "ruler" at another time to get a length of L_0 . Einstein defined such a test of the moving rod to have a constant positive velocity v moving in the direction of increasing x and with the rod oriented parallel to the resting frame's x axis. If end A is where the moving rod is at x_1 and end B is at x_2 , let $x_1 < x_2$ without loss of generality (if not, swap the subscripts and proceed). Let the resting frame's time t at its origin (*i.e.*, $x = 0$) be transformed to the moving frame's time τ by $\tau = \gamma t$. With $v < c$ in the first frame, length contraction still maintains the moving length $\zeta_2 - \zeta_1 > 0$ as the resting length $x_2 - x_1 > 0$. Miller had two observers fixed in the moving frame get the ζ coordinates of ends A and B of the rod at the same

coordinate time of the moving frame. Miller stated the two spatially separated observers mark the position of the ends of the moving rod simultaneously, which is incorrect as the observers are marking synchronously in τ coordinate time.

Miller derived his Equation (C) shown below and inserted it into another transformation equation to get the length contraction equation. He assumed $t_1 = t_2$ to get:

$$\tau_2 - \tau_1 = -\frac{v}{c^2}(\zeta_2 - \zeta_1) \quad (C)$$

No mathematical fallacy was derived, such as division by zero, but it is a physical fallacy as the time differences and spatial differences are individually positive. Miller's equation has the right side negative compared to the left side that is positive. All transformation equations in Einstein's 1905 paper are derived by Einstein from his second postulate of relativity, which Miller's derivation indicates a contradiction that shows Einstein's postulate is not technically or completely correct. One resolution of the apparent contradiction is that the lengths and times are in the moving frame 2 attached to the moving rod. Then, the velocity of the moving rod in frame 2 is zero, making the case that $\tau_1 = \tau_2$, or that an instant of coordinate time in frame 1 is an instant of coordinate time in frame 2, which now contradicts Einstein's claim of simultaneity versus non-simultaneity between frames. Either way, Miller's derivation contradicts the second postulate or Einstein's simultaneity claim.

One can test directly if Einstein's second postulate is totally correct in one direction. According to special relativity, length contraction is $\Delta L_{\text{resting}} = \gamma \Delta L_{\text{moving}}$ and time dilation is $\gamma \Delta \tau_{\text{resting}} = \Delta \tau_{\text{moving}}$. Let the integer $k = 299,792,458$, $\Delta L_{\text{resting}} = 1$ meter, $\Delta \tau_{\text{resting}} = 1$ second. As the speed of light is currently defined:

$$c = k \frac{\text{meters}}{\text{second}} = k \frac{\gamma \Delta L_{\text{moving}}}{\Delta \tau_{\text{moving}} / \gamma} = k \gamma^2 \frac{\Delta L_{\text{moving}}}{\Delta \tau_{\text{moving}}} = c' \quad (6)$$

This means that the speed of light in a constant moving inertial frame has shorter meters and longer second intervals than a resting frame such that one-way light speeds do not have the same universal constant as $k \neq k\gamma^2$. For example, the orbital speed of Earth at apogee and perigee is respectively about 29,300 and 30,300 m/s, which the numerical difference of $k\gamma^2$ would be about 0.199 between these orbital locations. Is Einstein's second postulate always correct and is BIPM's mandate correct that light speed must be exactly 299,792,458 m/s, or is special relativity correct that light speed will have an annual variation on Earth due to its orbital speeds?

Equation (6) demonstrates that special relativity maintains a variable light speed depending on a moving inertial frame's speed relative to a "resting" reference frame. This finding contradicts the second postulate.

A simple test can demonstrate if photons obey vector velocity addition. Set a laser to point horizontally at a partially silvered mirror that is angled at 45 degrees relative to the local plumb line. The reflected beam is aimed vertically to a

hemisphere mirror that is a distance, d , of about 10 meters above the partially silvered mirror, so that the hemisphere is centered along the line of transmission. The vertical beam is reflected from the hemisphere of 2 cm radius, r , to the partially silvered mirror below, and some light is transmitted through it to the floor below. Observe if the impact point varies over time as Earth rotates. The hemisphere mirror has a sideways displacement due to the velocity $\mathbf{v}$ of Earth around the Sun, forcing the beam to miss the nadir of the hemisphere mirror by an angle of $\theta \approx \sin(\theta) = v\Delta t/r = vd/(cr)$ in radians. The reflected beam from the hemisphere mirror should touch the floor by a maximum of $d\theta = 50$ cm from the plumb line assuming a maximum $v \approx 30,000$ m/s from Earth's orbit. Even assuming a perpendicular projection of Earth's velocity to the 10 m arm to be a slower speed of 15,000 m/s, this would still cause a 25 cm displacement from the plumb line. If no observed displacement occurs over a day, then the beam is constantly touching the nadir of the hemisphere mirror. This would prove photons obey vector addition of velocity as the beam changes direction and magnitude to remain on the nadir due to the changing horizontal displacement from Earth's orbital velocity (*i.e.*, $\mathbf{v}_{\perp}\Delta t$).

The laboratory setup on Earth's surface for this test is in a noninertial frame. All tides show a surface laboratory is noninertial as the walls rotate due to Earth's spin. When high precision is not needed, Newton's equations of motion are satisfactory. For high precision, artificial force terms need to be included in the theory to account for the frame's movement. For example, a Foucault pendulum rotates its plane from Earth's spinning where the rate of change in a vector relative to space (*i.e.*, fixed stars) is equal to the rate of change of that vector in the body frame plus the cross product of Earth's rotation rate with the vector (the last product is related to the Coriolis force).

The Coriolis force is too small compared to photon speed, and the freely falling Earth centered frame does not include Earth's orbital velocity. The worst deviation of the Coriolis effect is on the equator, which has a tangential speed of 465 m/s. Orient a one-meter leg longitudinally, and Earth's rotation would displace the far end 1.55 μm . The hypotenuse that the photon travels is longer than 1 m by $1.20\text{E}-12$ m. When technology measures light speed with more than 12 significant figures, the Coriolis force will affect a measured speed.

The next freely falling frame is the center of mass for the Earth-Moon system, denoted here as the E-M barycenter. Earth's rotational axis is tilted 23.44° from the ecliptic plane, and the E-M barycenter is located approximately 3/4 radially from Earth's center to Earth's surface, but the E-M barycenter plane is 5.14° from the ecliptic and the Earth rotates around it in 27.32 sidereal days. Again, the E-M barycenter is not able to provide Earth's orbital velocity, so the next freely falling frame would be the solar nonrotating barycenter (SNB) to be a sufficiently inertial frame. The velocity, V , of the E-M barycenter in the SNB frame can be found at any time and the orientation of the local vertical can be calculated after many transformations and rotations. The sine projection of the orbital velocity to the local vertical will be the perpendicular component needed to

predict the directional offset that a beam of photons will miss the nadir. One example is given above resulting in a floor displacement of 25 cm if the perpendicular velocity is 15000 m/s to the local vertical. If no reflected beam is displaced over a day, then the photons must be getting an additional velocity from the mirror added to its photon emission velocity so that the beam has the directional change and magnitude to keep touching the nadir.

A more rigorous demonstration reveals that photons precisely obey vector velocity addition. The Laser Interferometer Gravitational-Wave Observatory (LIGO) is a large-scale physics experiment with three observatories to detect cosmic gravitational waves [19] Twin observatories are near Hanford, Washington, and Livingston, Louisiana, with 4 km long arms within ultrahigh vacuum chambers allowing laser beams to detect gravity waves. VIRGO, located in Italy, joined LIGO in May 2007. **Table 1** lists the locations and azimuth in degrees counter-clockwise of each arm from local east at each facility.

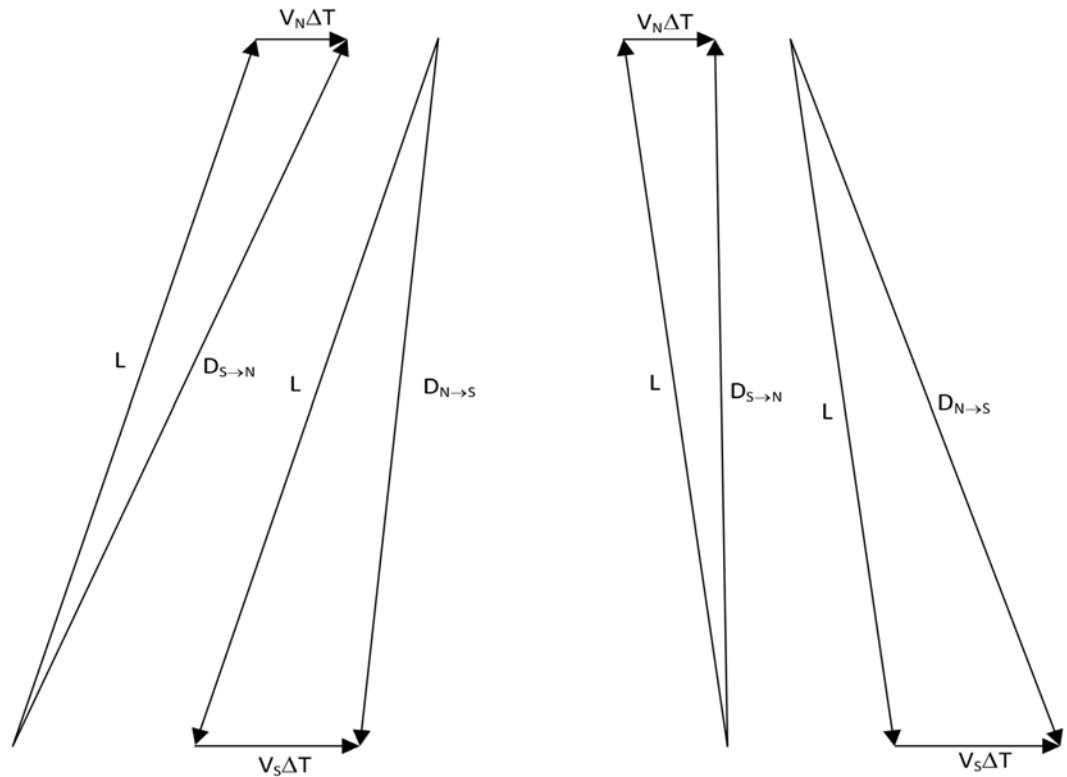
Each LIGO facility uses a continuous laser beam that is amplified from 40 watts to 750 watts with power reflecting mirrors. The signals are also magnified with signal recycling mirrors. LIGO has enhanced vibration absorption mechanisms to remove ground vibrations, tremors, solid Earth tides, etc., to isolate the signals. To increase the arm lengths from 4 km, Fabry-Perot cavities are installed at the beam splitting mirror and at the hanging reflection mirror at the end of each arm so that 280 reflections inside the cavities increase the effective arm length to almost 1120 km. An ultrahigh vacuum is maintained so that any gaseous molecule is removed promptly to avoid interference with the photon beams. Also, one of the split signals is inverted so destructive interference is created when recombining the two beams. The original Michelson-Morley interferometer gives constructive interference, but this enhancement allows far easier detection of variations after merging signals. “At its most sensitive state, LIGO will be able to detect a change in distance between its mirrors 1/10,000th the width of a proton.” [19]. (*i.e.*, resolution of $1.7\text{E}-19$ m.)

There were several runs, which LIGO operated continuously: 1) 12 September 2015 to 19 January 2016, 2) 30 November 2016 to 25 August 2017, 3) 1 April 2019 to 30 September 2019, and 4) 1 November 2019 until 27 March 2020, covering over 22 months. In a freely falling, nonrotating frame, which Einstein stated is sufficiently inertial, choose Earth’s center of mass as the origin and pick a time during LIGO’s operation. Designate this inertial Earth center-of-mass, nonrotating frame as ECN.

As all facilities are in the northern hemisphere, the Earth’s rotation will displace the southern end of all arms further eastward than the northern end of the arm. **Figure 4** shows this effect where the left side of every triangle is the arbitrary azimuth of the arm, but Earth’s rotation displaces the opposite end of the arm eastward, or to the right in each figure. The 4 km arm is L with the azimuth at emission (going northward) or at reflection (going southward), but the photon’s actual displacement due to Earth’s rotational effect is D . Note that V_s is greater than V_N depending on the geodetic latitude of each arm’s end.

Table 1. LIGO observatories with arm orientations.

LIGO	Location		Local Azimuth	
Facility	Latitude	Longitude	X Arm	Y Arm
Hanford	46°27'19"N	119°24'28"W	126°	216°
Livingston	30°33'46"N	90°46'27"W	198°	288°
Virgo	43°37'53"N	10°30'16"E	71°	161°

**Figure 4.** One-way photon displacement from rotation.

The roundtrip distance that photons travel along the x-axis is not equal to the roundtrip distance of the y-axis photons. Both x-axis and y-axis arms are joined at the same point where the beam is split and where the two split beams are recombined, but the opposite ends of both arms are at different latitudes where Earth's rotation from west to east will give different tangential velocities of these opposite ends in the inertial ECN frame. Not only is the outgoing distance unequal to the incoming distance for each arm in the ECN frame, but the difference in the tangential velocities at the opposite ends in the ECN frame prevents the roundtrip distances of the photons in the x-axis arm to be the same as the y-axis arm. All LIGO output over 22 months of continuous operation have resulted in the split beams arriving at the same simultaneous time upon interception (except for the many spontaneous data glitches and the rare detections of gravity waves). The only way for split beams to travel over different distances in the ECN frame and arrive at identical time spans is for photons to have different

one-way velocities upon emission.

LIGO observatories are enormous versions of the Michelson-Morley interferometer, which produced a null result when searching for an ether to explain light propagation [20]. One standard answer is length contraction along the axis parallel to the constant velocity will shorten the arm so the roundtrip traverse is identical to the distance traversed in the perpendicular arm. When the split beams are recombined, complete constructive interference occurs. Only length contraction from special relativity is applied to the arm parallel to the velocity of the apparatus. Time dilation should also be applied to the photons traversing the parallel arm, which would make that photon speed different than c . Equation (6) predicts from special relativity that photon speed is a constant of $\gamma^2 c$ in the parallel arm and different than the standard c in the perpendicular arm. If one applied special relativity only to the arm parallel to the Earth's orbital velocity, then destructive interference will be the output. The transverse beam will have a different arrival time at the recombination point due to a standard c speed than the parallel beam with $\gamma^2 c$ speed, even if the roundtrip traverses are identical after length contraction is applied to the parallel arm. The 22 months of continuous output of simultaneous arrival at the recombination point conflicts with this scenario.

The LIGO observatories are noninertial facilities and fixed on Earth's rotating surface. The exact transformation between ECN and the rotating Earth Center, Earth Fixed frame (ECEF) in geographical coordinates is the rate of change of a vector as shown below as an operator equation acting on a vector [21]:

$$\left(\frac{d}{dt} \right)_{ECN} = \left(\frac{d}{dt} \right)_{ECEF} + \boldsymbol{\omega} \times \quad (7)$$

where $\boldsymbol{\omega}$ is the angular velocity of Earth's rotation in the cross product. UT1 will be the worldwide time standard for both ECN and ECEF as maintained by BIPM by synchronizing all time standards in all timing institutes. Photons are emitted at the splitting location to traverse either in the x-axis arm or y-axis arm. At the opposite ends of the axes, the hanging mirrors absorb the photons and emit new photons from the excited atoms to return to the recombination point. Let $R_{x,o}$, $R_{y,o}$, $R_{x,r}$ and $R_{y,r}$ denote the displacement vector the photons traverse with footnote x denoting the x-axis, y the y-axis, o for outward direction from the splitting point, and r for the return to the recombination point. The second postulate states "Any ray of light moves in the 'resting' coordinate system with the definite velocity c , which is independent of whether the ray was emitted by a resting or by a moving body." ([10], §2, lines 9-11). When applying the operator Equation (7) to any of the four displacements, the ECN derivative should be the standard c velocity within an inertial frame, but the cross product term differs with direction and with geodetic latitude of the emission point that is the base of each displacement vector. Note that $\boldsymbol{\omega} \times \mathbf{R}_{x,o} \neq \boldsymbol{\omega} \times \mathbf{R}_{y,o}$ due to differences in the azimuths of the outgoing photons in the arms, and $\boldsymbol{\omega} \times \mathbf{R}_{x,r} \neq \boldsymbol{\omega} \times \mathbf{R}_{y,r}$ due to different latitudes where the photons were emitted as located at the base of the

vectors. With the reputed sensitivity of $\Delta L = 1.7\text{E}-19$ m, LIGO is affected by the Coriolis effect, but nothing on the website or in the literature gives any indication that LIGO compensates for it. The Coriolis effect is not ignorable with LIGO's precision that exceeds 13 significant figures. The 22 months of continuous observations of simultaneous arrival of photons through each arm in the ECEF frame require that the one-way photon speeds within the ECEF frame, dR/dt , vary within each arm and direction so that the same time interval within ECEF remains the same in LIGO. That also occurs in the ECN frame as LIGO's photon emitters (*e.g.*, splitters and mirrors) move within ECN. LIGO shows that the theory must be revised. Also, the perpendicular component of Earth's orbital velocity relative to each axis is different and constantly changing, causing each beam to miss the arm's opposite end up to a maximum of 40 cm. LIGO makes no compensation for Earth's orbital velocity affecting each arm's beam. One explanation for equal time transmissions to occur with unequal displacements in inertial frames is photon velocities obey vector velocity addition with moving photon sources. A null result in the laser test using a hemisphere mirror will also reveal that photon velocity is the sum of the standard velocity and the velocity of the emitting source.

Vector addition in three dimensions of two displacements is $\mathbf{Z} = \mathbf{X} + \mathbf{Y}$. As LIGO demonstrated equal time spans for photons to traverse through either arm as $\Delta t_x = \Delta t_y$ (or $dt_x = dt_y = dt$), then the derivative of the addition law of displacements is:

$$\frac{d\mathbf{Z}}{dt} = \frac{d\mathbf{X}}{dt} + \frac{d\mathbf{Y}}{dt} \text{ or } \mathbf{c}' = \mathbf{c} + \mathbf{v} \quad (8)$$

So, photons are emitted with a combined one-way speed of the standard light speed plus the velocity of the emitter relative to the resting frame. This verifies by (8) how photons obey vector velocity addition according to (4) and (5) from the ends of the uniformly moving rod.

For photons traveling by a roundtrip over a finite number of traverses, the average speed of the photons is the standard speed of light, c . This is already shown for $n = 2$ traverses as $(c_{AB} + c_{BA})/2 = 2c/2 = c$ in Equations (4) and (5). No $n = 1$ case exists for a roundtrip, because a reflection requires a minimum of two traverses. Circular tracks are made up of contiguous atoms as a photon is emitted by an atom and absorbed by the nearby atom so that the segments on the order of Avogadro's number approach a circular path. The proof of an average speed of photons in a roundtrip can be done by induction. The initial case for $n = 2$ is true by (4) and (5). Assume the average speed over a roundtrip excursion for photons is c with n traverses. Let $V_n = v_1 + v_2 + \dots + v_n$, which is the vector sum of all the velocities from emitters 1 through n that starts from the origin and ends at the end of leg n . The average speed over $n + 1$ traverses is:

$$\frac{\sum_{i=1}^{n+1} c_i + \sum_{i=1}^{n+1} v_i}{n+1} = \frac{(n)c}{n+1} + \frac{V_n}{n+1} + \frac{c}{n+1} + \frac{v_{n+1}}{n+1} = c \quad (9)$$

The vector sum of V_n and v_{n+1} is the velocity from the origin to the end of leg

n and through the start of leg $n + 1$ back to the origin. The sum of V_n and v_{n+1} equals zero with no net displacement by $n = 2$ case. Reducing the roundtrip from $n + 1$ legs into 2 legs proves the $n + 1$ case by (9). This proof illustrates that the second postulate of relativity is too restrictive and only applies to the case when the photons complete a roundtrip circuit, which the average speed is the standard c .

6. Issues of Varying Photon Speeds

If photon speeds obey vector velocity addition, then a faster speed would elongate the wavelength without changing the frequency so that $c' = f\lambda'$, where c' is any photon speed, f is the associated frequency and λ' is the wavelength. This would mean that photons behave now the same in a vacuum as in a transparent medium. The photon speed in a medium is $c/n = c'$ from experiments (n is the index of refraction and $n \geq 1$ with 1 for a vacuum), and only the speed and wavelength changed in the medium. When photons penetrate different layers of transparent media, the speed and wavelength change with each material, but no waves stack up at the interfaces between materials. This also applies to photons exiting a material into a vacuum or vice versa.

If photon speeds in a vacuum conform to $c' = f\lambda'$, then photons gain speed the farther they move from a gravitated body. In the Pound-Rebka test, photons that traveled up 22.5 m to the absorber had a longer wavelength and a shorter wavelength when traveling down. As gravity weakens with a photon moving away from the mass, there is less deceleration on the photon, which can then move faster. Beams bent by the Sun demonstrate this. Despite the tiny size of a photon, the side of the photon opposite to the Sun has a slightly greater velocity than the side facing the Sun, which makes the beam bend around the Sun. This is much like a tank with its left tread traveling slightly faster than the right tread, so the tank will turn right from a straight heading. This effect impacts all cosmological models and observations.

For example, a redshift in stellar transmissions is an increase in the measured wavelength, which is inferred that the frequency decreased when assuming a universal constant light speed. If photons gain speed the farther away they go from a star, a longer wavelength is created in all stellar emissions. If the net velocity of the star relative to Earth is great enough in the opposite direction, the effect would overcome the lengthening of the wavelength by gravity to produce a shorter wavelength seen on Earth. Under this condition, it would produce the inferred “blueshifts” that are infrequently recorded. If the stars and galaxies have a symmetric normal distribution of velocities toward or away from Earth, then most stellar photons will have lengthened wavelengths (*i.e.*, “redshift”) and fewer photons will have shorter wavelengths (*i.e.*, “blueshift”) as the needed stellar velocities to overcome a gravitational lengthening would be in one tail of the distribution. This may rectify a controversy over the Big Bang model where some observed galaxies have “redshifts” that place them outside the sphere of observa-

bility based on the probable age of the universe according to Hubble's law. In any case, this will require careful analysis by cosmologists as this research indicates "blueshift" observations are due to stellar objects moving from Earth—not towards Earth.

Also, the photon deflection of 1.75" when grazing the Sun can be obtained from Newton's solution of the two-body problem by allowing the Sun and photon to move in hyperbolic orbits against each other. When starting with a photon infinitely far from the Sun, their barycenter is also infinitely far between them. The same gravity the Sun pulls on the photon is also the same pull of the photon exerted on the Sun. This is a two-body problem, which the two bodies execute hyperbolic orbits about the barycenter. As the photon approaches the solar system, the barycenter is moving inside the Sun. The details are obtained with the photon and solar deflection measured against the orbital asymptotes to produce a total deflection of 1.75". [22] Previous classical derivations assumed the Sun was stationary and only the photon is in orbit, which produces half of the observed deflection.

If particles obey vector velocity addition, then a high particle speed could cause misinterpretation of a particle as being different. This may affect at least two particles in the standard model of particle physics. The muon and tauon are identical in all characteristics to the electron, except their masses are different according to special relativity. If a high-energy particle after a collision can exceed the standard light speed, then $E = mc^2$ may misinterpret high speed electrons as muons or tauons for the same electron having a rest mass. A faster photon speed than the standard light speed c would explain all the entanglement experiments where two or more objects at significantly long distances, d , interact with each other within a measured short time interval, Δt (i.e., $d > c\Delta t$).

If high speed or high energy particles can exceed the standard light speed, then a different explanation of high energy γ rays impacting the atmosphere should be considered. High energy cosmic radiation impacts atmospheric molecules creating pions that decay into muons within 26 ns. Electromagnetically, these muons interact with other atmospheric particles before disappearing with a typical lifetime of about 2.194 μ s. With a velocity just below c , it would take about 50 μ s to reach sea level. The current explanation is that time dilation allows the muons to reach sea level while classical mechanics would predict hardly any would be detected [23]. If electron velocity can exceed c , the cosmic rays may collide with the atmosphere to eject electrons with higher velocities up to say $25c$, which would allow those electrons on average to reach sea level without time dilation.

If photon speeds obey vector velocity addition, then some constants in physics are no longer constants. If photon speeds change c in a vacuum, then permittivity of free space, ϵ_0 , must be a variable as the product $4\pi\epsilon_0c^2$ is defined to be exactly $1E7$ [24]. The definition of the fine structure constant is $\alpha = e^2/(2\epsilon_0hc)$ where e is the elementary charge and h is Planck's constant with ϵ_0 and c defined. Electric charge e is $1.602176634 \times 10^{-19}$ C, which is a static value for electrons

and protons. Then, α equals $2\pi c^2/(h \text{ 1E7})$, which implies that either the fine structure constant could vary or Plank's constant h could vary inversely as c varies. Electromagnetic concepts should be revised since a moving photon emitter could change the speed of photons relative to the reference frame and change ϵ_0 . Current interpretation is that any moving photon source has no effect to change the speed of light, which is usually adequate for results of 5 or 6 significant figures and the emitters move with a velocity < 1000 m/s. Maxwell's electromagnetic equations assume c and ϵ_0 are constants. Curls and divergences are special derivatives that would produce additional terms from the products containing c or ϵ_0 . It is beyond the scope of this paper to delve into the general derivation by expanding Maxwell's equations when c and ϵ_0 are variables.

If photon speeds obey vector velocity addition, then the meter can not be defined by a one-way photon transmission, as the velocity of the emitter relative to the observer must be compensated in the reference frame. One may have to reestablish the physical meter as the international standard for length.

If photon speeds can exceed c by the emission from a moving satellite, then an alternative explanation of the flyby anomaly is possible. Several deep space satellites have been observed to have a small increase in velocity when nearing the perigee of a flyby of Earth, which often is in the range of 3 to 11 mm/s or about $2\text{E}-6 = dV/V$. An empirical equation for the anomalous flyby velocity change has been proposed [25], which is

$$\frac{dV}{V} = \frac{2\omega_E R_E (\cos \theta_i - \cos \theta_0)}{c} \quad (10)$$

where ω_E is the Earth's angular velocity, R_E is the Earth's equatorial radius, and θ_i and θ_0 are the inbound and outbound equatorial angles of the spacecraft. The tangential velocity at the equator due to Earth's rotation, $\omega_E R_E$ is 465 m/s. Depending on the cosine projection, Equation (10) is $3.1\text{E}-6$ or slightly less. Equation (10) predicts that the transmitted photon arrived at the ground station slightly sooner than anticipated, which would indicate the photons had a slightly faster velocity ($\sim(1 + 3\text{E}-6)c$). The effect is tiny, but present, so long as other effects such as atmospheric drag do not cover the effect.

If satellite transmissions obey vector velocity addition, then prograde satellites from the west of a ground station will have a slightly faster transmission velocity and satellites east of the ground station will have a slightly slower transmission velocity. The Global Positioning System (GPS) could improve its GPS satellite orbits by making this compensation in its pseudoranges, which are equivalently the time difference between transmission and reception of the GPS signals. This would also affect ultraprecise timing institutions, which use a GPS satellite in common view to compare between the institutions. Typically, Lab A differences its time with the GPS time from a common view satellite over several minutes for smoothing, and Lab B does the same with the same GPS satellite. Then, each lab exchanges the data. $\text{Lab}_A - \text{GPS}$ and $\text{Lab}_B - \text{GPS}$ data are calculated, and these sets of data are differenced to get $\text{Lab}_A - \text{Lab}_B$. As shown in the previous

paragraph, the effect can be as large as $3\text{E-}6$. This would affect ultraprecise timing institutions that maintain picosecond precision with their atomic clocks.

Some may argue that prior tests measured the same standard speed of light emitted by a moving particle. For example, γ rays from the decay of π^0 mesons with more than 6 GeV were measured absolutely by timing over a known distance [26]. The test was intended to measure $c + kv$, and the result was $k = (-3 \pm 13) \times 10^{-5}$ for mesons moving near light speed ($\gamma > 45$). The team used two detectors spaced 31.450 m apart to measure the time interval the γ rays traveled. The first detector absorbed the photons from high-speed γ rays and emitted new photons at the usual speed of light. The measured speed recorded by the second detector after light passed through the first detector (*i.e.*, absorbed and reemitted at c) was the standard light speed. This and other tests must be reexamined carefully to ensure that the photon speeds were directly measured without interception to eliminate any misinterpretations of the results.

If photons obey vector velocity addition, then an additional interpretation of the Sagnac effect is needed for some applications. Light is inserted into a ring interferometer and splits in opposite directions. The beams exit the ring at the start/end point and undergo interference. The destructive interference determines how much the ring interferometer rotated after the beams were split. The diagram on the left of Figure 5 [27] shows a nonrotating ring of radius R would output constructive interference as each beam would travel the same length of $2\pi R$ and exit at the same time. If the ring interferometer rotated as shown on the right side of Figure 5, one beam would travel further than the other, so both beams would exit with destructive interference.

It is easy to see as an external observer that one beam takes a longer trip than the oppositely traveling beam when the interferometer rotates, being the standard answer of the Sagnac effect. The rotating interferometer is a frame that is sufficiently inertial when comparing the constant angular velocity ω with the speed of light. The ring is a conduit that bends both beams to traverse in a

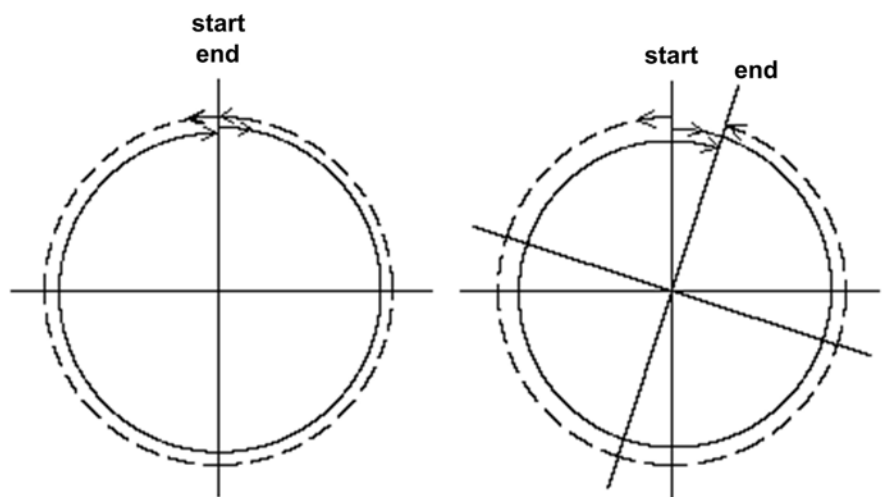


Figure 5. Sagnac effect in circular loops.

circular path. A perpendicular acceleration is required to force the light beams to traverse a circle in the rotating interferometer. Such an acceleration is no different than the perpendicular gravity that exists for the laboratory's light speed test, as any acceleration to the linear paths will be equal to both beams, so the acceleration effects cancel out when differencing the two beams.

Internally, Equations (1) and (2) give the lengths that the beams traverse in the circular interferometer with counterclockwise length for L_{AB} and clockwise length L_{BA} . For a circular traverse, set $L = 2\pi R$ and $v = \omega R$. In internal navigation applications, $c_{AB} \approx c_{BA} \approx c$. The time difference between the combined beams is:

$$\Delta t \approx \frac{2\pi R}{c} \left[\frac{c}{c - \omega R} - \frac{c}{c + \omega R} \right] = \frac{4\pi\omega R^2}{c^2 - \omega^2 R^2} \quad (11)$$

If an observer is fixed with the rotating ring, special relativity requires that each beam originates at the same constant light speed c , and each beam travels the circumference of $2\pi R$ inside the inertial rotating frame. This set of assumptions of special relativity for the rotation of the ring undergoing constant angular velocity, ω , would predict both beams exit simultaneously (*i.e.*, constructive interference in the output), but reality contradicts that. Equations (4) and (5) show that the inertial ring interferometer undergoing constant angular velocity ω will impart different one-way velocities to the light beams at the splitter for net velocities of $c + \omega R$ and $c - \omega R$. This explanation will obtain Equation (11) in the rotating ring frame as output for the internal observer. The identical observed interference of output is witnessed for both observers—the external observer fixed to the frame and the internal observer fixed to the circular interferometer. Thus, light obeys vector velocity addition in both linear and rotating reference frames.

As stated before, the second postulate is an excellent approximation, but not exact. The prediction of general relativity for the Mercury perihelion shift is 42.98"/cy, but Meisner, Thorne and Wheeler ([28], p. 1112) determined the residual relativistic perihelion shift was 41.4 ± 0.9 "/cy that does not encompass general relativity, which they gave no explanation. A post-Newtonian approximation with coordinate time predicted 40.73"/cy, which is within the error boundary [29]. Shapiro *et al.* [30] chose Icarus for its large orbital eccentricity and periodic close approaches to Earth for observations to compare against general relativity's orbital perihelion shift for a highly eccentric orbit. A dimensionless parameter λ determined how well perihelion predictions compared with observations (unity for general relativity and zero for Newtonian theory). Their first solution for estimating the orbital elements of Icarus and the λ factor was $\lambda = 0.75 \pm 0.08$, which excluded general relativity, but the post-Newtonian approximation of 0.83 fits [29] [30]. Shapiro *et al.* [30] wrote, "we must conclude that either general relativity is incorrect or some other aspect of either our theoretical model or the observations differed from our presumptions." After many more solutions using different models and even varying the angular spread between background stars, that team concluded that the FK4 star catalog had a

distortion that caused the anomaly, but subsequent star catalogs such as FK5 show no such regional distortion in comparison to the FK4, which coordinate transformations from FK4 axes are rotations to later star catalogs' axes.

If the one-way speed of photons obeys vector velocity addition due to the velocity of the photon emitter, then the quantity $F = c^2t^2 - x^2 - y^2 - z^2$ is not invariant. The values $F = \pm 1$ define two-sheeted hyperboloids and $F = 0$ yields their asymptotes, which had defined the trajectories of a light ray ([10], p. 268). Minkowski assumed these are invariances to define and describe his four-dimensional space-time model (later used in general relativity), which are based on Poincaré's group properties of the Lorentz transformation, and he used Lorentz's F to seek other invariant quantities under the Lorentz group, as (x, y, z, ict) ([10], p. 238-243). Any derivation based on these quantities is only approximate and limited to only the case when photons complete a roundtrip traverse. So, many derivations using these terms as invariances in quantum mechanics must be scrutinized for valid predictions of particles with one-way velocities.

This is an incomplete list of possible ramifications if the one-way photon speed is not a universal constant in a vacuum. In any case, it will take time to consider these results thoroughly. It would not be surprising if other accepted concepts of physics may need revisions.

7. Conclusions

The word "exact" should be deleted in any updated definition of the meter. The original definition of the meter was $1\text{E}-7$ of the meridian length from the equator to the north pole at sea level. Due to the irregular shape of the Earth, such a meter would differ depending on the choice of the meridian, even if one did get the quarter meridian length divided exactly into $1\text{E}7$ units. The International Prototype Meter was the replacement in 1889, which all other national meter prototypes were exactly compared. However, technology improved precision that, by 1927, the definition of the meter was updated to support the prototype with two cylinders to circumvent sagging during measurement and stipulated the temperature to prevent expansion of the prototype. As the BIPM and CGPM acknowledge that gravity can shorten the meter by $1.1\text{E}-16$ m for every meter the test is elevated in altitude, that alone admits the current meter's definition is never exact. Even an exact definition of the nautical mile of 1852 m is incorrect as it is 1855.324 m due to improvements in geodesic measurements of Earth's equatorial radius. That proves "exact" values must be scrutinized regularly for accuracy and improved precision to redefine any wrong values.

As discovered in the analysis, the one-way speed of photons depends on the velocity of the emitter in the reference frame. This was revealed in both the theoretical analysis of photons traversing the length of Einstein's uniformly moving rod and in LIGO's simultaneous interference on a rotating Earth. Assuming a universal constant speed c , the total distance in the stationary frame that photons move in both directions to traverse a uniformly moving rod ex-

ceeds the total distance when the rod is stationary. Length contraction is not enough to rectify the difference in distances, yet the roundtrip time is the same for either the stationary rod or the uniformly moving rod as required by the first postulate of relativity. LIGO shows equal roundtrip times through the x-axis and y-axis arms of participating observatories that are oriented with various azimuths on a rotating Earth. The rotation causes the north ends of each arm to move with slower eastward speeds and less eastward displacements than the south ends, so that the roundtrip distances traversed in an inertial reference frame are different for each arm. Furthermore, the rate of change of a vector (Equation (7)) reveals that the cross product term applied to photon displacements will yield different velocity contributions between the x-axis and y-axis arms due to different orientations. Those contributions are added to the photon velocity in LIGO's frame to produce the photon velocity in the inertial frame. The outgoing velocities will differ with the returning velocities through each arm due to the cross product terms, which makes the photon speed in the inertial frame different than c . Excluding the data glitches and rare gravitational wave detections, LIGO demonstrated the beams arrived simultaneously at the recombination point for 22 months of continuous operation. However, the average speed of roundtrip traverses is the standard photon speed, c , that would allow simultaneous arrival of the split beams at the recombination point, provided the one-way speed of the photons obeys vector velocity addition. For circular interferometers undergoing a constant angular velocity, the external observer easily sees that the output of destructive interference occurs when two beams moving oppositely inside the interferometer cover different distances. For the internal observer fixed with the uniformly rotating interferometer, the output is the same if the interferometer's splitter imparts an additional velocity to the photons moving oppositely to create the same destructive interference recorded by the external observer.

The proposed test using a laser beam, half-silvered flat mirror, and hemisphere mirror will prove if photons do exhibit vector velocity addition. Due to Earth's orbital velocity and Earth rotation, the velocity of the test apparatus will undergo different daily velocities and directions by the changing perpendicular velocity component from Earth's orbital velocity. Photons emitted from the half-silvered mirror will only impact the nadir if their additional velocities contributed by the emitting atoms are the same as the hemisphere mirror is subjected throughout the day. Photons emitted from the hemisphere mirror will hit the same point on the floor if they have the additional velocity of the hemisphere mirror as the floor.

In-depth analysis indicates photon speeds obey vector velocity addition in one-way traverses. In roundtrip tests, the total roundtrip times are always the same. It is proven by induction that the average speed is exactly c in roundtrip tests in either moving or stationary frames. Every precise test measuring photon speeds has been with roundtrip traverses using a reflective surface, which always

produced the apparent constant for photon speeds. This was found in both Einstein's thought experiment with the rod and in LIGO's continuously constant reception times. An updated definition of the meter must require a roundtrip test of photons and a fixed distance between the emitter and the reflector during the test to ensure the elapsed transmission time is always the same to construct a meter.

Also, gravity affects the speed of the photons in the vertical direction to be faster the farther away a photon is from the body of mass. For example, the solar deflection of beams near the Sun's surface is bent around the Sun. The side of the photon nearest to the Sun has a slower velocity than the side away from the Sun, which causes the photon trajectory to bend slightly towards the Sun. This is similar to a tank with a left tread moving faster than the right tread so that it turns to the right compared to a straight path. An updated definition of the meter should require the test be conducted in the horizontal plane. If the test is conducted on Earth's surface, the differences due to altitude changes would be in multiples of $1\text{E}-7$ nm, which is well below the current practical limits of construction. If the meter is attempted in highly different gravitation than Earth's local gravity (*e.g.* another planet or in outer space), then the gravitational effect to construct a meter must be included in the definition.

It is understood that the SI second is implied, but not stated in the current definition of the meter. Other seconds as time units are defined for specific uses. The laboratory time should be coordinate time synchronized to the master time standard. According to special relativity, proper time recorded by moving clocks will produce longer second units than coordinate seconds.

The following suggestion is offered as a possible revision to the current definition of the meter to eliminate the inadequacies discussed in this paper.

The metre is the length between a light source and reflector that is accomplished near Earth's surface in a roundtrip traverse by light in a vacuum during a time interval of $2/299,792,458$ of an SI second in coordinate time where the distance between the source and reflector is fixed relative to the Earth during the test and is oriented perpendicular to the local gravity.

If more precise testing of the photon's speed in a vacuum obtains more significant figures, the number in the denominator of the definition's time interval should be updated.

Conflicts of Interest

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

References

- [1] (1989) Oxford English Dictionary. 2nd Edition, Clarendon Press, Oxford, Vol. 9, 697.
- [2] Rapport sur le choix d'une unité de mesure, lu à l'Académie des sciences, le 19 mars 1791. (in French)

- [3] Maxwell, J.C. (1873) *A Treatise on Electricity and Magnetism*. MacMillan and Co., London, 3.
- [4] Michelson, A.A. (1893) *Astronomy & Astrophysics*, **12**, 556-560.
- [5] Marion, J.B. (1982) *Physics for Science and Engineering*. CBS College Publishing, New York, 3.
- [6] (1983) 17th General Conference of Weights and Measures. Resolution 1.
- [7] BIPM (2003) *Mise en Pratique*.
- [8] Taylor, B.N. and Thompson, A. (2008) *The International System of Units (SI) United States Version of the English Text of the Eighth Edition (2006) of the BIPM Publication Le Système International d'Unités (SI) (Special Publication 330)*. National Institute of Standards and Technology (NIST), Gaithersburg.
- [9] Einstein, A. (1905) *Annals of Physics*, **17**, 891-921.
<https://doi.org/10.1002/andp.19053221004>
- [10] Miller, A.I. (1981) *Albert Einstein's Special Theory of Relativity: Emergence (1905) and Early Interpretation (1905-1911)*. Addison-Wesley Pub. Co. Inc., Reading.
<https://doi.org/10.1063/1.2914975>
- [11] Markowitz, W., Hall, R.G., Essen, L. and Parry, J.V.L. (1958) *Physical Review Letters*, **1**, 105-107. <https://doi.org/10.1103/PhysRevLett.1.105>
- [12] Aria, F. and Lewandowski, W. (2011) *Metrologia*, **48**, E01.
- [13] Riehle, F. (2015) *Comptes Rendus Physique*, **16**, 506-515.
<https://doi.org/10.1016/j.crhy.2015.03.012>
- [14] Pound, R.V. and Rebka Jr., G.A. (1959) *Physical Review Letters*, **3**, 439-441.
<https://doi.org/10.1103/PhysRevLett.3.439>
- [15] Department of Defense (1991) *World Geodetic System 1984, DMA Technical Report 8359.2. 2nd Edition*, U.S. Government Printing Office, Washington DC.
- [16] Moritz, H. (1980) *Bulletin Géodésique*, **54**, 395-405.
<https://doi.org/10.1007/BF02521480>
- [17] Dahl, P.F. (1997) *Flash of the Cathode Rays: A History of J J Thomson's Electron*. CRC Press, Boca Raton, 10. <https://doi.org/10.1201/9781482268461>
- [18] Nahin, P.J. (2002) *Oliver Heaviside*. JHU Press, Baltimore.
<https://doi.org/10.56021/9780801869099>
- [19] LIGO caltech.edu Website (Operated by Caltech and MIT), *Facts Subpage Lists the Extreme Engineering to Construct LIGO and Its Operation*.
- [20] Michelson, A.A. and Morley, E.W. (1887) *American Journal of Science*, **34**, 333-345.
<https://doi.org/10.2475/ajs.s3-34.203.333>
- [21] Goldstein, H. (1980) *Classical Mechanics*. 2nd Edition, Addison-Wesley Pub. Co., Boston, 174-180.
- [22] Deines, S.D. (2016) *International Journal of Applied Mathematics and Theoretical Physics*, **2**, 52-56.
- [23] Covarrubias, E. (2018) *Relativistic Effect of Cosmogenic Muons*. California State University, Stanislaus.
- [24] Lorrain, P. and Corson, D.R. (1970) *Electromagnetic Fields and Waves*. 2nd Edition, W. H. Freeman and Co., San Francisco, 250.
- [25] Anderson, J.D., *et al.* (2008) *Physical Review Letters*, **100**, Article ID: 091102.
<https://doi.org/10.1103/PhysRevLett.100.091102>
- [26] Alväger, T.A., Farley, F.J.M., Kjellman, J. and Waller, I. (1964) *Physics Letters*, **12**,

- 260-262. [https://doi.org/10.1016/0031-9163\(64\)91095-9](https://doi.org/10.1016/0031-9163(64)91095-9)
- [27] The Sagnac Effect. <https://www.mathpages.com/rr/s2-07/2-07.htm>
- [28] Misner, C.W., Thorne, K.S. and Wheeler, J.A. (1973) Gravitation. W. H. Freeman and Co., San Francisco, 1112-1113.
- [29] Deines, S.D. (2017) *International Journal of Applied Mathematics and Theoretical Physics*, **3**, 61-73. <https://doi.org/10.11648/j.ijamtp.20170303.14>
- [30] Shapiro, I.I., Smith, W.B., Ash, M.E. and Herrick, S. (1971) *The Astronomical Journal*, **76**, 588-606. <https://doi.org/10.1086/111168>

Mathematical Aspects of $SU(2)$ and $SO(3, \mathbb{R})$ Derived from Two-Mode Realization in Coordinate-Invariant Form

Alfred Wünsche

Institut für Physik, Humboldt-Universität, Newtonstr. 15, Berlin, Germany

Email: alfred.wuensche@physik.hu-berlin.de

How to cite this paper: Wünsche, A. (2023) Mathematical Aspects of $SU(2)$ and $SO(3, \mathbb{R})$ Derived from Two-Mode Realization in Coordinate-Invariant Form. *Journal of Modern Physics*, 14, 361-413. <https://doi.org/10.4236/jmp.2023.143022>

Received: January 16, 2023

Accepted: February 25, 2023

Published: February 28, 2023

Copyright © 2023 by author(s) and Scientific Research Publishing Inc. This work is licensed under the Creative Commons Attribution International License (CC BY 4.0). <http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

Some mathematical aspects of the Lie groups $SU(2)$ and $SO(3, \mathbb{R})$ in realization by two pairs of boson annihilation and creation operators and in the parametrization by the vector parameter φ instead of the Euler angles (α, β, γ) and the vector parameter c of Fyodorov are developed. The one-dimensional root scheme of $SU(2)$ is embedded in two-dimensional root schemes of some higher Lie groups, in particular, in inhomogeneous Lie groups and is represented in text and figures. The two-dimensional fundamental representation $D^{\left(\frac{1}{2}\right)}$ of $SU(2)$ is calculated and from it the composition law for the product of two transformations and the most important decompositions of general transformations in special ones are derived. Then the transition from representation $D^{\left(\frac{1}{2}\right)}$ of $SU(2)$ to $D^{(1)}$ of $SO(3, \mathbb{R})$ is made where in addition to the parametrization by vector φ the convenient parametrization by vector c is considered and the connections are established. The measures for invariant integration are derived for $SO(3, \mathbb{R})$ and for $SU(2)$. The relations between 3D-rotations of a unit sphere to fractional linear transformations of a plane by stereographic projection are discussed. All derivations and representations are tried to make in coordinate-invariant way.

Keywords

Boson Operators, Lie Algebra, Root Diagram, Invariant Integration, Hamilton-Cayley Identity, Cayley-Gibbs-Fyodorov Parametrization, Composition Law, Quaternion, Stereographic Projection, Fractional Linear Transformation

Remarks about Notations

Occasionally, we use for explanations of coordinate-invariant expression with three-dimensional tensors an index form including the Levi-Civita or fully anti-symmetric symbols ε_{ijk} defined by

$$\varepsilon_{ijk} = \varepsilon_{jki} = \varepsilon_{kij} = -\varepsilon_{ikj} = -\varepsilon_{kji} = -\varepsilon_{jik} \equiv \varepsilon^{ijk}, \quad \varepsilon_{123} = \varepsilon^{123} = 1. \quad (1)$$

Our most important notations deviating from an essential part of English literature are the following very rational ones for scalar product, vector product and dyadic product of two vectors $\mathbf{a}$ and $\mathbf{b}$

$$\tilde{\mathbf{a}}\mathbf{b}, [\mathbf{a}, \mathbf{b}], \mathbf{b} \cdot \mathbf{a}, \text{ or } \tilde{\mathbf{a}}\mathbf{b} \equiv \tilde{a}_i b^i, [\mathbf{a}, \mathbf{b}]_i \equiv \varepsilon_{ijk} a^j b^k, (\mathbf{b} \cdot \mathbf{a})^i \equiv b^i \tilde{a}_j, \quad (2)$$

and for the volume product of three vectors $\mathbf{a}, \mathbf{b}$ and $\mathbf{c}$

$$[\mathbf{a}, \mathbf{b}, \mathbf{c}] = [\mathbf{a}, \mathbf{b}]\mathbf{c} = \mathbf{a}[\mathbf{b}, \mathbf{c}] = \mathbf{a}[\mathbf{b}]\mathbf{c} \equiv \varepsilon_{ijk} a^i b^j c^k. \quad (3)$$

They agree fully or partially with many weighty sources, in particular, most completely with Rosenfeld [1] but also widely with, e.g., [2] [3] [4] [5] [6] and many weighty others.

The action of operators $\mathbf{A}, \mathbf{B}, \dots$ onto vectors $\mathbf{x}, \mathbf{y}, \dots$ from the left and co-vectors $\tilde{\mathbf{x}}, \tilde{\mathbf{y}}, \dots$ from the right we denote by

$$\mathbf{y} = \mathbf{A}\mathbf{x}, \leftrightarrow y^i = A_j^i x^j, \quad \tilde{\mathbf{z}} = \tilde{\mathbf{y}}\mathbf{A}, \leftrightarrow \tilde{z}_j = \tilde{y}_i A_j^i. \quad (4)$$

A bilinear form $\tilde{\mathbf{y}}\mathbf{A}\mathbf{x}$ in vectors $\mathbf{x}$ and co-vectors $\tilde{\mathbf{y}}$ is then

$$\tilde{\mathbf{y}}\mathbf{A}\mathbf{x} = \tilde{y}_i A_j^i x^j. \quad (5)$$

Operators in arbitrary n -dimensional spaces obey a Hamilton-Cayley identity which in cases of $n = 1, 2, 3$ takes on the forms (e.g., [6] [7])

$$\begin{aligned} 1\text{-dim.: } 0 &= \mathbf{A} - \langle \mathbf{A} \rangle \mathbf{I}, \\ 2\text{-dim.: } 0 &= \mathbf{A}^2 - \langle \mathbf{A} \rangle \mathbf{A} + [\mathbf{A}] \mathbf{I}, \\ 3\text{-dim.: } 0 &= \mathbf{A}^3 - \langle \mathbf{A} \rangle \mathbf{A}^2 + [\mathbf{A}] \mathbf{A} - |\mathbf{A}| \mathbf{I}, \end{aligned} \quad (6)$$

where $\mathbf{I} \equiv \mathbf{A}^0$ is the corresponding identity operator in considered dimension. Unspecifically for dimension, we introduced for the first three invariants with respect to similarity transformation $\mathbf{A}' = \mathbf{B}\mathbf{A}\mathbf{B}^{-1}$ notations by special symbols according to (other notations, e.g., Fyodorov [6] [7] and Lagally [8])

$$\begin{aligned} \langle \mathbf{A} \rangle &\equiv A_i^i, \quad (\text{trace}), \\ [\mathbf{A}] &\equiv \frac{1}{2}(\langle \mathbf{A} \rangle^2 - \langle \mathbf{A}^2 \rangle), \quad (\text{second invariant}), \\ |\mathbf{A}| &\equiv \frac{1}{6}(\langle \mathbf{A} \rangle^3 - 3\langle \mathbf{A} \rangle \langle \mathbf{A}^2 \rangle + 2\langle \mathbf{A}^3 \rangle), \quad (\text{determinant}). \end{aligned} \quad (7)$$

The second invariant in the form (7) is at once the determinant of two-dimensional operators. By introduction of new higher operator invariants this series may be continued to higher dimension but in this case it is difficult to invent for them new bracket symbols (but $\|\mathbf{A}\|$ for determinant seems to be acceptable to add for 4D-case and $|\mathbf{A}|$ becomes then third invariant and inversely $\|\mathbf{A}\|$ va-

nishes in 3D-case) and an index numbering for invariants is a possible solution. The Hamilton-Cayley identities allow to introduce specifically for dimension complementary operators $\bar{A}$ to A and with their help the corresponding inverse operators A^{-1} in the cases $n=1,2,3$ are

$$A^{-1} = \frac{I}{\langle A \rangle} \equiv \frac{\bar{A}}{\langle A \rangle}, \quad A^{-1} = \frac{\langle A \rangle I - A}{[A]} \equiv \frac{\bar{A}}{[A]}, \quad A^{-1} = \frac{[A]I - \langle A \rangle A + A^2}{|A|} \equiv \frac{\bar{A}}{|A|}. \quad (8)$$

Symmetric and antisymmetric parts of an operators A can be determined in spaces with a given invariant symmetric scalar product (Euclidean or pseudo-Euclidean ones with a nonsingular metric tensor $g_{ij} = g_{ji}$, $|g| \neq 0$) from

$$A_{ik} \equiv g_{ij} A_k^j \quad \text{in coordinate form} \quad A_{ik} = B_{ik} + C_{ik} \quad \text{with} \quad B_{ik} = \frac{1}{2}(A_{ik} + A_{ki}) = B_{ki},$$

$$C_{ik} = \frac{1}{2}(A_{ik} - A_{ki}) = -C_{ki} \quad \text{or in index form (upper index } \tau \text{ means transposition } A_k^\tau \equiv A_{ki})$$

$$A = B + C, \quad B = \frac{1}{2}(A + A^\tau) = B^\tau, \quad C = \frac{1}{2}(A - A^\tau) = -C^\tau. \quad (9)$$

Specifically, three-dimensional antisymmetric second-rank tensors $C_{ik} = -C_{ki}$ can be mapped in unique way onto (axial) vectors c^j according to

$$C = -C^\tau : C \equiv [c], \quad \text{or} \quad C_{ik} = \varepsilon_{ijk} c^j \Leftrightarrow c^j = \frac{1}{2} \varepsilon^{ijk} C_{ik}, \quad (10)$$

with $[c]$ built from vector c as alternative notation for the antisymmetric operator C with correspondences

$$Cx = [c]x = [c, x], \quad yC = y[c] = [y, c], \quad yCx = y[c]x = [y, c, x]. \quad (11)$$

In considered spaces we can relate in unique way covariant vectors $\tilde{y}$ with contravariant vectors y , in coordinate form by $\tilde{y}_i = g_{ij} y^j$, and may define quadratic forms $\tilde{x}Bx \equiv xBx$ according to

$$\tilde{x}Bx = x^i g_{ij} B_k^j x^k = \tilde{x}_j B_k^j x^k = x^i B_{ik} x^k. \quad (12)$$

The explained notations are convenient for two- and three-dimensional coordinate-invariant calculations¹.

1. Introduction

The aim of present article is to review and to continue to develop mathematical

¹In the 5-th edition of the remarkable monograph of Lagally [8] (my first book to this topics) revised by the Editor W. Franz he changed older notations in favor of the notations of Gibbs (mostly used in English literature) that was also in the sense of the late Lagally. This is said in the Preface. Now, I find some of the older or some alternative notations more suited for coordinate-invariant calculation, for example, scalar products as well as application of operators onto vectors without a point between the factors but dyadic products with a separating point. It is favorable to denote vector products by square brackets such as the “products” in Lie algebras due to some (not incidentally) analogous relations. In multiple vector products or in combination with other products due to non-associativity brackets are often needed and are then already present and, by experience, written complicated expressions become mostly shorter. Quantum mechanics also does not use a point in scalar products and in the action of operators onto state vectors (e.g., Dirac’s notations). My additional notations for operator invariants make the formulae easier to read. In physical texts with respect to the kind of letters I do not apply very strong rules.

aspects for the classical and quantum-mechanical description of the polarization of two-mode quasilplane quasimonochromatic light beams by means of the Lie group $SU(2)$. In a possible continuation it is intended to apply this to the investigation of the transformation of polarized light beams in reflection and refraction problems and to discuss the quantum-mechanical conditions for unpolarized light beams. Classically and for vacuum which we only consider, it is mostly sufficient to investigate a two-dimensional polarization matrix composed from the electric field and with components perpendicular to beam propagation from which can be determined the Stokes parameters which vanish for unpolarized light beams. Quantum-statistically this can be done from a density operator for, at least, two mutually orthogonal boson modes and to determine their expectation values for the polarization matrix.

There are two well-known realizations of the Lie group $SU(2)$, first, the translation of the classical formula for the angular momentum into a vector operator according to the rules of transition to canonical quantum mechanics and, second, the two-mode realization by the two-dimensional fundamental representation of $SU(2)$ acting onto two quantum-mechanical states which correspond to two orthogonal polarization vectors in the operator representation of the electric field.

The quantum-statistical description of two-mode light polarization rests on the Jordan-Schwinger realization of the Lie algebra operators to $SU(2)$ by quadratic combinations of the boson annihilation and creation operators of the two modes (Jordan 1935 [9], and of Schwinger 1952 [10], the last republished in [11]) in the specific application to polarization. This description was developed in short form by Jauch and Rohrlich [12] in 1955. The monograph of Peřina [13] from 1971 contains a chapter about the polarization properties of light in which the polarization matrix (called coherence matrix) is quantum-statistically defined. In the same year 1971, Prakash and Chandra [14] determined the general form of two-mode unpolarized light beams by its definition of $SU(2)$ invariance and Agarwal [15] determined the quasiprobabilities of such light beams. After some time of stagnation the rigorous quantum-statistical description of light polarization was further developed by many authors and author groups and was reviewed in the article of Luis and Sánchez-Soto [16]. V. Peřinová, A. Lukš and J. Peřina [17] consider and refer the application of $SU(2)$ operators to atomic coherent states. The best-known realization of $SU(2)$ not considered here is the application to quantum-mechanical angular momentum (e.g., [3] [18] [19]).

With coordinate-invariant methods initiated in second half of last century mainly by F.I. Fyodorov [4] [5] [6] we develop some mathematical aspects of the groups $SU(2)$ and for $SO(3, \mathbb{R})$ in parametrization by the three-dimensional vector φ instead of the Euler angles (α, β, γ) that is rarely to find compared with the huge “group” literature from which we cite the early works of Lyubarski [20], Gürsey, [21], Behrends, Dreitlein, Fronsdal and Lee, [22], Wybourne [23], Gilmore [24], Barut and Rączka [25] and Hamermesh [26]. The group $SO(3, \mathbb{R})$

can be with advantage also parameterized by a three-dimensional vector $\mathbf{c}$ which simplifies mainly the composition law for two and more transformations and was developed with a few predecessors also mainly by Fyodorov. This is discussed here in details and applied for the calculation of the invariant (Haar) measure in both mentioned parametrization (Section 16). In Section 5 it is shown how one can determine representation matrices if one knows a complete set of basis operators, in our case for the fundamental two-dimensional representation of $SU(2)$ with the vector parameter $\boldsymbol{\varphi}$. Some important decompositions (disentanglements) of these matrices are calculated in Section 10 and Section 11 considers shortly the representation of $SU(2)$ by quaternions. A comparison of the two mentioned parametrizations by vector $\boldsymbol{\varphi}$ and Euler angles (α, β, γ) is made in **Appendix A**.

2. Short Quantum Description of Two-Mode Light Polarization of Beams by $SU(2)$ Transformations

The group $SU(2)$ can be applied in quantum optics for the theory of polarized and unpolarized light beams and for the lossless beamsplitter, in each case when two amplitudes can be transformed into each other or into two other amplitudes.

First we consider a light beam in an isotropic medium (here vacuum) which may be composed of two partial beams with two possible polarizations described by normalized, in general, complex-valued polarization vectors $\mathbf{e}_1$ and $\mathbf{e}_2$. The electric field $\mathbf{E}(\mathbf{r}, t)$ of such a beam can be represented in the following way

$$\mathbf{E}(\mathbf{r}, t) = (\mathbf{e}_1 A_1 + \mathbf{e}_2 A_2) e^{i(k_0 r - \omega_0 t)} + (\mathbf{e}_1^* A_1^* + \mathbf{e}_2^* A_2^*) e^{-i(k_0 r - \omega_0 t)}, \quad \omega_0 = c|\mathbf{k}_0|. \quad (2.1)$$

The mean value of the wave vector of the beam is denoted by $\mathbf{k}_0$ and its mean frequency by ω_0 . Both are assumed here to be real-valued. The complex-valued amplitudes to two possible (mean) polarizations $\mathbf{e}_1$ and $\mathbf{e}_2$ are denoted by A_1 and A_2 and they depend slowly from position and time $(\mathbf{r}, t)$ that we do not explicitly write since it does not play a role in our further considerations. In quantum-mechanical context the electric field $\mathbf{E}(\mathbf{r}, t)$ becomes an operator and, correspondingly, the amplitudes $((A_\mu, A_\mu^*), \mu = 1, 2)$ too proportional to pairs of boson annihilation creation operators $((a_\mu, a_\mu^\dagger), \mu = 1, 2)$. The polarization vectors are orthogonal to the wave vector $\mathbf{k}_0 = \mathbf{k}_0^*$ and should satisfy the orthonormalization conditions

$$\mathbf{k}_0 \mathbf{e}_\mu = 0, \quad \mathbf{e}_\mu \mathbf{e}_\nu^* = \delta_{\mu\nu}, \quad (\mu, \nu = 1, 2). \quad (2.2)$$

The description of the beam polarization by vectors $\mathbf{e}_1$ and $\mathbf{e}_2$ is not obligatory and we can use also other two polarization vectors $\mathbf{e}'_1$ and $\mathbf{e}'_2$ which are connected with the primary ones and satisfying the relations (2.2) by

$$(\mathbf{e}'_1, \mathbf{e}'_2) = (\mathbf{e}_1, \mathbf{e}_2) S = (\mathbf{e}_1, \mathbf{e}_2) \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} = (\mathbf{e}_1 S_{11} + \mathbf{e}_2 S_{21}, \mathbf{e}_1 S_{12} + \mathbf{e}_2 S_{22}), \quad (2.3)$$

The general transformation S of this kind proves to be a two-dimensional uni-

tary unimodular transformation. Then the new amplitudes A'_1 and A'_2 to polarization vectors $\mathbf{e}'_1$ and $\mathbf{e}'_2$ in the beam are connected with the primary ones by

$$(\mathbf{e}'_1, \mathbf{e}'_2) \begin{pmatrix} A'_1 \\ A'_2 \end{pmatrix} = (\mathbf{e}_1, \mathbf{e}_2) \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} \begin{pmatrix} A_1 \\ A_2 \end{pmatrix} = (\mathbf{e}_1, \mathbf{e}_2) \begin{pmatrix} A_1 \\ A_2 \end{pmatrix}, \quad (2.4)$$

and the primary amplitudes are connected with the new amplitudes by the same matrix S but in the way

$$\begin{pmatrix} A_1 \\ A_2 \end{pmatrix} = S \begin{pmatrix} A'_1 \\ A'_2 \end{pmatrix} = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} \begin{pmatrix} A'_1 \\ A'_2 \end{pmatrix} = \begin{pmatrix} S_{11}A'_1 + S_{12}A'_2 \\ S_{21}A'_1 + S_{22}A'_2 \end{pmatrix}. \quad (2.5)$$

The two complex polarization vectors $(\mathbf{e}_1, \mathbf{e}_2)$ contain 4 real components and by conditions (2.2) they are restricted to 3 independent real components. The general state of polarization of a light beam is a partially polarized state. An unpolarized state is a state which remains invariant with respect to all transformations S of the $SU(2)$ -group. From Section 5 on we consider a parametrization of the unitary unimodular matrices S by a three-dimensional vector $\boldsymbol{\varphi}$.

Another device where the $SU(2)$ transformations may be applied with advantage is a beamsplitter between two isotropic media. In this case we have an incident wave A_1^i from medium “1” and (or) an incident wave A_2^i from medium “2” of coupled modes from which result a reflected or refracted wave A_1^r and A_2^r in media “1” and “2” in dependence from which side we see the incident wave. In this case it is interesting to consider, in general, polarized incident waves and to calculate from them the reflected and refracted waves. The transformations of the beam are here actively made but inner details of the action of beamsplitter (e.g., a layer) are not involved. The correspondences to a singular beam with two possible polarizations are here

$$(A_1, A_2) \leftrightarrow (A_1^i, A_2^i), \quad (A_1^r, A_2^r) \leftrightarrow (A_1^r, A_2^r). \quad (2.6)$$

One may, however, consider in case of a beamsplitter only one incident wave from medium “1” or medium “2” with two possible polarizations but in this case we have a reflected or refracted partially polarized wave in both media that means, at least, 6 possible polarizations (or 8 in case of incident waves from both sides) and the consideration are not in full analogy to a simple beam.

3. The Group $SU(2)$ Embedded into the Symplectic Group $Sp(4, \mathbb{R})$ of All Quadratic Combinations of Two Pairs of Boson Operators

Lie algebras were created as the tool to describe the local properties of Lie groups in the neighborhood of the identity element. It seems that pairs of boson annihilation and creation operators in quadratic combinations can be considered as elementary building stones of series of Lie algebras, in particular $A_n = SU(n+1)$, $B_n = SO(2n+1, \mathbb{R})$, $C_n = Sp(2n, \mathbb{R})$, $D_n = SO(2n, \mathbb{R})$ (e.g., [23] [24] [25] and many others). For a more general insight into the Lie group of polarization

transformations of two modes which leads to a realization of the two-dimensional unitary unimodular group $SU(2)$ it is favorable to embed them into the group of all possible transformations made by quadratic combinations of two pairs of annihilation and creation operators of a two-mode system.

Pairs of boson annihilation and creation operators $(a_\mu, a_\mu^\dagger)$, $(\mu = 1, \dots, n)$ comprise $2n$ operators which obey the following commutation relations (I identity operator in (Hilbert) representation space)

$$[a_\mu, a_\nu^\dagger] = \delta_{\mu\nu} I, \quad [a_\mu, a_\nu] = [a_\mu^\dagger, a_\nu^\dagger] = 0, \quad (\mu, \nu = 1, \dots, n). \quad (3.1)$$

These boson annihilation and creation operators result from (Hermitean) canonical operators Q_μ and P_μ by the definitions

$$a_\mu \equiv \frac{Q_\mu + iP_\mu}{\sqrt{2\hbar}}, \quad a_\mu^\dagger \equiv \frac{Q_\mu - iP_\mu}{\sqrt{2\hbar}}, \quad (Q_\mu, P_\mu) = (Q_\mu^\dagger, P_\mu^\dagger), \quad (3.2)$$

with the commutation relations

$$[Q_\mu, P_\nu] = i\hbar I \delta_{\mu\nu}, \quad (\mu, \nu = 1, 2, \dots, n). \quad (3.3)$$

The relations (3.1), (3.2) and (3.3) form the Heisenberg-Weyl algebra of an n -mode system. As important partial set of quadratic combinations of n pairs of boson operators may be considered the number operators N_i defined by

$$N_\mu \equiv a_\mu^\dagger a_\mu = a_\mu a_\mu^\dagger - I, \quad (\mu = 1, \dots, n), \quad N \equiv N_1 + N_2 + \dots + N_n, \quad (3.4)$$

The operator N is the total number operator and is the sum of all partial number operators.

Due to our main interest here for the description of polarized and unpolarized light beams we consider now two pairs of boson annihilation and creation operators $(a_1, a_1^\dagger)$ and $(a_2, a_2^\dagger)$. We introduce three Hermitean operators (J_1, J_2, J_3) by the definitions (e.g., Schwinger [10], p. 545, Jauch and Rohrlich [12], pp. 40-49, Messiah [19], Chap. XIII)

$$\begin{aligned} J_1 &\equiv \frac{1}{2}(a_1 a_2^\dagger + a_1^\dagger a_2) = \frac{1}{2\hbar}(Q_1 Q_2 + P_1 P_2) = J_1^\dagger, \\ J_2 &\equiv \frac{i}{2}(a_1 a_2^\dagger - a_1^\dagger a_2) = \frac{1}{2\hbar}(Q_1 P_2 - P_1 Q_2) = J_2^\dagger, \\ J_3 &\equiv \frac{1}{2}(a_1 a_1^\dagger - a_2 a_2^\dagger) = \frac{1}{4\hbar}(Q_1^2 + Q_2^2 - P_1^2 - P_2^2) \equiv \frac{1}{2}(N_1 - N_2) = J_3^\dagger. \end{aligned} \quad (3.5)$$

They satisfy the abstract commutation relations for a Lie algebra $su(2)$

$$[J_2, J_3] = iJ_1, \quad [J_3, J_1] = iJ_2, \quad [J_1, J_2] = iJ_3, \quad \Leftrightarrow \quad [J_i, J_j] = i\varepsilon_{ijk} J_k, \quad (3.6)$$

where ε_{ijk} is the Levi-Civita symbol (or Levi-Civita pseudo-tensor) and here ad once they are the structure coefficients of the Lie algebra $su(2)$. From (3.6) follows

$$J_k = -\frac{i}{2}\varepsilon_{ijk} [J_i, J_j] = -\frac{i}{2}\varepsilon_{kij} (J_i J_j - J_j J_i) = -i\varepsilon_{kij} J_i J_j, \quad (3.7)$$

and therefore (the operator $\varepsilon_{ijk} J_i J_j J_k$ is analogous to the fully antisymmetric volume product)

$$\mathbf{J}^2 \equiv J_k J_k = -i \varepsilon_{ijk} J_i J_j J_k. \quad (3.8)$$

In addition to (3.5) we introduce raising and lowering operators J_- and J_+ as usual by

$$J_- \equiv J_1 - iJ_2 = a_1 a_2^\dagger, \quad J_+ \equiv J_1 + iJ_2 = a_1^\dagger a_2, \quad J_\pm = J_\mp^\dagger, \quad (3.9)$$

which satisfy the commutation relations

$$[J_3, J_\pm] = \pm J_\pm, \quad [J_+, J_-] = 2J_3. \quad (3.10)$$

The operator C defined by (see also (3.8))

$$C \equiv \mathbf{J}^2 = (J_1)^2 + (J_2)^2 + (J_3)^2 = \frac{1}{2}(J_+ J_- + J_- J_+) + (J_3)^2, \quad (3.11)$$

commutes with all operators of the Lie algebra $su(2)$

$$[\mathbf{J}^2, J_k] = 0, \quad \Rightarrow \quad [\mathbf{J}^2, J_\pm] = 0, \quad [\mathbf{J}^2, J_3] = 0, \quad (3.12)$$

and is usually taken as the Casimir operator to the Lie algebra $su(2)$. It can be represented after substitution of (J_1, J_2, J_3) according to (3.5) using the number operators defined in (3.4) in the following way

$$C \equiv \mathbf{J}^2 = \frac{N}{2} \left(\frac{N}{2} + I \right). \quad (3.13)$$

For the finite-dimensional irreducible representations which are parameterized by a discrete parameter j it is proportional to the identity operator I of the corresponding representation space of dimension $n = 2j + 1$ and for these representations it can be specialized to

$$\mathbf{J}^2 = j(j+1)I, \quad \left(j = 0, \frac{1}{2}, 1, \frac{3}{2}, \dots \right). \quad (3.14)$$

The Casimir operator C does not belong to operators of the Lie algebra $su(2)$. The number operator $N \equiv N_1 + N_2$ also commutes with all operators of the Lie algebra $su(2)$

$$[N, J_k] = 0, \quad (3.15)$$

and can be taken in addition to the operators of $su(2)$ forming the Lie algebra $u(2)$ to unitary transformations of the group $U(2)$.

With the 4 annihilation and creation operators $(a_1, a_2, a_1^\dagger, a_2^\dagger)$ one may form 16 quadratic combination from which only 10 are linearly independent. These are the 4 squared operators $(a_1^2 \equiv 2K_{1,-}, a_2^2 \equiv 2K_{2,-}, a_1^{\dagger 2} \equiv 2K_{1,+}, a_2^{\dagger 2} \equiv 2K_{2,+})$ and the remaining 12 squared operators reduce to 6 independent operators due to the mutual commutation relations that means totally to 10 independent operators.

From the quadratic combinations we may separate the Lie algebra of the two-mode squeezing group determined by the following three basic Hermitean operators (K_1, K_2, K_0)

$$K_1 \equiv \frac{1}{2}(a_1 a_2 + a_1^\dagger a_2^\dagger) = \frac{1}{2\hbar}(Q_1 Q_2 - P_1 P_2) = K_1^\dagger,$$

$$\begin{aligned}
K_2 &\equiv \frac{i}{2}(a_1 a_2 - a_1^\dagger a_2^\dagger) = -\frac{1}{2\hbar}(Q_1 P_2 + P_1 Q_2) = K_2^\dagger, \\
K_0 &\equiv \frac{1}{2}(a_1 a_1^\dagger + a_2^\dagger a_2) = \frac{1}{2}(a_1^\dagger a_1 + a_2 a_2^\dagger) \\
&= \frac{1}{4\hbar}(Q_1^2 + P_1^2 + Q_2^2 + P_2^2) = \frac{1}{2}(N_1 + N_2 + I) = K_0^\dagger.
\end{aligned} \tag{3.16}$$

They satisfy the abstract commutation relations for a Lie algebra $su(1,1)$

$$[K_1, K_2] = -iK_0, \quad [K_2, K_0] = iK_1, \quad [K_0, K_1] = iK_2, \tag{3.17}$$

In addition to the three operators (K_1, K_2, K_0) we introduce in analogy to $su(2)$ the operators K_- and K_+ by linear combinations

$$K_- \equiv K_1 - iK_2 = a_1 a_2, \quad K_+ \equiv K_1 + iK_2 = a_1^\dagger a_2^\dagger, \quad K_\pm = K_\pm^\dagger, \tag{3.18}$$

which satisfy the commutation relations

$$[K_0, K_\pm] = \pm K_\pm, \quad [K_-, K_+] = 2K_0. \tag{3.19}$$

By the formal substitutions

$$J_1 \leftrightarrow iK_1, \quad J_2 \leftrightarrow iK_2, \quad J_3 \leftrightarrow K_0, \quad J_+ \leftrightarrow iK_+, \quad J_- \leftrightarrow iK_-, \tag{3.20}$$

they make the transition to the commutation relations (3.6) for a Lie algebra $su(2)$. The operator C' defined by

$$C' \equiv \mathbf{K}^2 = K_0^2 - K_1^2 - K_2^2 = K_0^2 - \frac{1}{2}(K_- K_+ + K_+ K_-), \tag{3.21}$$

commutes with all operators K_k of the Lie algebra $su(1,1)$

$$[\mathbf{K}^2, K_k] = 0, \quad \Rightarrow \quad [\mathbf{K}^2, K_\pm] = 0, \quad [\mathbf{K}^2, K_0] = 0, \tag{3.22}$$

and is usually taken as the Casimir operator of the Lie algebra $su(1,1)$. For irreducible representations it is proportional to the identity operator I of the representation space and is signified by a parameter k (Bargmann index) according to

$$C' \equiv \mathbf{K}^2 = k(k-1)I. \tag{3.23}$$

In the realization of $su(1,1)$ by (3.16) the operator C' takes on the form

$$C' = \frac{1}{4}\{(N_1 - N_2)^2 - I\} = \left(\frac{|N_1 - N_2| + I}{2}\right)\left(\frac{|N_1 - N_2| - I}{2}\right), \tag{3.24}$$

which using (3.5) can be also represented

$$C = \left(J_3 + \frac{1}{2}I\right)\left(J_3 - \frac{1}{2}I\right). \tag{3.25}$$

In comparison, the Casimir operator C of $su(2)$ in (3.13) can be represented by

$$C' = \left(K_0 - \frac{1}{2}I\right)\left(K_0 + \frac{1}{2}I\right). \tag{3.26}$$

The operator J_3 in the realization (3.5) commutes with all operators K_k of the Lie algebra $su(1,1)$ in the realizations (3.16) and the operator K_0 with all operators J_l of the Lie algebra $su(2)$ and we have

$$[J_3, K_k] = 0, \quad [K_0, J_l] = 0. \quad (3.27)$$

The two sets of 4 operators (J_1, J_2, J_3, K_0) and (K_1, K_2, K_0, J_3) form the extended Lie algebras $u(2)$ and $u(1,1)$ of only unitary but not necessarily unimodular operators, respectively.

The 6 operators $(J_-, J_+, J_3, K_-, K_+, K_0)$ are not closed as a Lie algebra since the commutators between the $su(2)$ operators (J_-, J_+, J_3) and the $su(1,1)$ operators (K_-, K_+, K_0) lead to new operators according to

$$\begin{aligned} [J_-, K_-] &= -a_1^2 \equiv -2K_{1,-}, & [J_+, K_+] &= a_1^{\dagger 2} \equiv 2K_{1,+}, \\ [J_+, K_-] &= -a_2^2 \equiv -2K_{2,-}, & [J_-, K_+] &= a_2^{\dagger 2} \equiv 2K_{2,+}. \end{aligned} \quad (3.28)$$

which cannot be represented as linear combinations of these 6 operators. Closing them to a new Lie algebra can be obtained by adding two new groups of operators $(K_{\mu,1}, K_{\mu,2}, K_{\mu,0})$, $(\mu=1,2)$, formed from the pairs of annihilation and creation operators $(a_\mu, a_\mu^\dagger)$, $(\mu=1,2)$,

$$K_{\mu,1} = \frac{1}{4}(a_\mu^2 + a_\mu^{\dagger 2}), \quad K_{\mu,2} = \frac{i}{4}(a_\mu^2 - a_\mu^{\dagger 2}), \quad K_{\mu,0} = \frac{1}{4}(a_\mu a_\mu^\dagger + a_\mu^\dagger a_\mu). \quad (3.29)$$

From (3.29) we combine the operators $K_{\mu,1}$ and $K_{\mu,2}$ to new operators $K_{\mu,-}$ and $K_{\mu,+}$ according to $(\mu=1,2)$

$$K_{\mu,-} \equiv K_{\mu,1} - iK_{\mu,2}, \quad K_{\mu,+} \equiv K_{\mu,1} + iK_{\mu,2}, \quad (3.30)$$

leading explicitly to

$$\begin{aligned} K_{1,-} &\equiv \frac{1}{2}a_1^2, & K_{1,+} &\equiv \frac{1}{2}a_1^{\dagger 2}, & K_{1,0} &\equiv \frac{1}{4}(a_1 a_1^\dagger + a_1^\dagger a_1) = \frac{1}{2}\left(N_1 + \frac{1}{2}I\right) \equiv \frac{1}{2}N'_1, \\ K_{2,-} &\equiv \frac{1}{2}a_2^2, & K_{2,+} &\equiv \frac{1}{2}a_2^{\dagger 2}, & K_{2,0} &\equiv \frac{1}{4}(a_2 a_2^\dagger + a_2^\dagger a_2) = \frac{1}{2}\left(N_2 + \frac{1}{2}I\right) \equiv \frac{1}{2}N'_2. \end{aligned} \quad (3.31)$$

They satisfy the commutation relations of an $su(1,1)$ -algebra in analogy to (3.19)

$$[K_{\mu,0}, K_{\mu,\pm}] = \pm K_{\mu,\pm}, \quad [K_{\mu,-}, K_{\mu,+}] = 2K_{\mu,0}, \quad (\mu=1,2). \quad (3.32)$$

For completeness there remains to calculate the commutation relation of $(K_{\mu,0}, K_{\mu,+}, K_{\mu,-})$ with the operators (K_0, K_+, K_-) and with (J_+, J_-, J_3) for which we find

$$\begin{aligned} [K_0, K_{\mu,0}] &= 0, & [K_0, K_{\mu,\pm}] &= \pm K_{\mu,\pm}, \\ [K_+, K_{\mu,0}] &= -\frac{1}{2}K_+, & [K_+, K_{\mu,+}] &= 0, & [K_+, K_{1,-}] &= -J_-, \\ [K_+, K_{2,-}] &= -J_+, & [K_-, K_{\mu,0}] &= +\frac{1}{2}K_-, & [K_-, K_{\mu,-}] &= 0, \\ [K_-, K_{1,+}] &= +J_+, & [K_-, K_{2,+}] &= +J_-. \end{aligned} \quad (3.33)$$

and

$$\begin{aligned} [J_3, K_{\mu,0}] &= 0, & [J_3, K_{1,\pm}] &= \pm K_{1,\pm}, & [J_3, K_{2,\pm}] &= \mp K_{2,\pm}, \\ [J_\mp, K_{1,0}] &= \pm \frac{1}{2}J_\mp, & [J_\mp, K_{2,0}] &= \mp \frac{1}{2}J_\mp. \end{aligned} \quad (3.34)$$

The operators $K_{1,0}$ and $K_{2,0}$ are connected with J_3 and K_0 by

$$J_3 = K_{1,0} - K_{2,0} = \frac{1}{2}(N_1 - N_2), \quad K_0 = K_{1,0} + K_{2,0} = \frac{1}{2}(N_1 + N_2 + I). \quad (3.35)$$

Thus the operators $K_{1,0}$ and $K_{2,0}$ are already contained as linear combinations of the operators J_3 and K_0 or N_1 and N_2 and must not separately be taken into account for closing the Lie algebra of the above mentioned 6 operators. The operators $(K_{\mu,-}, K_{\mu,+}, K_{\mu,0}), (\mu = 1, 2)$ form two bases of two Lie algebras $su(1,1)$ of squeezing operators of the two modes with indices “1” and “2” separately.

The 10 operators $(J_-, J_+, J_3, K_-, K_+, K_0, K_{1,-}, K_{1,+}, K_{2,-}, K_{2,+})$ form a possible basis of the 10-parameter Lie algebra $sp(4, \mathbb{R})$ with $\frac{10 \times 9}{2} = 45$ independent commutation relations. An informative overview about the structure of the Lie algebra with its commutation relations give the root diagrams that we investigate in the next Section.

4. The Root Diagrams for the Homogeneous and Inhomogeneous Symplectic Group $Sp(4, \mathbb{R})$ of Two Pairs of Boson Operators

For a first overview about the structure of a Lie group it is very useful to consider the root diagram (e.g., [21] [22] [23] [24] [25]) of its Lie algebra which describes the neighborhood of the identity operator I of the Lie group. We suppose that the compact part of the Lie group in the neighborhood of the identity element can be parameterized by vectors $\xi \equiv (\xi^1, \dots, \xi^d)$ with d independent components where d is called the dimension of the Lie group and we require that the vector $\xi = \mathbf{0}$ describes the identity element. We denote the group operators by $A(\xi)$. In the neighborhood of the identity element $A(\mathbf{0}) = I$ an arbitrary group operator $A(\xi)$ can be expanded in a Taylor series according to

$$A(\xi) = A(\mathbf{0}) + \frac{\partial A}{\partial \xi^i}(\mathbf{0}) \xi^i + \dots \equiv I + X_i \xi^i + \dots, \quad (i = 1, \dots, d). \quad (4.1)$$

The operators $X_i, (i = 1, \dots, d)$ are infinitesimal operators of the Lie group and the set of all possible linear combinations of these operators forms the Lie algebra of dimension d . We give here some definitions and a few basic results to the general theory of Lie algebras and discuss later the root diagrams from Equation (4.16) (refrootdef) on.

The commutator $[X, Y]$ of two arbitrary elements X and Y takes on the role of multiplication in the Lie algebra and the result Z has to be an element of the Lie algebra for closing it ($\langle A \rangle$ denotes the trace of an operator A)

$$Z = [X, Y] \equiv XY - YX, \quad \Rightarrow \quad \langle Z \rangle = 0. \quad (4.2)$$

From the definition of the commutator follows immediately

$$[[X, Y], Z] + [[Y, Z], X] + [[Z, X], Y] = 0, \quad (4.3)$$

which is called the Jacobi identity. It takes on the role of the associative law for,

e.g., the algebra of real and complex numbers or the algebra of matrices. Now we choose a basis of d linearly independent operators E_i in such way that arbitrary operators of the Lie algebra can be represented as

$$X = E_i x^i, \quad Y = E_j y^j, \quad Z = E_k z^k, \quad (i=1, \dots, d), \quad (4.4)$$

where $x^i, y^j, z^k, \dots$ are vector components (in analogy to vectors $\mathbf{x} = \mathbf{e}_i x^i$ and later with operators $\mathbf{A}$ to $\mathbf{y} = \mathbf{A}\mathbf{x} = \mathbf{A}\mathbf{e}_i x^i = \mathbf{e}_k A_i^k x^i = \mathbf{e}_k y^k$). Then from (4.2) we find for the commutator $Z = [X, Y]$

$$E_k z^k = Z = [X, Y] = [E_i x^i, E_j y^j] = [E_i, E_j] x^i y^j \equiv E_k c_{ij}^k x^i y^j, \quad (4.5)$$

where we have introduced coefficients c_{ij}^k by definition

$$[E_i, E_j] \equiv E_k c_{ij}^k, \quad \Rightarrow \quad c_{ij}^k = -c_{ji}^k, \quad (4.6)$$

which are called the structure coefficients of the Lie algebra with respect to the chosen basis E_k . From (4.6) follows

$$E_k z^k = Z = [X, Y] = [E_i, E_j] x^i y^j = E_k c_{ij}^k x^i y^j, \quad \Rightarrow \quad z^k = c_{ij}^k x^i y^j. \quad (4.7)$$

From the Jacobi identity (4.3) follows then for arbitrary three basis operators E_i, E_j, E_l follows then

$$\begin{aligned} 0 &= [[E_i, E_j], E_l] + [[E_j, E_l], E_i] + [[E_l, E_i], E_j] \\ &= [E_m c_{ij}^m, E_l] + [E_m c_{jl}^m, E_i] + [E_m c_{li}^m, E_j] \\ &= E_n (c_{ij}^m c_{ml}^n + c_{jl}^m c_{mi}^n + c_{li}^m c_{mj}^n), \end{aligned} \quad (4.8)$$

and therefore

$$c_{ij}^m c_{ml}^n + c_{jl}^m c_{mi}^n + c_{li}^m c_{mj}^n = 0. \quad (4.9)$$

By contraction over the indices $j = n$ and then interchanging the free indices i and l follow the two equations

$$\begin{aligned} c_{in}^m c_{ml}^n + c_{nl}^m c_{mi}^n + c_{li}^m c_{mn}^n &= 0, \\ c_{ln}^m c_{mi}^n + c_{ni}^m c_{ml}^n + c_{il}^m c_{mn}^n &= 0, \end{aligned} \quad (4.10)$$

where we used that from the antisymmetry of the structure coefficients in the lower indices follows for the sum terms in (4.10)

$$c_{in}^m c_{ml}^n = c_{lm}^n c_{ni}^m = c_{ln}^m c_{mi}^n, \quad c_{nl}^m c_{mi}^n = c_{ml}^n c_{ni}^m = c_{ni}^m c_{ml}^n, \quad c_{li}^m c_{mn}^n = -c_{il}^m c_{mn}^n. \quad (4.11)$$

Forming the difference of the Equations (4.10) and using the symmetry (4.11) the first two sum terms cancel and from the third sum terms using again (4.11) results

$$c_{li}^m c_{mn}^n = c_{il}^m c_{mn}^n = -c_{li}^m c_{mn}^n = 0, \quad (4.12)$$

and each of the equations (10) simplifies to

$$c_{in}^m c_{ml}^n + c_{nl}^m c_{mi}^n = c_{in}^m c_{ml}^n - c_{ln}^m c_{mi}^n = 0, \quad (4.13)$$

from which follows

$$\gamma_{il} \equiv c_{ni}^m c_{ml}^n = c_{nl}^m c_{mi}^n \equiv \gamma_{li}. \quad (4.14)$$

The symmetric tensor $\gamma_{il} = \gamma_{li}$ is called the Killing form and with its help one may define a bilinear symmetric scalar product of two operators $X = E_i x^i$ and $Y = E_l y^l$ written (XY) as follows

$$\begin{aligned}(XY) &\equiv (E_i x^i E_l y^l) = (E_i E_l) x^i y^l \equiv \gamma_{il} x^i y^l = (YX), \\ \gamma_{il} &\equiv (E_i E_l) = (E_l E_i) = \gamma_{li}.\end{aligned}\quad (4.15)$$

The second-rank symmetric tensor γ_{il} is a kind of metric tensor for the Lie algebra and plays an important role for the distinction of different kinds of Lie algebras (e.g., Levy-Maltsev theorem, [23] [25]).

Now comes into play the Cartan subalgebra of the Lie algebra which is the linear space of the maximum of commuting operators of the Lie algebra with operators usually denoted by $\mathbf{H} \equiv (H_1, \dots, H_r)$. The number r of independent operators of the Cartan subalgebra is called the rank of the Lie algebra. From linear combinations of the Lie-algebra operators X_i one may select by linear combinations d operators E_α which are eigenvectors of the operator of the Cartan subalgebra in the sense

$$[\mathbf{H}, E_\alpha] = \alpha E_\alpha, \quad (\text{or } [H_i, E_\alpha] = \alpha_i E_\alpha), \quad (\alpha \equiv (\alpha_1, \dots, \alpha_r)). \quad (4.16)$$

The vectorial eigenvalues α are called the root vectors of the Lie algebra and their dimensionality is equal to the rank of the Lie algebra or the dimension r of the Cartan subalgebra. Only the vectorial eigenvalue $\alpha = \mathbf{0}$ is r -fold degenerate and their root vectors are linear combinations of the operators H_i . The root diagram represents the $d - r$ root vectors in the r -dimensional space plus the r operators H_i of the Cartan subalgebra in the center. The basis operators of the Cartan subalgebra are not uniquely determined and can be defined in different variants ways of the theory. The commuting operators H_i of the Cartan subalgebra plus the root vectors determine already a certain amount of all commutation relations. For the remaining commutation relations of the periphery the theory of Lie algebras derives relations from the Jacobi identity which restrict their possibilities. This is only a minimum of the many well-known relations for Lie algebras (see, e.g., [23] [24] [25] and many others).

For some generalizations we extend the mainly here considered Lie group $SU(2)$. In **Figure 1** we represent the two-dimensional root vectors of the Lie algebra $sp(4, \mathbb{R})$ to the symplectic group $Sp(4, \mathbb{R})$ as arrows in two bases of the Cartan subalgebra (J_3, K_0) and (N_1, N_2) . For example, the arrow in the left-hand partial picture to the operator J_- means the commutator relation $[(J_3, K_0), J_-] = (-1, 0) J_-$ and the right-hand partial picture the commutator relation $[(N_1, N_2), J_-] = (-1, 1) J_-$.

Each pair of boson annihilation and creation vectors $(a, a^\dagger)$ or corresponding canonical operators (Q, P) with the commutation relations $[a, a^\dagger] = I$ or $[Q, P] = i\hbar I$ forms also a Lie algebra called Heisenberg-Weyl algebra which is of dimension zero and reduces to a point. However, if we take in addition to pairs of boson operators of the Heisenberg-Weyl algebra the corresponding number operators $N_i \equiv a_i^\dagger a_i, (i = 1, \dots, n)$ then due to commutation relations

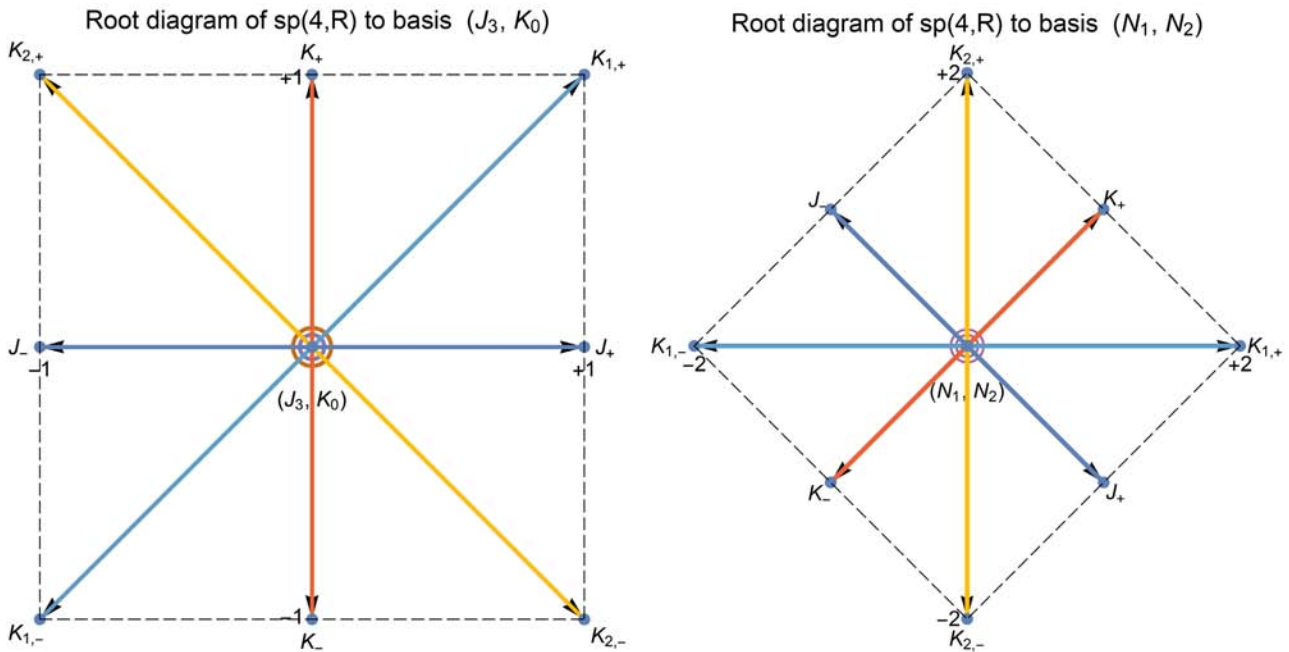


Figure 1. Root diagrams of Lie algebra to the homogeneous symplectic group $Sp(4, \mathbb{R})$ in different bases. It is 10-dimensional and therefore a basis possesses 10 operators. In first basis of the Cartan subalgebra we have

$\left(J_3 = \frac{1}{2}(N_1 - N_2), K_0 = \frac{1}{2}(N_1 + N_2 + I) \sim \frac{1}{2}(N_1 + N_2) \right)$ and in the second basis $(N_1 \equiv a_1^\dagger a_1, N_2 \equiv a_2^\dagger a_2)$ where N_1 and N_2 are the number operators to the two modes. The identity operator I in the Cartan subalgebra in the center of the diagram does not play a role since it commutes with all operators of the Lie group and does not provide a contribution to the roots.

$[N_i, a_j] = -\delta_{ij} a_j$, $[N_i, a_j^\dagger] = +\delta_{ij} a_j^\dagger$, $[N_i, N_j] = 0$ the Heisenberg-Weyl algebras taken together with the number operators also form Lie algebras of corresponding rank. The same is, for example, by combination of Lie algebras $sp(2n, \mathbb{R})$ or $su(n+1)$ representable by pairs of annihilation and creation operators we find new algebras which we call inhomogeneous Lie algebras $i.sp(2n, \mathbb{R})$ or $i.su(n+1)$, respectively.

In **Figure 2** we represent the root diagrams of $i.sp(4, \mathbb{R})$ in different basis systems. Besides the operators of the Cartan subalgebra in the center they contain there the identity operator I for closing them. In **Figure 3** we represent the root diagram of a subalgebra of $i.sp(4, \mathbb{R})$ in **Figure 2** and make from it the transition to the root diagram of the Lie algebra $su(3)$. In **Figure 4** the root diagram for $SU(3)$ of the right-hand **Figure 3** is stretched in direction of the ordinate to the canonical form in a way that it takes on the maximal symmetry of a regular hexagon. One may check that operator $N \equiv N_1 + N_2 + N_3$, $N_1 \equiv a_1^\dagger a_1$, $N_2 \equiv a_2^\dagger a_2$, $N_3 \equiv a_3^\dagger a_3$, commutes with all operators $J_\pm^{(12)}, J_\pm^{(13)}, J_\pm^{(23)}$ to the Lie algebra $su(3)$ constructed from the inhomogeneous group $I.Sp(4, \mathbb{R})$ in described way and illustrated in **Figure 4**

$$[N, J_\pm^{(12)}] = [N, J_\pm^{(13)}] = [N, J_\pm^{(23)}] = 0, \quad (N \equiv N_1 + N_2 + N_3). \quad (4.17)$$

Therefore one may add to the operator $K_0 = \frac{1}{2}(N_1 + N_2 + I)$ a multiple of

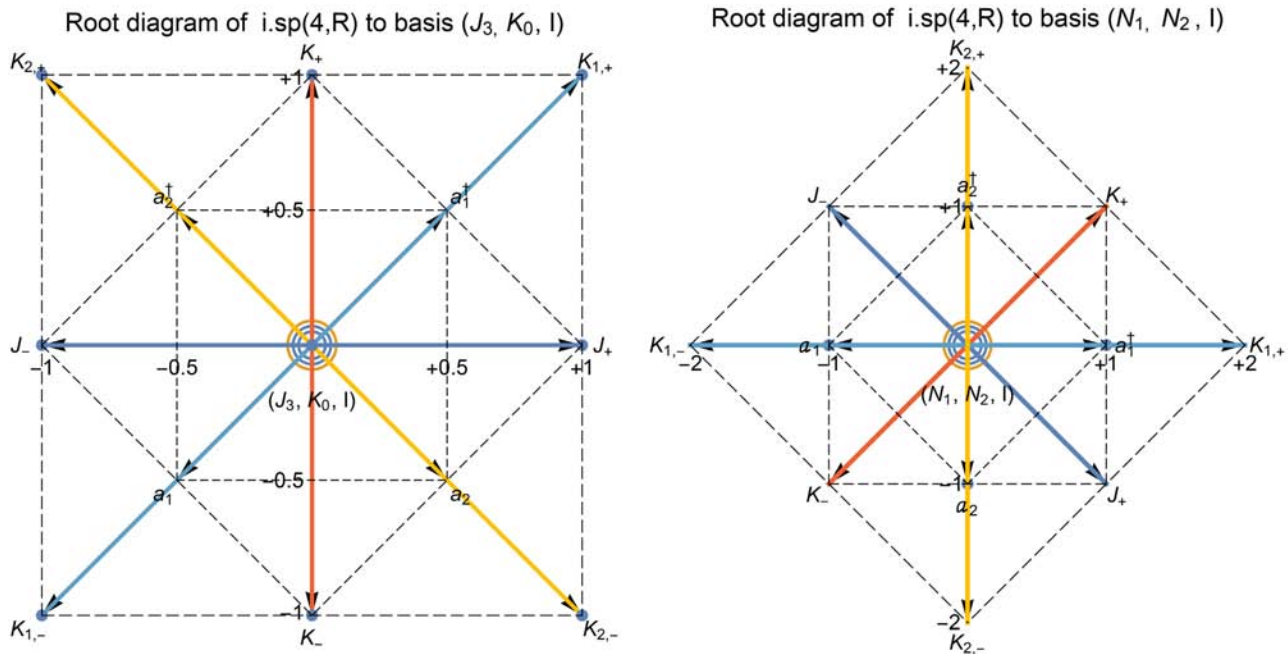


Figure 2. Root diagram of Lie algebra to the inhomogeneous symplectic group $ISp(4, \mathbb{R})$. These root diagrams contain in addition to the homogeneous symplectic group $Sp(4, \mathbb{R})$ the pairs of operators $(a_i, a_i^\dagger)$ and $(a_i, a_i^\dagger)$ with commutation relations $[a_i, a_i^\dagger] = [a_2, a_2^\dagger] = I$. Therefore, the identity operator I belongs to the Lie algebra and since it commutes with all operators it belongs to the Cartan subalgebra in the center of the diagrams. The diagram is quasi two-dimensional since no root operator possesses a component in direction of I perpendicular to the paper plane.

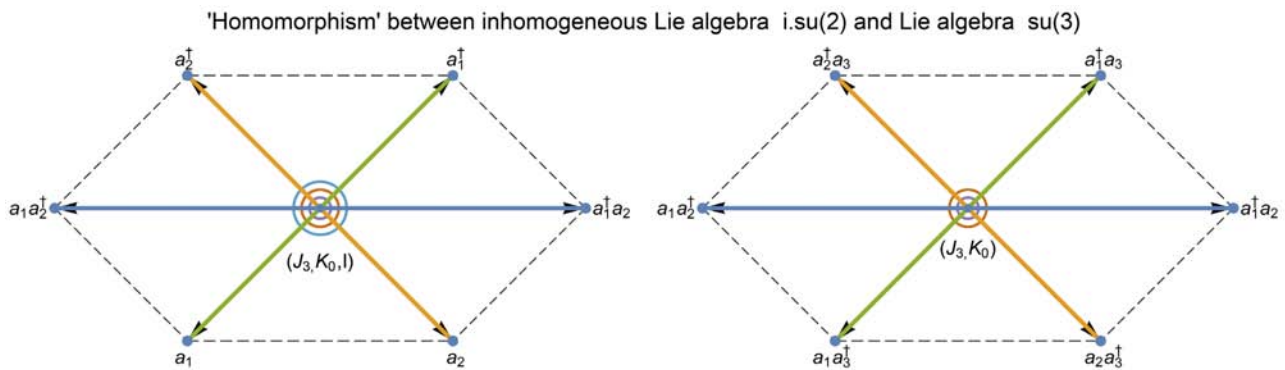


Figure 3. Isomorphism of root diagram to Lie subalgebra of $isp(4, \mathbb{R})$ to the Lie algebra $su(3)$ of homogeneous unitary group $SU(3)$. The root diagram of this Lie algebra on the left-hand side is part of the root diagram in **Figure 2**. The operators in the center are defined by $J_3 \equiv \frac{1}{2}(N_1 - N_2)$, $K_0 \equiv \frac{1}{2}(N_1 + N_2)$ with $N_1 \equiv a_1^\dagger a_1$, $N_2 \equiv a_2^\dagger a_2$. Since the operators J_3 and K_0 commute with the operators $(a_3, a_3^\dagger)$ both root diagrams are equivalent.

$N = N_1 + N_2 + N_3$ and for the same reason a multiple of the identity operator I without changing the root diagram of $su(3)$ that is represented on the right-hand picture of **Figure 4**. If we add to this scheme now the annihilation and creation operators $(a_i, a_i^\dagger)$, $(i=1, 2, 3)$ which do not commute with the operator $N = N_1 + N_2 + N_3$ then we may obtain the root diagram for the Lie algebra to

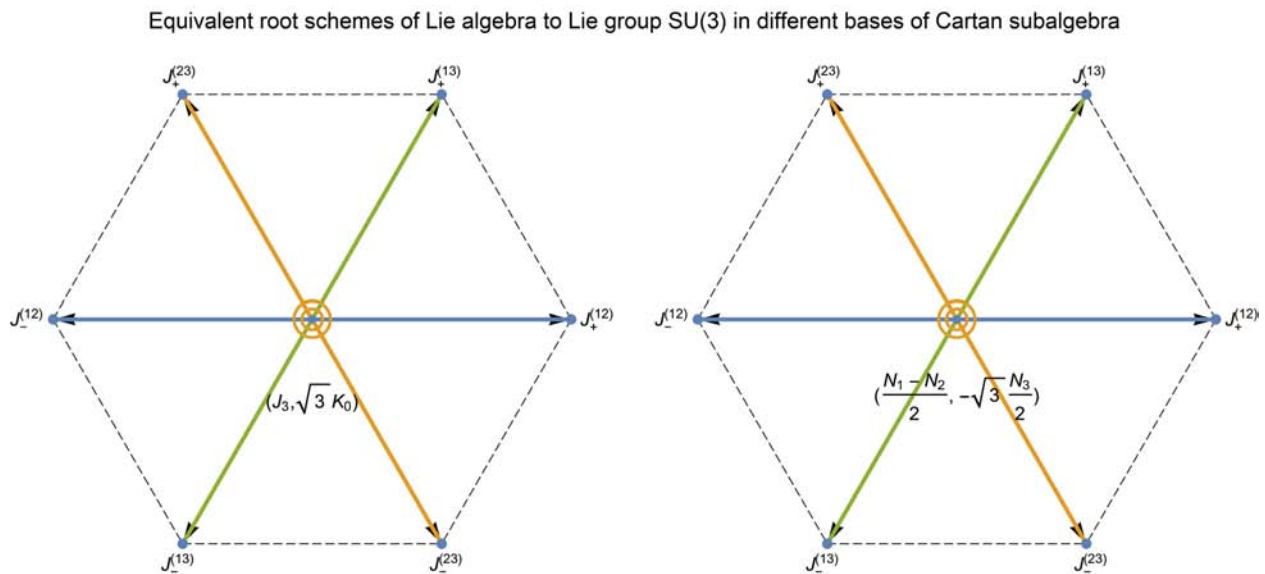


Figure 4. Root diagram of Lie algebra to the unitary unimodular group $SU(3)$ in two equivalent bases. The operators in the center are defined by $J_3 \equiv \frac{1}{2}(N_1 - N_2)$, $K_0 \equiv \frac{1}{2}(N_1 + N_2 + I)$ with $N_1 \equiv a_1^\dagger a_1$, $N_2 \equiv a_2^\dagger a_2$ and additionally $N_3 \equiv a_3^\dagger a_3$. The two schemes are equivalent because the number operator $N \equiv N_1 + N_2 + N_3$ commutes with all operators of the Lie algebra to $SU(3)$ in considered realization and thus the operators of the Cartan subalgebra in the center can be substituted using the given identity. These diagrams correspond to the decontorted right-hand diagram of **Figure 3** in perpendicular direction which leads to highest symmetry of the root diagram for $SU(3)$ that becomes clear if we take into account $J_-^{(12)} = a_1 a_2^\dagger$, $J_+^{(12)} = a_1^\dagger a_2$, $J_-^{(13)} = a_1 a_3^\dagger$, $J_+^{(13)} = a_1^\dagger a_3$, $J_-^{(23)} = a_2 a_3^\dagger$, $J_+^{(23)} = a_2^\dagger a_3$.

the inhomogeneous group $I.SU(3)$ represented in **Figure 5**. It contains 15 operators, 3 of the Cartan subalgebra in the center and 12 in the periphery and is very similar to the root diagram of the Lie algebra g_2 to the exceptional Lie group G_2 with 14 basis operators (e.g., [23]). The root scheme to the next unitary unimodular group $SU(4)$ is three-dimensional and thus of rank 3. It possesses 3 operators in the center belonging to the Cartan subalgebra and 12 in the periphery which with their tips form in the most symmetric way the 12 corners of a cuboctahedron which is a semi-regular polyhedron.

In cases when the identity operator I belongs to the Cartan subalgebra in addition to $2n$ annihilation and creation operators (inhomogeneous groups) the root scheme, nevertheless, remains quasi of rank n since operator I commutes with all operators and does not provide a contribution to a higher rank $n+1$.

A conclusion is that by transition from known series of Lie groups to their inhomogeneous partner groups some problems emerge with their root diagrams and algebraic properties (traces, see also **Appendix B**).

5. Fundamental Representation $D^{\left(\frac{1}{2}\right)}$ of $SU(2)$ in Parametrization by a Three-Dimensional Vector φ

We now derive the fundamental representation of $SU(2)$ in the basis of the

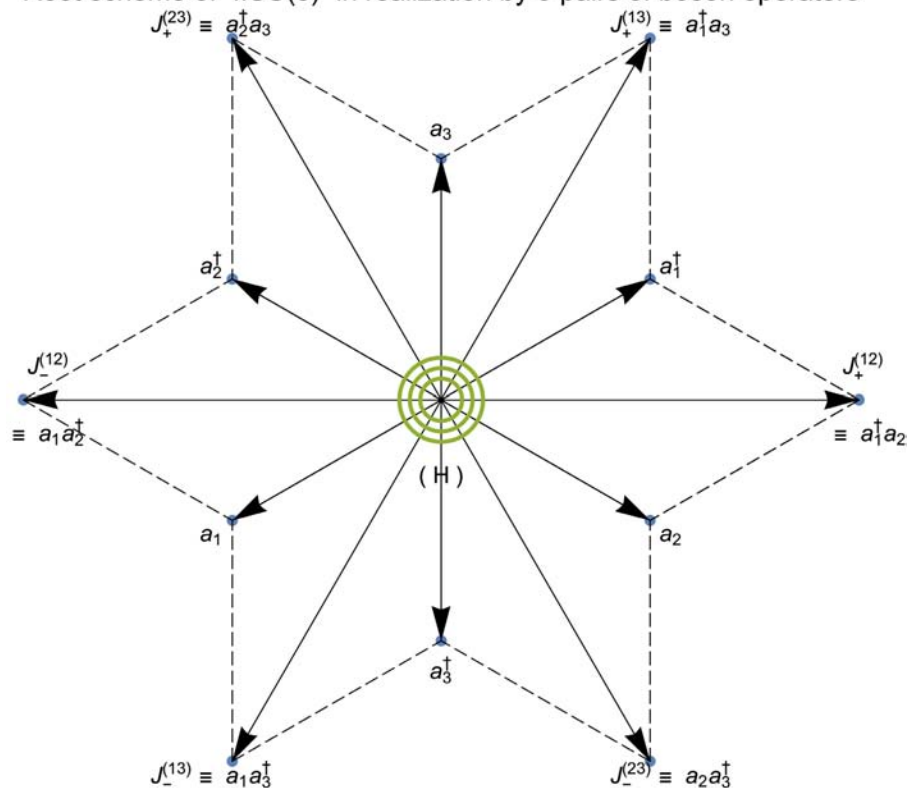
Root scheme of $I.SU(3)$ in realization by 3 pairs of boson operators

Figure 5. Root scheme of Lie algebra of inhomogeneous group $I.SU(3)$ in realization by 3 pairs of boson annihilation and creation operators. The 3 operators (H) of the Cartan subalgebra are defined by $(H) = \left(J \equiv \frac{1}{2}(N_1 - N_2), K \equiv \frac{\sqrt{3}}{6}(N_1 + N_2 - 2N_3), I \right)$ with

$N_1 \equiv a_1^\dagger a_1$, $N_2 \equiv a_2^\dagger a_2$, $N_3 \equiv a_3^\dagger a_3$. The distance from the center to the tips of the star is equal to 1. The diagram is very similar to that of the exceptional group G_2 and becomes the same if we omit the operator I in the center since it does not contribute to a third dimension. On the other side all commutators should belong to the diagram.

operators $(a_1^\dagger, a_2^\dagger)$ and clarify in this way the transformation properties of the involved basic quantities such as boson operators and polarization vectors. For this purpose, we use a real three-dimensional vector parameter $\varphi = (\varphi_1, \varphi_2, \varphi_3)$ and represent the general element x of the Lie algebra $su(2)$ in the following way

$$x \equiv J\varphi = J_1\varphi_1 + J_2\varphi_2 + J_3\varphi_3 = \frac{1}{2}(J_+\varphi_- + J_-\varphi_+) + J_3\varphi_3, \quad (5.1)$$

$$\varphi_k = \varphi_k^*, \quad x = x^\dagger, \quad \varphi_+ \equiv \varphi_1 + i\varphi_2, \quad \varphi_- \equiv \varphi_1 - i\varphi_2.$$

In **Appendix A**, we establish the connection to the Euler angles as parameters. The transition from the Lie algebra $su(2)$ to the Lie group $SU(2)$ and its inversion is made by the exponential mapping

$$x = J\varphi \leftrightarrow X \equiv e^{ix} = \exp(iJ\varphi), \quad \Leftrightarrow \quad x = x^\dagger \leftrightarrow X^{-1} = X^\dagger. \quad (5.2)$$

The construction of the fundamental representation of $SU(2)$ in the basis

of the operators $(a_1^\dagger, a_2^\dagger)$ requires to consider a part of the Lie algebra of the inhomogeneous unitary unimodular group $ISU(2)$ with the operators $(J_1, J_2, J_3, a_1, a_2, a_1^\dagger, a_2^\dagger, I)$ and their commutation relations as a possible set of basis operators. Instead of writing down all commutation relations, we use the necessary ones in the following mapping of (J_1, J_2, J_3) onto matrices (s_1, s_2, s_3) in the two-dimensional fundamental representation of $su(2)$

$$\begin{aligned} [J_1, (a_1^\dagger, a_2^\dagger)] &= (a_1^\dagger, a_2^\dagger) \frac{1}{2} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = (a_1^\dagger, a_2^\dagger) s_1, \quad s_1 = \frac{1}{2} \sigma_1, \\ [J_2, (a_1^\dagger, a_2^\dagger)] &= (a_1^\dagger, a_2^\dagger) \frac{1}{2} \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} = (a_1^\dagger, a_2^\dagger) s_2, \quad s_2 = \frac{1}{2} \sigma_2, \\ [J_3, (a_1^\dagger, a_2^\dagger)] &= (a_1^\dagger, a_2^\dagger) \frac{1}{2} \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} = (a_1^\dagger, a_2^\dagger) s_3, \quad s_3 = \frac{1}{2} \sigma_3. \end{aligned} \quad (5.3)$$

The matrices s_k are essentially (multiplied by factor 2) the Pauli spin matrices σ_k which possess the properties

$$\sigma_k \sigma_l = \delta_{kl} I + i \varepsilon_{jkl} \sigma_j, \quad \Rightarrow \quad \sigma_k \sigma_l + \sigma_l \sigma_k = 2 \delta_{kl} I, \quad [\sigma_k, \sigma_l] = i 2 \varepsilon_{jkl} \sigma_j. \quad (5.4)$$

The most direct relation to the Pauli spin matrices is one reason that we construct the fundamental representation of the Lie algebra $su(2)$ in the basis $(a_1^\dagger, a_2^\dagger)$ and not in the adjoint basis (a_1, a_2) .

From (5.3) follows for the representation of the Lie algebra operators $x = J\varphi$

$$[x, (a_1^\dagger, a_2^\dagger)] = (a_1^\dagger, a_2^\dagger) \frac{1}{2} \begin{pmatrix} \varphi_3 & \varphi_- \\ \varphi_+ & -\varphi_3 \end{pmatrix} \quad (5.5)$$

This is the mapping $x \rightarrow \mathbf{x}$ of the operators x on two-dimensional matrices $\mathbf{x}$ according to

$$\mathbf{x} \rightarrow \mathbf{x} = \frac{1}{2} \begin{pmatrix} \varphi_3 & \varphi_- \\ \varphi_+ & -\varphi_3 \end{pmatrix}, \quad \mathbf{x}^2 \rightarrow \mathbf{x}^2 = \frac{1}{4} |\varphi|^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \frac{1}{4} |\varphi|^2 I, \quad (5.6)$$

By means of the well-known operator expansion (e.g., [19])

$$e^A B e^{-A} = \sum_{n=0}^{\infty} \frac{1}{n!} \underbrace{[A, [A, \dots [A, B] \dots]]}_{n \times} = B + [A, B] + \frac{1}{2!} [A, [A, B]] + \dots \quad (5.7)$$

and using the Hamilton-Cayley identity for two-dimensional operators A in (5.6) we find the corresponding mapping $e^{i\mathbf{x}} \rightarrow e^{i\mathbf{x}}$ into the Lie group (note the difference between $\mathbf{x}$ and x)

$$e^{i\mathbf{x}} = \sum_{n=0}^{\infty} \frac{i^n}{n!} \mathbf{x}^n \rightarrow e^{i\mathbf{x}} = \sum_{n=0}^{\infty} \frac{i^n}{n!} x^n, \quad e^{i\mathbf{x}} (a_1^\dagger, a_2^\dagger) e^{-i\mathbf{x}} = (a_1^\dagger, a_2^\dagger) e^{i\mathbf{x}}. \quad (5.8)$$

In described way we obtain from (5.3) the two-dimensional fundamental representation $D^{\left(\frac{1}{2}\right)}$ of $SU(2)$ by unitary unimodular matrices $S \equiv S(\varphi)$ in the basis of creation operators $(a_1^\dagger, a_2^\dagger)$

$$\begin{aligned} (a_1^\dagger, a_2^\dagger) &\equiv \exp(iJ\varphi) (a_1^\dagger, a_2^\dagger) \exp(-iJ\varphi) = (a_1^\dagger, a_2^\dagger) S = (a_1^\dagger, a_2^\dagger) \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} \\ &= (a_1^\dagger S_{11} + a_2^\dagger S_{21}, a_1^\dagger S_{12} + a_2^\dagger S_{22}), \quad [S] = 1. \end{aligned} \quad (5.9)$$

In analogy we find from (5.9) for the transformation of annihilation operators

$$\begin{pmatrix} a'_1 \\ a'_2 \end{pmatrix} \equiv \exp(i\mathbf{J}\boldsymbol{\varphi}) \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \exp(-i\mathbf{J}\boldsymbol{\varphi}) = S^\dagger \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} S_{22} & -S_{12} \\ -S_{21} & S_{11} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} \quad (5.10)$$

$$= \begin{pmatrix} S_{22}a_1 - S_{12}a_2 \\ -S_{21}a_1 + S_{11}a_2 \end{pmatrix}.$$

According to

$$N'_1 + N'_2 \equiv a_1'^\dagger a'_1 + a_2'^\dagger a'_2 = (a_1'^\dagger, a_2'^\dagger) \begin{pmatrix} a'_1 \\ a'_2 \end{pmatrix} = a_1^\dagger a_1 + a_2^\dagger a_2 \equiv N_1 + N_2, \quad (5.11)$$

these transformations possess the total number operator $N \equiv N_1 + N_2$ as basic invariant.

The form of the matrix S is

$$S = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} = \exp(i\mathbf{s}\boldsymbol{\varphi}), \quad (\mathbf{s}\boldsymbol{\varphi} \equiv s_1\varphi_1 + s_2\varphi_2 + s_3\varphi_3), \quad (5.12)$$

and is explicitly found in described way using the Hamilton-Cayley identity and with abbreviations φ_+ and φ_- (compare, e.g., Gilmore [24], p. 150, Equation (6.4¹))

$$S \equiv S(\boldsymbol{\varphi}) = \begin{pmatrix} \cos\left(\frac{|\boldsymbol{\varphi}|}{2}\right) + i\frac{\varphi_3}{|\boldsymbol{\varphi}|}\sin\left(\frac{|\boldsymbol{\varphi}|}{2}\right) & i\frac{\varphi_-}{|\boldsymbol{\varphi}|}\sin\left(\frac{|\boldsymbol{\varphi}|}{2}\right) \\ i\frac{\varphi_+}{|\boldsymbol{\varphi}|}\sin\left(\frac{|\boldsymbol{\varphi}|}{2}\right) & \cos\left(\frac{|\boldsymbol{\varphi}|}{2}\right) - i\frac{\varphi_3}{|\boldsymbol{\varphi}|}\sin\left(\frac{|\boldsymbol{\varphi}|}{2}\right) \end{pmatrix}, \quad SS^\dagger = I, \quad (5.13)$$

$$|\boldsymbol{\varphi}| \equiv \sqrt{\varphi_+^2 + \varphi_-^2 + \varphi_3^2} = \sqrt{\varphi_1^2 + \varphi_2^2 + \varphi_3^2}, \quad \varphi_\pm \equiv \varphi_1 \pm i\varphi_2 = \varphi_\mp^*.$$

The relations of unitarity $SS^\dagger = I$ and of unimodularity $[S] = 1$ in addition are more explicitly ($[S]$ denotes determinant of two-dimensional matrix or operator S)

$$S^{-1} = S^\dagger, \quad \Leftrightarrow \quad S_{11} = S_{22}^*, \quad S_{12} = -S_{21}^*, \quad (5.14)$$

$$[S] = S_{11}S_{22} - S_{12}S_{21} = S_{11}S_{11}^* + S_{12}S_{12}^* = 1.$$

The character of the representation or trace of the representation matrices is ($\langle S \rangle$ denotes trace of a matrix or operator S)

$$\langle S \rangle \equiv S_{11} + S_{22} = 2\cos\left(\frac{|\boldsymbol{\varphi}|}{2}\right) = \frac{\sin(|\boldsymbol{\varphi}|)}{\sin\left(\frac{|\boldsymbol{\varphi}|}{2}\right)}, \quad (5.15)$$

with $\sin\left(\frac{|\boldsymbol{\varphi}|}{2}\right)$ expressed by the matrix elements using the unimodularity and unitarity (5.14) of S

$$\sin\left(\frac{|\boldsymbol{\varphi}|}{2}\right) = \pm \sqrt{1 - \left(\frac{S_{11} + S_{22}}{2}\right)^2} = \pm \sqrt{S_{12}S_{12}^* - \left(\frac{S_{11} - S_{11}^*}{2}\right)^2}. \quad (5.16)$$

We mention still that using

$$\begin{pmatrix} \varphi_3 & \varphi_- \\ \varphi_+ & -\varphi_3 \end{pmatrix}^2 = \begin{pmatrix} \varphi_3^2 + \varphi_- \varphi_+ & 0 \\ 0 & \varphi_+ \varphi_- + \varphi_3^2 \end{pmatrix} = |\boldsymbol{\varphi}|^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad (5.17)$$

and additionally using the identity (5.7) the matrix (5.13) together with the chosen basis $(a_1^\dagger, a_2^\dagger)$ can be straightforwardly derived also as follows

$$\begin{aligned} e^{i\boldsymbol{J}\boldsymbol{\varphi}}(a_1^\dagger, a_2^\dagger)e^{-i\boldsymbol{J}\boldsymbol{\varphi}} &= (a_1^\dagger, a_2^\dagger) \sum_{n=0}^{\infty} \frac{1}{n!} \left(\frac{i}{2}\right)^n \begin{pmatrix} \varphi_3 & \varphi_- \\ \varphi_+ & -\varphi_3 \end{pmatrix}^n \\ &= (a_1^\dagger, a_2^\dagger) \left\{ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \sum_{m=0}^{\infty} \frac{(-1)^m}{(2m)!} \left(\frac{|\boldsymbol{\varphi}|}{2}\right)^{2m} \right. \\ &\quad \left. + i \begin{pmatrix} \varphi_3 & \varphi_- \\ \varphi_+ & -\varphi_3 \end{pmatrix} \frac{1}{|\boldsymbol{\varphi}|} \sum_{m=0}^{\infty} \frac{(-1)^m}{(2m+1)!} \left(\frac{|\boldsymbol{\varphi}|}{2}\right)^{2m+1} \right\} \\ &= (a_1^\dagger, a_2^\dagger) \left\{ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \cos\left(\frac{|\boldsymbol{\varphi}|}{2}\right) + i \begin{pmatrix} \varphi_3 & \varphi_- \\ \varphi_+ & -\varphi_3 \end{pmatrix} \frac{1}{|\boldsymbol{\varphi}|} \sin\left(\frac{|\boldsymbol{\varphi}|}{2}\right) \right\}, \end{aligned} \quad (5.18)$$

that is identical with (5.13). The components of $\boldsymbol{J}$ and their relations to annihilation and creation operators are explained in (3.5) and (3.9) and furthermore $\varphi_{\pm} \equiv \varphi_1 \pm i\varphi_2$ is used.

Other bases, for example, (a_1, a_2) , lead only to a reordering with changing signs in the matrices in (5.13). In explained way one may calculate also other irreducible and reducible representations even with other dimension if one possesses a suited basis.

6. Determination of a Basic Range of Vector Parameter $\boldsymbol{\varphi}$ for $SU(2)$

We now determine a basic range for the vector parameter $\boldsymbol{\varphi}$. Two vector parameters $\boldsymbol{\varphi}$ and $\boldsymbol{\varphi}'$ which lead to the same explicitly given matrix S in (5.13) are called equivalent and this is denoted by $\boldsymbol{\varphi} \sim \boldsymbol{\varphi}'$. We show that the transformations of the parameter $\boldsymbol{\varphi}$ according to

$$\boldsymbol{\varphi} \rightarrow \boldsymbol{\varphi}' = \boldsymbol{\varphi} + n4\pi \frac{\boldsymbol{\varphi}}{|\boldsymbol{\varphi}|} = (|\boldsymbol{\varphi}| + n4\pi) \frac{\boldsymbol{\varphi}}{|\boldsymbol{\varphi}|} \sim \boldsymbol{\varphi}, \quad (n = 0, \pm 1, \pm 2, \dots), \quad (6.1)$$

leave the operators S in (5.13) unchanged. From (6.1) follows for modulus and direction of $\boldsymbol{\varphi}'$

$$|\boldsymbol{\varphi}'| = |\boldsymbol{\varphi}| + n4\pi = \pm(|\boldsymbol{\varphi}| + n4\pi), \quad \Leftrightarrow \quad \frac{\boldsymbol{\varphi}'}{|\boldsymbol{\varphi}'|} = \frac{|\boldsymbol{\varphi}| + n4\pi}{|\boldsymbol{\varphi}| + n4\pi} \frac{\boldsymbol{\varphi}}{|\boldsymbol{\varphi}|} = \pm \frac{\boldsymbol{\varphi}}{|\boldsymbol{\varphi}|}, \quad (6.2)$$

with correlated signs. In connection with $\cos\left(\frac{|\boldsymbol{\varphi}'|}{2}\right) = \cos\left(\frac{|\boldsymbol{\varphi}|}{2}\right)$ and

$\sin\left(\frac{|\boldsymbol{\varphi}'|}{2}\right) = \pm \sin\left(\frac{|\boldsymbol{\varphi}|}{2}\right)$ we see that the transformation (6.1) preserves the matrices

(5.13). The minimal difference of two different equivalent points is obtained if we set $n = \pm 1$ in (6.1) and it is

$$\varphi' - \varphi = \pm 4\pi \frac{\varphi}{|\varphi|}, \quad |\varphi' - \varphi| = 4\pi, \quad \Rightarrow \quad S' = S. \quad (6.3)$$

Such equivalent points possess opposite directions considered from the center. Therefore, one may choose as basic range of inequivalent parameters φ a three-dimensional ball² of radius 2π

$$0 \leq |\varphi| \leq 2\pi. \quad (6.4)$$

Inner points of this three-dimensional ball possess equivalent points only outside it. The whole surface (sphere S^2) of this three-dimensional ball $\varphi' = 2\pi \frac{\varphi}{|\varphi|}$, independently on the direction of $\frac{\varphi}{|\varphi|}$, corresponds to the negatively taken identity matrix $S = -I$ that means topologically to one point of the group manifold.

We consider now the transformation

$$\varphi \rightarrow \varphi' = \varphi + (2n+1)2\pi \frac{\varphi}{|\varphi|} \sim \varphi + 2\pi \frac{\varphi}{|\varphi|} = (|\varphi| + 2\pi) \frac{\varphi}{|\varphi|}, \quad (6.5)$$

It is only necessary to investigate the case $n=0$ since the remaining part is identical with the already considered transformation (6.1). From

$$|\varphi'| = |\varphi| + 2\pi, \quad \frac{\varphi'}{|\varphi'|} = \frac{\varphi}{|\varphi|}, \quad (6.6)$$

follows

$$\cos\left(\frac{|\varphi'|}{2}\right) = -\cos\left(\frac{|\varphi|}{2}\right), \quad \sin\left(\frac{|\varphi'|}{2}\right) = -\sin\left(\frac{|\varphi|}{2}\right), \quad \Rightarrow \quad S' = -S. \quad (6.7)$$

Thus the transformation (6.5) is the transition from the matrices (5.13) to their negative matrices.

The inversion of the matrices is made by the transformation

$$\varphi \rightarrow \varphi' = -\varphi, \quad \Rightarrow \quad S' = S^{-1} = S^\dagger. \quad (6.8)$$

In applications the vector parameter φ does not possess the same weight (or Haar measure) independently on the modulus $|\varphi|$ and therefore not the same topology as a usual three-dimensional sphere with equal weight for all φ and with volume $V = \frac{4\pi}{3}R^3$ for radius R . As a second possible fundamental range of inequivalent parameters $|\varphi|$ one may also choose

$$2\pi \leq |\varphi| < 4\pi. \quad (6.9)$$

Therefore, the invariant measure for the described basic ranges of parameters $|\varphi|$ should be vanishing for all $|\varphi| = n2\pi, (n=0,1,2,\dots)$. We come back to this important problem in Section 16 when we discuss the derivation of an invariant

²We distinguish between a three-dimensional ball $\mathbb{B}$ with its two-dimensional surface S^2 called sphere. In two-dimensional case the analogous notions are a (round) disc $\mathbb{D}$ with its one-dimensional border, the circle S^1 .

measure over the group $SU(2)$. This invariant measure has to become vanishing for $|\varphi| = 2\pi$ and $|\varphi| = 4\pi$ such as for the center $|\varphi| = 0$.

7. Inversion of the Mapping $\varphi \rightleftharpoons \mathbf{S}(\varphi)$

From the matrix $\mathbf{S}$ one may determine the parameter $\varphi = (\varphi_1, \varphi_2, \varphi_3)$ up to the equivalence $\varphi \sim \varphi'$ given in (6.1). First from (5.15) follows

$$\begin{aligned} \frac{|\varphi|}{2} &= \arccos\left(\frac{\langle S \rangle}{2}\right) = \arccos\left(\frac{S_{11} + S_{22}}{2}\right) = \pm \arcsin\left(\sqrt{1 - \left(\frac{S_{11} + S_{22}}{2}\right)^2}\right), \\ \Rightarrow \frac{\frac{\varphi}{2}}{\sin\left(\frac{\varphi}{2}\right)} &= \frac{\arcsin\left(\sqrt{1 - \left(\frac{S_{11} + S_{22}}{2}\right)^2}\right)}{\sqrt{1 - \left(\frac{S_{11} + S_{22}}{2}\right)^2}}. \end{aligned} \quad (7.1)$$

Therefore

$$\begin{aligned} \varphi_1 &= -i \frac{\arcsin\left(\sqrt{1 - \left(\frac{S_{11} + S_{22}}{2}\right)^2}\right)}{\sqrt{1 - \left(\frac{S_{11} + S_{22}}{2}\right)^2}} (S_{12} + S_{21}), \\ \varphi_2 &= - \frac{\arcsin\left(\sqrt{1 - \left(\frac{S_{11} + S_{22}}{2}\right)^2}\right)}{\sqrt{1 - \left(\frac{S_{11} + S_{22}}{2}\right)^2}} (S_{12} - S_{21}), \\ \varphi_3 &= -i \frac{\arcsin\left(\sqrt{1 - \left(\frac{S_{11} + S_{22}}{2}\right)^2}\right)}{\sqrt{1 - \left(\frac{S_{11} + S_{22}}{2}\right)^2}} (S_{11} - S_{22}), \end{aligned} \quad (7.2)$$

or for $\varphi_- \equiv \varphi_1 - i\varphi_2$ and $\varphi_+ \equiv \varphi_1 + i\varphi_2$ instead of φ_1 and φ_2

$$\begin{aligned} \varphi_- &= -i2 \frac{\arcsin\left(\sqrt{1 - \left(\frac{S_{11} + S_{22}}{2}\right)^2}\right)}{\sqrt{1 - \left(\frac{S_{11} + S_{22}}{2}\right)^2}} S_{12}, \\ \varphi_+ &= -i2 \frac{\arcsin\left(\sqrt{1 - \left(\frac{S_{11} + S_{22}}{2}\right)^2}\right)}{\sqrt{1 - \left(\frac{S_{11} + S_{22}}{2}\right)^2}} S_{21}. \end{aligned} \quad (7.3)$$

Thus from the matrix (5.13) the vector parameter φ can be determined with the indeterminacy of φ described by (6.1).

8. Eigenvalues and Eigenvectors of $SU(2)$ in the Two-Dimensional Fundamental Representation with Vector Parameter φ

In this Section we determine the right-hand and left-hand eigenvectors (spinors) $\mathbf{x}$ and $\tilde{\mathbf{x}}$ of the matrix S in (5.13) to the well-known eigenvalues

$$\lambda = \exp\left(\pm i \frac{|\varphi|}{2}\right) \text{ according to}$$

$$\begin{aligned} S\mathbf{x}_1 &= \lambda_1 \mathbf{x}_1, & \tilde{\mathbf{x}}_1 S &= \lambda_1 \tilde{\mathbf{x}}_1, \\ S\mathbf{x}_2 &= \lambda_2 \mathbf{x}_2, & \tilde{\mathbf{x}}_2 S &= \lambda_2 \tilde{\mathbf{x}}_2, \end{aligned} \quad (8.1)$$

We consider $\mathbf{x}_1$ and $\mathbf{x}_2$ as column vectors and $\tilde{\mathbf{x}}_1$ and $\tilde{\mathbf{x}}_2$ as row vectors. From scalar multiplication of the first equation with $\tilde{\mathbf{x}}_2$ and of the third equation with $\mathbf{x}_1$ and forming the difference of the obtained equations follows for $\lambda_1 \neq \lambda_2$ the orthogonality of $\mathbf{x}_1$ and $\tilde{\mathbf{x}}_2$ and similarly of $\mathbf{x}_2$ and $\tilde{\mathbf{x}}_1$. This leads to the following representation of S

$$S = \lambda_1 \frac{\mathbf{x}_1 \cdot \tilde{\mathbf{x}}_1}{\tilde{\mathbf{x}}_1 \mathbf{x}_1} + \lambda_2 \frac{\mathbf{x}_2 \cdot \tilde{\mathbf{x}}_2}{\tilde{\mathbf{x}}_2 \mathbf{x}_2} \equiv \lambda_1 \Pi_1 + \lambda_2 \Pi_2, \quad \tilde{\mathbf{x}}_2 \mathbf{x}_1 = \tilde{\mathbf{x}}_1 \mathbf{x}_2 = 0. \quad (8.2)$$

The operators Π_1 and Π_2 are one-dimensional projection operators with the properties

$$\Pi_1^2 = \Pi_2^2 = I, \quad \Pi_1 \Pi_2 = \Pi_2 \Pi_1 = 0, \quad \langle \Pi_1 \rangle = \langle \Pi_2 \rangle = 1. \quad (8.3)$$

Since S are unitary unimodular matrices their eigenvalues λ are complex numbers on the unit circle in the complex plane. The unimodularity of the matrix S makes the product of its eigenvalues equal to 1 and thus complex conjugate in two-dimensional case. Concretely, one finds from the eigenvalue equation

$$\lambda^2 - \langle S \rangle \lambda + [S] = \lambda^2 - 2 \cos\left(\frac{|\varphi|}{2}\right) \lambda + 1 = 0, \quad (8.4)$$

the following well-known solutions for the eigenvalues

$$\lambda_1 = \exp\left(+i \frac{|\varphi|}{2}\right), \quad \lambda_2 = \exp\left(-i \frac{|\varphi|}{2}\right), \quad (8.5)$$

as already mentioned. Inserting these eigenvalues into (8.1) in the concrete representation (5.13) one obtains equations with the following solutions for the two-dimensional eigenvectors in a non-normalized form

$$\begin{aligned} \mathbf{x}_1 &= \begin{pmatrix} \varphi_- \\ |\varphi| - \varphi_3 \end{pmatrix}, & \tilde{\mathbf{x}}_1 &= (\varphi_+, |\varphi| - \varphi_3), & \tilde{\mathbf{x}}_1 \mathbf{x}_1 &= 2|\varphi|(|\varphi| - \varphi_3), & \tilde{\mathbf{x}}_1 \mathbf{x}_2 &= 0, \\ \mathbf{x}_2 &= \begin{pmatrix} -\varphi_- \\ |\varphi| + \varphi_3 \end{pmatrix}, & \tilde{\mathbf{x}}_2 &= (-\varphi_+, |\varphi| + \varphi_3), & \tilde{\mathbf{x}}_2 \mathbf{x}_2 &= 2|\varphi|(|\varphi| + \varphi_3), & \tilde{\mathbf{x}}_2 \mathbf{x}_1 &= 0. \end{aligned} \quad (8.6)$$

In general, they are complex vectors. There are possibilities to represent these eigenvector in another way and to choose other proportionality factors. As a more symmetrical and normalized form of the eigenvectors (8.6) one may

choose

$$\mathbf{u}_1 \equiv \frac{\mathbf{x}_1}{\sqrt{\tilde{\mathbf{x}}_1 \mathbf{x}_1}}, \quad \mathbf{u}_2 \equiv \frac{\mathbf{x}_2}{\sqrt{\tilde{\mathbf{x}}_2 \mathbf{x}_2}}, \quad \tilde{\mathbf{u}}_1 \equiv \frac{\tilde{\mathbf{x}}_1}{\sqrt{\tilde{\mathbf{x}}_1 \mathbf{x}_1}}, \quad \tilde{\mathbf{u}}_2 \equiv \frac{\tilde{\mathbf{x}}_2}{\sqrt{\tilde{\mathbf{x}}_2 \mathbf{x}_2}}, \quad (8.7)$$

which leads to

$$\begin{aligned} \mathbf{u}_1 &= \begin{pmatrix} \sqrt{\frac{1}{2} \sqrt{\frac{\varphi_-}{\varphi_+}} \left(1 + \frac{\varphi_3}{|\varphi|}\right)} \\ \sqrt{\frac{1}{2} \sqrt{\frac{\varphi_+}{\varphi_-}} \left(1 - \frac{\varphi_3}{|\varphi|}\right)} \end{pmatrix}, \quad \tilde{\mathbf{u}}_1 = \begin{pmatrix} \sqrt{\frac{1}{2} \sqrt{\frac{\varphi_+}{\varphi_-}} \left(1 + \frac{\varphi_3}{|\varphi|}\right)} \\ \sqrt{\frac{1}{2} \sqrt{\frac{\varphi_-}{\varphi_+}} \left(1 - \frac{\varphi_3}{|\varphi|}\right)} \end{pmatrix}, \\ \mathbf{u}_2 &= \begin{pmatrix} \sqrt{\frac{1}{2} \sqrt{\frac{\varphi_-}{\varphi_+}} \left(1 - \frac{\varphi_3}{|\varphi|}\right)} \\ -\sqrt{\frac{1}{2} \sqrt{\frac{\varphi_+}{\varphi_-}} \left(1 + \frac{\varphi_3}{|\varphi|}\right)} \end{pmatrix}, \quad \tilde{\mathbf{u}}_2 = \begin{pmatrix} \sqrt{\frac{1}{2} \sqrt{\frac{\varphi_+}{\varphi_-}} \left(1 - \frac{\varphi_3}{|\varphi|}\right)} \\ -\sqrt{\frac{1}{2} \sqrt{\frac{\varphi_-}{\varphi_+}} \left(1 + \frac{\varphi_3}{|\varphi|}\right)} \end{pmatrix}, \end{aligned} \quad (8.8)$$

with the orthonormality relations

$$\tilde{\mathbf{u}}_1 \mathbf{u}_1 = 1, \quad \tilde{\mathbf{u}}_2 \mathbf{u}_2 = 1, \quad \tilde{\mathbf{u}}_1 \mathbf{u}_2 = 0, \quad \tilde{\mathbf{u}}_2 \mathbf{u}_1 = 0. \quad (8.9)$$

In degenerate case $|\varphi| = 0$ one finds $\lambda_1 = \lambda_2 = 1$ and thus $S = I$ and in degenerate case $|\varphi| = 2\pi$ one has $\lambda_1 = \lambda_2 = -1$ corresponding to $S = -I$ that means only to one group transformation.

9. Composition Law for Vector Parameters φ

Corresponding to Products of $SU(2)$ Transformations

The group $SU(2)$ is described by the vector parameter φ , for example, in the fundamental representation by the matrix S given in (5.13) with respect to a basis discussed in Section 5. We now consider the composition of two such matrices S_1 and S_2 with the vector parameters φ_1 and φ_2 to the product matrix $S = S_1 S_2$ and ask how the vector parameter φ to the matrix S is connected with the vector parameters φ_1 and φ_2 according to the correspondences

$$S = S_1 S_2 \leftrightarrow \varphi, \quad S_1 \leftrightarrow \varphi_1, \quad S_2 \leftrightarrow \varphi_2. \quad (9.1)$$

From the multiplication of two such matrices of the form (5.13) and reorganization of the obtained terms one easily finds the following scalar and vector equation ($|\varphi_1| < 4\pi, |\varphi_2| < 4\pi, |\varphi| < 4\pi$)

$$\begin{aligned} \cos\left(\frac{|\varphi|}{2}\right) &= \cos\left(\frac{|\varphi_1|}{2}\right) \cos\left(\frac{|\varphi_2|}{2}\right) - \frac{\varphi_1 \varphi_2}{|\varphi_1| |\varphi_2|} \sin\left(\frac{|\varphi_1|}{2}\right) \sin\left(\frac{|\varphi_2|}{2}\right), \\ \frac{\varphi}{|\varphi|} \sin\left(\frac{|\varphi|}{2}\right) &= \frac{\varphi_1}{|\varphi_1|} \sin\left(\frac{|\varphi_1|}{2}\right) \cos\left(\frac{|\varphi_2|}{2}\right) + \frac{\varphi_2}{|\varphi_2|} \cos\left(\frac{|\varphi_1|}{2}\right) \sin\left(\frac{|\varphi_2|}{2}\right) \\ &\quad - \frac{[\varphi_1, \varphi_2]}{|\varphi_1| |\varphi_2|} \sin\left(\frac{|\varphi_1|}{2}\right) \sin\left(\frac{|\varphi_2|}{2}\right), \end{aligned} \quad (9.2)$$

where $[\varphi_1, \varphi_2]$ denotes the vector product of the vectors φ_1 and φ_2 . These two equations can be resolved with respect to the vector φ in unique way that provides the following formula (notation $\langle, \rangle$ see below)

$$\frac{\varphi}{2} = \frac{\frac{\varphi_1}{|\varphi_1|} \sin\left(\frac{|\varphi_1|}{2}\right) \cos\left(\frac{|\varphi_2|}{2}\right) + \frac{\varphi_2}{|\varphi_2|} \cos\left(\frac{|\varphi_1|}{2}\right) \sin\left(\frac{|\varphi_2|}{2}\right) - \frac{[\varphi_1, \varphi_2]}{|\varphi_1||\varphi_2|} \sin\left(\frac{|\varphi_1|}{2}\right) \sin\left(\frac{|\varphi_2|}{2}\right)}{\left| \frac{\varphi_1}{|\varphi_1|} \sin\left(\frac{|\varphi_1|}{2}\right) \cos\left(\frac{|\varphi_2|}{2}\right) + \frac{\varphi_2}{|\varphi_2|} \cos\left(\frac{|\varphi_1|}{2}\right) \sin\left(\frac{|\varphi_2|}{2}\right) - \frac{[\varphi_1, \varphi_2]}{|\varphi_1||\varphi_2|} \sin\left(\frac{|\varphi_1|}{2}\right) \sin\left(\frac{|\varphi_2|}{2}\right) \right|} \quad (9.3)$$

$$\cdot \arccos\left(\cos\left(\frac{|\varphi_1|}{2}\right) \cos\left(\frac{|\varphi_2|}{2}\right) - \frac{\varphi_1 \varphi_2}{|\varphi_1||\varphi_2|} \sin\left(\frac{|\varphi_1|}{2}\right) \sin\left(\frac{|\varphi_2|}{2}\right)\right) \equiv \left\langle \frac{\varphi_1}{2}, \frac{\varphi_2}{2} \right\rangle.$$

One may call this formula the composition law for the chosen vector parameter of the group $SU(2)$. The factor on first line of the right-hand gives the direction of the new vector φ which as we have seen in Section 4 possesses another meaning for $SU(2)$ in comparison to the rotation group and the factor on the second line the modulus of $\frac{\varphi}{2}$. For the new parameter $\frac{\varphi}{|\varphi|}$ we find

$$\frac{\varphi}{|\varphi|} = \frac{\frac{\varphi_1}{|\varphi_1|} \sin\left(\frac{|\varphi_1|}{2}\right) \cos\left(\frac{|\varphi_2|}{2}\right) + \frac{\varphi_2}{|\varphi_2|} \cos\left(\frac{|\varphi_1|}{2}\right) \sin\left(\frac{|\varphi_2|}{2}\right) - \frac{[\varphi_1, \varphi_2]}{|\varphi_1||\varphi_2|} \sin\left(\frac{|\varphi_1|}{2}\right) \sin\left(\frac{|\varphi_2|}{2}\right)}{\sqrt{1 - \left(\cos\left(\frac{|\varphi_1|}{2}\right) \cos\left(\frac{|\varphi_2|}{2}\right) - \frac{\varphi_1 \varphi_2}{|\varphi_1||\varphi_2|} \sin\left(\frac{|\varphi_1|}{2}\right) \sin\left(\frac{|\varphi_2|}{2}\right)\right)^2}}. \quad (9.4)$$

The above composition formulae are somehow similar (but not identical) to formulae for spherical trigonometry of the surface (sphere) of a three-dimensional ball but here play a role also the inner points. In addition, in applications our three-dimensional ball possesses different weights of its points in dependence on the modulus $|\varphi|$ and thus another topology as a usual three-dimensional ball with equal weight measure for all its inner points. For example, as discussed the whole surface (sphere) of our three-dimensional ball with $|\varphi|$ corresponds to the operator $-I$ that means to only one point and therefore the (Haar) measure has to vanish for $|\varphi|$. In Section 14-15, we consider another parametrization where the composition law for $SO(3, \mathbb{R})$ takes on a simpler form.

We mention that Fyodorov [5] (§. 29, p. 447) and [6] (§. 3, p. 18) introduced a special notation $\langle, \rangle$ for the composition of parameters corresponding to the product of two group elements, in our case of φ_1 and φ_2 to φ according to

$$\varphi = \langle \varphi_1, \varphi_2 \rangle, \quad \langle \varphi_1, \varphi_2 \rangle \neq \langle \varphi_2, \varphi_1 \rangle, \quad (\text{in general}). \quad (9.5)$$

For the composition of three parameters holds an associative law

$$\langle \langle \varphi_1, \varphi_2 \rangle, \varphi_3 \rangle = \langle \varphi_1, \langle \varphi_2, \varphi_3 \rangle \rangle \equiv \langle \varphi_1, \varphi_2, \varphi_3 \rangle, \quad (9.6)$$

corresponding to the associative law for the multiplication of group elements. Therefore, one may omit the inner brackets. The introduced symbols are very convenient and hardly come into conflict with other generally used symbols (only Hermitean scalar products are sometimes denoted by such brackets).

10. Decompositions of $S(\varphi)$ Matrices or Disentanglement Relations for the Group $SU(2)$

Beside the composition it is sometimes useful to decompose the general matrix $S(\varphi)$ in (5.13) into products of simpler matrices which is called disentanglement. The obtained decompositions are then true for arbitrary irreducible representations. With the two-dimensional fundamental representation of $SU(2)$ in representation by the vector parameter φ in Section 5 we have developed at once the mathematical means for the disentanglement of $SU(2)$ group operators that we present here. The method is the same as used, for example, in [27] [28] for $SU(1,1)$. We have to decompose the general matrices S of the fundamental representation derived in (5.13) into products of simpler matrices and have to look for the corresponding decompositions of the general group operators $\exp(i\mathbf{J}\varphi)$ into products of special group operators. Another method is the derivation of differential equations for the exponents in the decomposition formulae by introduction of an additional parameter and differentiation with respect to this parameter and then to solve the obtained differential equations that is made in a paper of Ban [29] based on Lie algebra methods (see also, e.g., [30] [31]).

The following more special matrices of (5.13) are mainly of interest

$$\begin{aligned}\exp\left(\frac{i}{2}J_+\varphi_-\right) &\rightarrow S(\varphi_-, 0, 0) = \begin{pmatrix} 1 & i\frac{\varphi_-}{2} \\ 0 & 1 \end{pmatrix}, \\ \exp\left(\frac{i}{2}J_-\varphi_+\right) &\rightarrow S(0, \varphi_+, 0) = \begin{pmatrix} 1 & 0 \\ i\frac{\varphi_+}{2} & 1 \end{pmatrix}, \\ \exp(iJ_3\varphi_3) &\rightarrow S(0, 0, \varphi_3) = \begin{pmatrix} \exp\left(i\frac{\varphi_3}{2}\right) & 0 \\ 0 & \exp\left(-i\frac{\varphi_3}{2}\right) \end{pmatrix}.\end{aligned}\tag{10.1}$$

The first two are oppositely triangular matrices and the third is a diagonal matrix.

If we take into account the most interesting decompositions of the general operator of $SU(2)$ into products of special operators with J_+ , J_- and J_3 separately in the arguments of the exponentials, we can make the following 6 product decompositions of the 2D matrix S using its unimodularity

$$\begin{aligned}S_{11}S_{22} - S_{12}S_{21} &= 1 \\ \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} &= \begin{pmatrix} 1 & 0 \\ \frac{S_{21}}{S_{11}} & 1 \end{pmatrix} \begin{pmatrix} 1 & S_{12}S_{11} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} S_{11} & 0 \\ 0 & \frac{1}{S_{11}} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \frac{S_{21}}{S_{11}} & 1 \end{pmatrix} \begin{pmatrix} S_{11} & 0 \\ 0 & \frac{1}{S_{11}} \end{pmatrix} \begin{pmatrix} 1 & \frac{S_{12}}{S_{11}} \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} S_{11} & 0 \\ 0 & \frac{1}{S_{11}} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ S_{21}S_{11} & 1 \end{pmatrix} \begin{pmatrix} 1 & \frac{S_{12}}{S_{11}} \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} \frac{1}{S_{22}} & 0 \\ 0 & S_{22} \end{pmatrix} \begin{pmatrix} 1 & S_{12}S_{22} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ \frac{S_{21}}{S_{22}} & 1 \end{pmatrix}\end{aligned}$$

$$= \begin{pmatrix} 1 & \frac{S_{12}}{S_{22}} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{S_{22}} & 0 \\ 0 & S_{22} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ \frac{S_{21}}{S_{22}} & 1 \end{pmatrix} = \begin{pmatrix} 1 & \frac{S_{12}}{S_{22}} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ S_{21}S_{22} & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{S_{22}} & 0 \\ 0 & S_{22} \end{pmatrix}. \quad (10.2)$$

Evidently, these decompositions are not specific for $SU(2)$ and are applicable for all two-dimensional unimodular matrices of the Special linear group $SL(2, \mathbb{C})$, for example, also similar for the two-dimensional non-unitary fundamental representation of $SU(1,1)$ and decompositions of matrices including triangular matrices were already known to Gauss [32]. The obtained disentanglement relations in the corresponding 6 considered orderings of group operators of $SU(2)$ are

$$\begin{aligned} & \exp \left\{ i \left(J_+ \frac{\varphi_-}{2} + J_- \frac{\varphi_+}{2} + J_3 \varphi_3 \right) \right\} \\ &= \exp \left(J_- \frac{S_{21}}{S_{11}} \right) \exp(J_+ S_{12} S_{11}) \exp(2J_3 \log(S_{11})) \\ &= \exp \left(J_- \frac{S_{21}}{S_{11}} \right) \exp(2J_3 \log(S_{11})) \exp \left(J_+ \frac{S_{12}}{S_{11}} \right) \\ &= \exp(2J_3 \log(S_{11})) \exp(J_- S_{21} S_{11}) \exp \left(J_+ \frac{S_{12}}{S_{11}} \right) \\ &= \exp(-2J_3 \log(S_{22})) \exp(J_+ S_{12} S_{22}) \exp \left(J_- \frac{S_{21}}{S_{22}} \right) \\ &= \exp \left(J_+ \frac{S_{12}}{S_{22}} \right) \exp(-2J_3 \log(S_{22})) \exp \left(J_- \frac{S_{21}}{S_{22}} \right) \\ &= \exp \left(J_+ \frac{S_{12}}{S_{22}} \right) \exp(J_- S_{21} S_{22}) \exp(-2J_3 \log(S_{22})), \end{aligned} \quad (10.3)$$

with the explicit form of the elements of the two-dimensional matrix S (see (5.13))

$$\begin{aligned} S_{11} &= \cos \left(\frac{|\varphi|}{2} \right) + i \frac{\varphi_3}{|\varphi|} \sin \left(\frac{|\varphi|}{2} \right), \quad S_{12} = i \frac{\varphi_-}{|\varphi|} \sin \left(\frac{|\varphi|}{2} \right), \\ S_{21} &= i \frac{\varphi_+}{|\varphi|} \sin \left(\frac{|\varphi|}{2} \right), \quad S_{22} = \cos \left(\frac{|\varphi|}{2} \right) - i \frac{\varphi_3}{|\varphi|} \sin \left(\frac{|\varphi|}{2} \right), \\ \varphi_{\pm} &\equiv \varphi_1 \pm i \varphi_2, \quad |\varphi| \equiv \sqrt{\varphi_+ \varphi_- + (\varphi_3)^2}. \end{aligned} \quad (10.4)$$

In the special case $\varphi_3 = 0$ corresponding to

$$\exp \left(\frac{i}{2} (J_+ \varphi_- + J_- \varphi_+) \right) \rightarrow S(\varphi_-, \varphi_+, 0) = \begin{pmatrix} \cos \left(\frac{|\varphi|}{2} \right) & i \frac{\varphi_-}{|\varphi|} \sin \left(\frac{|\varphi|}{2} \right) \\ i \frac{\varphi_+}{|\varphi|} \sin \left(\frac{|\varphi|}{2} \right) & \cos \left(\frac{|\varphi|}{2} \right) \end{pmatrix}, \quad (10.5)$$

$$|\varphi| \equiv \sqrt{\varphi_+ \varphi_-} = \sqrt{\varphi_1^2 + \varphi_2^2},$$

we obtain from (10.3) and (10.4) the disentanglement relations

$$\begin{aligned}
& \exp\left\{\frac{i}{2}(J_+\varphi_- + J_-\varphi_+)\right\} \\
&= \exp\left(iJ_-\frac{\varphi_+}{|\varphi|}\operatorname{tg}\left(\frac{|\varphi|}{2}\right)\right)\exp\left(iJ_+\frac{\varphi_-\sin(|\varphi|)}{2|\varphi|}\right)\left(\cos\left(\frac{|\varphi|}{2}\right)\right)^{2J_3} \\
&= \exp\left(iJ_-\frac{\varphi_+}{|\varphi|}\operatorname{tg}\left(\frac{|\varphi|}{2}\right)\right)\left(\cos\left(\frac{|\varphi|}{2}\right)\right)^{2J_3}\exp\left(iJ_+\frac{\varphi_-}{|\varphi|}\operatorname{tg}\left(\frac{|\varphi|}{2}\right)\right) \\
&= \left(\cos\left(\frac{|\varphi|}{2}\right)\right)^{2J_3}\exp\left(iJ_-\frac{\varphi_+\sin(|\varphi|)}{2|\varphi|}\right)\exp\left(iJ_+\frac{\varphi_-}{|\varphi|}\operatorname{tg}\left(\frac{|\varphi|}{2}\right)\right) \\
&= \left(\cos\left(\frac{|\varphi|}{2}\right)\right)^{-2J_3}\exp\left(iJ_+\frac{\varphi_-\sin(|\varphi|)}{2|\varphi|}\right)\exp\left(iJ_-\frac{\varphi_+}{|\varphi|}\operatorname{tg}\left(\frac{|\varphi|}{2}\right)\right) \\
&= \exp\left(iJ_+\frac{\varphi_-}{|\varphi|}\operatorname{tg}\left(\frac{|\varphi|}{2}\right)\right)\left(\cos\left(\frac{|\varphi|}{2}\right)\right)^{-2J_3}\exp\left(iJ_-\frac{\varphi_+}{|\varphi|}\operatorname{tg}\left(\frac{|\varphi|}{2}\right)\right) \\
&= \exp\left(iJ_+\frac{\varphi_-}{|\varphi|}\operatorname{tg}\left(\frac{|\varphi|}{2}\right)\right)\exp\left(iJ_-\frac{\varphi_+\sin(|\varphi|)}{2|\varphi|}\right)\left(\cos\left(\frac{|\varphi|}{2}\right)\right)^{-2J_3}, \quad |\varphi| \equiv \sqrt{\varphi_+\varphi_-},
\end{aligned} \tag{10.6}$$

where on the right-hand side the operator J_3 appears although it is not on the left-hand side. The matrices $S(\varphi_-, \varphi_+, 0)$ in (10.5) themselves do not form a group.

These formulae are important, for example, for the derivation of $SU(2)$ group-coherent states in the sense of Perelomov [32] and for their representation.

We consider now the following decompositions of the unimodular 2D matrix S into products of two unimodular matrices

$$\begin{aligned}
\begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} &= \begin{pmatrix} \sqrt{S_{11}S_{22}} & \sqrt{\frac{S_{11}}{S_{22}}}S_{12} \\ \sqrt{\frac{S_{22}}{S_{11}}}S_{21} & \sqrt{S_{11}S_{22}} \end{pmatrix} \begin{pmatrix} \sqrt{\frac{S_{11}}{S_{22}}} & 0 \\ 0 & \sqrt{\frac{S_{22}}{S_{11}}} \end{pmatrix} \\
&= \begin{pmatrix} \sqrt{\frac{S_{11}}{S_{22}}} & 0 \\ 0 & \sqrt{\frac{S_{22}}{S_{11}}} \end{pmatrix} \begin{pmatrix} \sqrt{S_{11}S_{22}} & \sqrt{\frac{S_{22}}{S_{11}}}S_{12} \\ \sqrt{\frac{S_{11}}{S_{22}}}S_{21} & \sqrt{S_{11}S_{22}} \end{pmatrix}.
\end{aligned} \tag{10.7}$$

They correspond to the following disentanglement of group operators with J_3 and with $J_\pm$

$$\begin{aligned}
\exp\left\{i\left(J_+\frac{\varphi_-}{2} + J_-\frac{\varphi_+}{2} + J_3\varphi_3\right)\right\} &= \exp\left\{i\left(J_+\frac{\varphi'_-}{2} + J_-\frac{\varphi'_+}{2}\right)\right\}\exp(iJ_3\varphi'_3) \\
&= \exp(iJ_3\varphi'_3)\exp\left\{i\left(J_+\frac{\varphi''_-}{2} + J_-\frac{\varphi''_+}{2}\right)\right\},
\end{aligned} \tag{10.8}$$

with the relations

$$\begin{aligned}
\frac{\varphi'_\pm}{2} &= \exp\left(\mp i \frac{\varphi'_3}{2}\right) \frac{\varphi_\pm}{\sqrt{\varphi_+ \varphi_-}} \arcsin\left(\frac{\sqrt{\varphi_+ \varphi_-}}{|\varphi|} \sin\left(\frac{|\varphi|}{2}\right)\right), \quad \exp\left(\pm i \frac{\varphi'_3}{2}\right) \frac{\varphi'_\pm}{\sqrt{\varphi'_+ \varphi'_-}} = \frac{\varphi_\pm}{\sqrt{\varphi_+ \varphi_-}}, \\
\frac{\varphi''_\pm}{2} &= \exp\left(\pm i \frac{\varphi'_3}{2}\right) \frac{\varphi_\pm}{\sqrt{\varphi_+ \varphi_-}} \arcsin\left(\frac{\sqrt{\varphi_+ \varphi_-}}{|\varphi|} \sin\left(\frac{|\varphi|}{2}\right)\right), \quad \exp\left(\mp i \frac{\varphi'_3}{2}\right) \frac{\varphi''_\pm}{\sqrt{\varphi''_+ \varphi''_-}} = \frac{\varphi_\pm}{\sqrt{\varphi_+ \varphi_-}}, \quad (10.9) \\
\exp\left(i \frac{\varphi'_3}{2}\right) &= \sqrt{\frac{\cos\left(\frac{|\varphi|}{2}\right) + i \frac{\varphi_3}{|\varphi|} \sin\left(\frac{|\varphi|}{2}\right)}{\cos\left(\frac{|\varphi|}{2}\right) - i \frac{\varphi_3}{|\varphi|} \sin\left(\frac{|\varphi|}{2}\right)}}, \quad \operatorname{tg}\left(\frac{\varphi'_3}{2}\right) = \frac{\varphi_3}{|\varphi|} \operatorname{tg}\left(\frac{|\varphi|}{2}\right).
\end{aligned}$$

From these relations follows

$$\begin{aligned}
\cos\left(\frac{\sqrt{\varphi'_+ \varphi'_-}}{2}\right) &= \cos\left(\frac{\sqrt{\varphi''_+ \varphi''_-}}{2}\right) = \sqrt{\cos^2\left(\frac{|\varphi|}{2}\right) + \frac{(\varphi_3)^2}{|\varphi|^2} \sin^2\left(\frac{|\varphi|}{2}\right)}, \\
\sin\left(\frac{\sqrt{\varphi'_+ \varphi'_-}}{2}\right) &= \sin\left(\frac{\sqrt{\varphi''_+ \varphi''_-}}{2}\right) = \frac{\sqrt{\varphi_+ \varphi_-}}{|\varphi|} \sin\left(\frac{|\varphi|}{2}\right). \quad (10.10)
\end{aligned}$$

The stable part in the decompositions (10.8) is the factor $\exp(iJ_3\varphi'_3)$. The inversion of (10.9) and (10.10) can be found using

$$\cos\left(\frac{|\varphi|}{2}\right) = \cos\left(\frac{\sqrt{\varphi'_+ \varphi'_-}}{2}\right) \cos\left(\frac{\varphi'_3}{2}\right) = \cos\left(\frac{\sqrt{\varphi''_+ \varphi''_-}}{2}\right) \cos\left(\frac{\varphi'_3}{2}\right), \quad (10.11)$$

which follows from combination of the relation for $\cos\left(\frac{\varphi'_3}{2}\right)$ in (10.9) with the relation for $\cos\left(\frac{\sqrt{\varphi'_+ \varphi'_-}}{2}\right)$ in (10.10).

11. Parametrization of $SU(2)$ by Quaternions

For some completeness we will shortly consider the parametrization of $SU(2)$ by quaternions which were introduced by W.R. Hamilton in the middle of the 19th century after searching for more general number systems than complex numbers (e.g. [33] [34] [35]).

A quaternion $r = (r_0, \mathbf{r})$ consists of a scalar part r_0 and of a vectorial part $\mathbf{r} \equiv (r_1, r_2, r_3)$ which both (by definition) are real in case of real quaternions. The associative but not commutative multiplication law in the quaternion algebra $\mathbb{H}$ for two quaternions $r = (r_0, \mathbf{r})$ and $s = (s_0, \mathbf{s})$ is

$$\begin{aligned}
rs &\equiv (r_0, \mathbf{r})(s_0, \mathbf{s}) = (r_0 s_0 - \mathbf{r} \cdot \mathbf{s}, s_0 \mathbf{r} + r_0 \mathbf{s} + [\mathbf{r}, \mathbf{s}]), \\
sr &\equiv (s_0, \mathbf{s})(r_0, \mathbf{r}) = (r_0 s_0 - \mathbf{r} \cdot \mathbf{s}, s_0 \mathbf{r} + r_0 \mathbf{s} - [\mathbf{r}, \mathbf{s}]), \quad (11.1)
\end{aligned}$$

where $\mathbf{r} \cdot \mathbf{s}$ denotes the scalar product and $[\mathbf{r}, \mathbf{s}]$ the vector product of two vectors $\mathbf{r}$ and $\mathbf{s}$. From (11.1) follows

$$\frac{rs + sr}{2} = (r_0 s_0 - \mathbf{r} \cdot \mathbf{s}, s_0 \mathbf{r} + r_0 \mathbf{s}), \quad \frac{rs - sr}{2} = (0, [\mathbf{r}, \mathbf{s}]), \quad (11.2)$$

The quaternion $\bar{r} = \overline{(r_0, \mathbf{r})} \equiv (r_0, -\mathbf{r})$ is called the conjugate quaternion to $r = (r_0, \mathbf{r})$ and the product $r\bar{r} = \bar{r}r = (r_0^2 + \mathbf{r}^2, \mathbf{0})$ is proportional to the identical quaternion $(1, \mathbf{0})$ and therefore $r^{-1} \equiv \frac{\bar{r}}{r_0^2 + \mathbf{r}^2}$ is the reciprocal quaternion to r . The nonnegative number $|r| \equiv \sqrt{r_0^2 + \mathbf{r}^2}$ is the modulus or norm of a quaternion. The multiplication law (11.1) can be realized by matrix multiplication, in lowest-dimensional case by special 2D matrices, for example, by the following correspondence

$$(r_0, \mathbf{r}) \equiv (r_0, (r_1, r_2, r_3)) \leftrightarrow \begin{pmatrix} r_0 - ir_3 & i(r_1 - ir_2) \\ i(r_1 + ir_2) & r_0 + ir_3 \end{pmatrix} \equiv S, \quad (11.3)$$

with the additive decomposition of S as follows³

$$\begin{aligned} S &= r_0 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + ir_1 \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} + ir_2 \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix} - ir_3 \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \\ &= r_0 I + ir_1 \sigma_1 + ir_2 \sigma_2 - ir_3 \sigma_3, \end{aligned} \quad (11.4)$$

where $(\sigma_1, \sigma_2, \sigma_3)$ are the three Pauli spin matrices σ_k explicitly given in (5.3). By comparison with (5.13), one finds the following correspondences between $SU(2)$ matrices S and real quaternions $(r_0, \mathbf{r})$

$$S \leftrightarrow (r_0, (r_1, r_2, r_3)) = \left(\cos\left(\frac{|\boldsymbol{\varphi}|}{2}\right), \frac{(\varphi_1, \varphi_2, -\varphi_3)}{|\boldsymbol{\varphi}|} \sin\left(\frac{|\boldsymbol{\varphi}|}{2}\right) \right). \quad (11.5)$$

This means that the squared modulus of the quaternion is the determinant $[S]$ of the two-dimensional matrix S which due to unimodularity is equal to 1 and the scalar part is half the trace $\langle S \rangle$ of the matrix S . Therefore, $SU(2)$ matrices correspond to real unit quaternions (modulus equal to 1 by definition) with 3 independent real parameters. The multiplication of matrices (5.13) or the quaternion multiplication (11.1) allow to establish the composition law of two $SU(2)$ transformations. Clearly, the noncommutative matrix multiplication is the more generally applicable operation in comparison to quaternion multiplication.

12. Regular Representation $D^{(1)}$ of $SU(2)$ as Basic Representation of $SO(3, \mathbb{R})$

In this Section we construct the regular (or adjoint) representation of $SU(2)$ which provides the group of inner automorphisms of $SU(2)$ and thus the transformation of the vector operator $\mathbf{J}$. It uses the operators of the abstract Lie algebra themselves as a basis and, therefore, is three-dimensional. If we use (J_+, J_3, J_-) as basis, we can construct the three-dimensional representation matrices to J_k from the commutators $[J_k, (J_+, J_3, J_-)]$ in analogy to (5.3) that

³This formula shows a blemish of beauty, the negative sign in front of σ_3 compared with the positive signs in front of σ_1 and σ_2 . We did not find a way to remove it and, likely, this is impossible. If we change, for example, $r_3 \rightarrow -r_3$ in (11.3) then the matrix multiplication is no more compatible with the definition (11.1) of the quaternion product.

we do not write down explicitly. From this realization we find in analogy to (5.12) the regular representation of $SU(2)$ in the form

$$(J'_+, J'_3, J'_-) \equiv \exp(i\mathbf{J}\boldsymbol{\varphi})(J_+, J_3, J_-)\exp(-i\mathbf{J}\boldsymbol{\varphi}) = (J_+, J_3, J_-)\mathbf{R}, \quad (12.1)$$

with the following three-dimensional matrix $\mathbf{R} = \mathbf{R}(\boldsymbol{\varphi})$ with particularly simple structure

$$\mathbf{R} = \begin{pmatrix} S_{11}S_{11} & -S_{11}S_{12} & -S_{12}S_{12} \\ -2S_{11}S_{21} & S_{11}S_{22} + S_{12}S_{21} & 2S_{12}S_{22} \\ -S_{21}S_{21} & S_{21}S_{22} & S_{22}S_{22} \end{pmatrix}, \quad (12.2)$$

$$\langle \mathbf{R} \rangle = 1 + 2 \cos(|\boldsymbol{\varphi}|) = \frac{\sin\left(3\frac{|\boldsymbol{\varphi}|}{2}\right)}{\sin\left(\frac{|\boldsymbol{\varphi}|}{2}\right)}, \quad [\mathbf{R}] = \langle \mathbf{R} \rangle, \quad |\mathbf{R}| = 1,$$

where $\langle \mathbf{R} \rangle$ denotes the trace, $[\mathbf{R}]$ the second invariant and $|\mathbf{R}|$ the determinant of the operator $\mathbf{R}$ (see also below) and where we took into account $S_{11}S_{22} - S_{12}S_{21} = 1$. It is a special complex unimodular matrix which is equivalent to a real orthogonal matrix as we will now show. Inserting (5.13) for S , we obtain an explicit form of this matrix which we do not write down. It is now straightforward to get the matrix $\mathbf{R}$ of the mapping $\exp(i\mathbf{J}\boldsymbol{\varphi}) \rightarrow \mathbf{R}$ in the basis of the operators (J_1, J_2, J_3) instead of (J_+, J_3, J_-) that can be represented in the vector form⁴

$$\mathbf{J}' \equiv \exp(i\mathbf{J}\boldsymbol{\varphi})\mathbf{J}\exp(-i\mathbf{J}\boldsymbol{\varphi}) = \mathbf{J}\mathbf{R}, \quad \Leftrightarrow \quad J'_l = J_k R_{kl}, \quad (12.3)$$

and explicitly in representation by vector components (sum convention)

$$R_{kl} \equiv \frac{\varphi_k \varphi_l}{|\boldsymbol{\varphi}|^2} + \left(\delta_{kl} - \frac{\varphi_k \varphi_l}{|\boldsymbol{\varphi}|^2} \right) \cos(|\boldsymbol{\varphi}|) + \varepsilon_{kij} \frac{\varphi_j}{|\boldsymbol{\varphi}|^2} \sin(|\boldsymbol{\varphi}|), \quad (12.4)$$

or written in coordinate-invariant form

$$\mathbf{R} = \frac{\boldsymbol{\varphi} \cdot \boldsymbol{\varphi}}{|\boldsymbol{\varphi}|^2} + \left(\mathbf{I} - \frac{\boldsymbol{\varphi} \cdot \boldsymbol{\varphi}}{|\boldsymbol{\varphi}|^2} \right) \cos(|\boldsymbol{\varphi}|) - \left[\frac{\boldsymbol{\varphi}}{|\boldsymbol{\varphi}|} \right] \sin(|\boldsymbol{\varphi}|), \quad (12.5)$$

$$|\boldsymbol{\varphi}| \equiv \sqrt{\varphi_1^2 + \varphi_2^2 + \varphi_3^2} \leq 2\pi.$$

The matrix $\mathbf{R} = \mathbf{R}(\boldsymbol{\varphi})$ in the representation R_{kl} corresponding to the basis (J_1, J_2, J_3) is a three-dimensional real orthogonal matrix with determinant equal to +1 (*i.e.* a proper rotation) that means

$$\langle \mathbf{R} \rangle = [\mathbf{R}] = 1 + 2 \cos(|\boldsymbol{\varphi}|), \quad |\mathbf{R}| = 1, \quad \mathbf{R}\mathbf{R}^T = \mathbf{I}, \quad (R_{kj}R_{jl} = \delta_{kl}). \quad (12.6)$$

The rotation axis is described by the unit axial vector $\frac{\boldsymbol{\varphi}}{|\boldsymbol{\varphi}|}$ and the rotation angle is $|\boldsymbol{\varphi}|$ or, equivalently, by the unit vector $-\frac{\boldsymbol{\varphi}}{|\boldsymbol{\varphi}|}$ and the rotation angle $-|\boldsymbol{\varphi}|$.

⁴We do not introduce a new notation for $\mathbf{R}$ in comparison to (12.2) because we consider it as the same operator in matrix representation with respect to different basis vectors.

In the form $J_k R_{kl}$ it describes by convention an anti-clockwise rotation of the vector J_k and in the form $R_{kl} r_l$ a clockwise rotation of the vector r_l about an angle $\varphi \equiv |\varphi|$ if the vector $\frac{\varphi}{|\varphi|}$ is perpendicular to the clock plane.

Rotations with the parameters φ and $\varphi + n2\pi \frac{\varphi}{|\varphi|}$, ($n = \pm 1, \pm 2, \dots$) lead to the same matrix R . We consider the following transformations of the vector parameter φ

$$\varphi \rightarrow \varphi' = \varphi + n2\pi \frac{\varphi}{|\varphi|} = (|\varphi| + n2\pi) \frac{\varphi}{|\varphi|} \sim \varphi, \quad (n = 0, \pm 1, \pm 2, \dots), \quad (12.7)$$

from which follows for modulus and direction of φ'

$$|\varphi'| = |\varphi| + n2\pi = \pm(|\varphi| + n2\pi), \quad \Leftrightarrow \quad \frac{\varphi'}{|\varphi'|} = \frac{|\varphi| + n2\pi}{|\varphi| + n2\pi} \frac{\varphi}{|\varphi|} = \pm \frac{\varphi}{|\varphi|}. \quad (12.8)$$

Inserted in (12.5) it leaves the rotation R unchanged.

Two other special cases are, in particular, the special case $n = -1$ in (6.1) for which follows

$$\varphi \rightarrow \varphi' = \varphi - 2\pi \frac{\varphi}{|\varphi|}, \quad \Rightarrow \quad |\varphi| \rightarrow 2\pi - |\varphi|, \quad \frac{\varphi}{|\varphi|} \rightarrow -\frac{\varphi}{|\varphi|}, \quad R \rightarrow R. \quad (12.9)$$

In comparison, for the transformation $\varphi \rightarrow \varphi - \pi \frac{\varphi}{|\varphi|}$ we find

$$\varphi \rightarrow \varphi' = \varphi - \pi \frac{\varphi}{|\varphi|}, \quad \Rightarrow \quad |\varphi| \rightarrow \pm(\pi - |\varphi|), \quad \frac{\varphi}{|\varphi|} \rightarrow \pm \frac{\varphi}{|\varphi|}, \quad R \rightarrow R^{-1}. \quad (12.10)$$

This allows to restrict a fundamental region of the parameters $\varphi = (\varphi_1, \varphi_2, \varphi_3)$ to the three-dimensional sphere $|\varphi| \leq \pi$ with identification of opposite points on the surface of the two-dimensional sphere $\mathbb{S}^2$ with $|\varphi| = \pi$ as its boundary. The very direct relation to covariant quantities of $SO(3, \mathbb{R})$ is an advantage of using the vector parameter $\varphi = (\varphi_1, \varphi_2, \varphi_3)$ in comparison to the Euler angles (α, β, γ) .

We can look to relation (12.3) also in another way. For this purpose we take the vector parameter φ_0 for an arbitrary element $x_0 \equiv J\varphi_0$ of the Lie algebra to $SU(2)$ and consider its transformation

$$J'\varphi_0 \equiv \exp(iJ\varphi)(J\varphi_0)\exp(-iJ\varphi) = JR\varphi_0 \equiv J\varphi'_0, \quad (12.11)$$

from which follows

$$\varphi'_0 = R\varphi_0, \quad (0 \leq |\varphi_0| < 2\pi). \quad (12.12)$$

This means that all elements of $SU(2)$ with vector parameters φ'_0 where $\varphi'_0 = R\varphi_0$ is obtained from φ_0 by an arbitrary rotation of the three-dimensional rotation group $SO(3, \mathbb{R})$ are equivalent and form one class within $SU(2)$. This transformation changes only the axis direction $\mathbf{n}_0 \equiv \frac{\varphi_0}{|\varphi_0|}$ and all elements of

$SU(2)$ with the same $|\varphi_0|$ are equivalent.

The relation between the three-dimensional matrix R_{kl} and the two-dimensional matrix S and its inversion in covariant form is (e.g., see [18] (Equations (2.16) and (2.32)), [25] (pp. 42-43) and [36] (Equation (2.2.15)))⁵

$$R_{kl} = \frac{1}{2} \langle \sigma_k S \sigma_l S^\dagger \rangle, \quad S = \pm \frac{I + \sigma_k R_{kl} \sigma_l}{\sqrt{[I + \sigma_k R_{kl} \sigma_l]}}, \quad (12.13)$$

where $\langle \dots \rangle$ and $[\dots]$ denote here the trace and determinant of two-dimensional matrices, respectively. Matrices $+S$ and $-S$ lead to the same R . Thus we have constructed the known 2-1 homomorphism of $SU(2)$ to

$SO(3, \mathbb{R}) = SU(2)/Z_2$ where Z_2 is the center of $SU(2)$ consisting of two elements I and $-I$. We can look to Equations (1)-(6) as to the inner automorphisms of the unitary unimodular group $SU(2)$ that means to the inner transformations of the operators (J_1, J_2, J_3) which leave unchanged the commutation relations. This is important for the coordinate-invariant interpretation.

13. Parametrization of Rotation Group $SO(3, \mathbb{R})$ by Vector φ in Coordinate-Invariant Description

The three-dimensional rotation operator R can be represented in the following exponential form

$$R \equiv e^\Phi, \quad \Rightarrow \quad e^{-\Phi} = R^{-1} = R^\top = e^{\Phi^\top}, \quad \Rightarrow \quad \Phi = -\Phi^\top, \quad (13.1)$$

which defines a real three-dimensional antisymmetric operator Φ . Three-dimensional anti-symmetric operators $\Phi = -\Phi^\top$ can be mapped onto three-dimensional axial vectors φ according to

$$\Phi \leftrightarrow \Phi_{kl} = \varepsilon_{kjl} \varphi_j = -\Phi_{lk}, \quad \varphi \leftrightarrow \varphi_j = \frac{1}{2} \varepsilon_{kjl} \Phi_{kl}, \quad (13.2)$$

from which follows

$$\langle \Phi \rangle = 0, \quad |\Phi| = 0, \quad [\Phi] \equiv \frac{1}{2} (\langle \Phi \rangle^2 - \langle \Phi^2 \rangle) = -\frac{1}{2} \langle \Phi^2 \rangle. \quad (13.3)$$

Only the second invariant $[\Phi]$ is in general non-vanishing whereas trace and determinant vanish and the Hamilton-Cayley identity for Φ reduces to

$$\Phi^3 + [\Phi] \Phi = 0, \quad (13.4)$$

with the consequence that all powers of Φ reduce to powers of Φ and Φ^2 multiplied by factors

$$\Phi^{2n+1} = (-[\Phi])^n \Phi, \quad \Phi^{2n+2} = (-[\Phi])^n \Phi^2, \quad (n = 0, 1, 2, \dots). \quad (13.5)$$

From the Taylor series $R = e^\Phi = \sum_{n=0}^{\infty} \frac{1}{n!} \Phi^n$ and analogously for $R^{-1} = e^{-\Phi} = R^\top$

⁵In representations of quantum field theory such as in [36], it is mostly given in more general form for the Lorentz group $SO(3,1) \sim SL(2, \mathbb{R}) \supset SO(3, \mathbb{R})$. The inversion of the relation analogously to second equation in (12.13) which is often absent can be found, e.g., in [25].

by substitution $\Phi \rightarrow -\Phi$ follows

$$R = I + \frac{\sin(\sqrt{[\Phi]})}{\sqrt{[\Phi]}} \Phi + \frac{1 - \cos(\sqrt{[\Phi]})}{[\Phi]} \Phi^2. \quad (13.6)$$

Due to relation $|B| = |e^A| = e^{\langle A \rangle}$ for an exponential operator $B \equiv e^A$ we check for determinant $|R|$

$$|R| = \exp(\langle \Phi \rangle) = \exp(0) = 1. \quad (13.7)$$

Using $|R| = 1$ with consequence $R^T = R^{-1} = \bar{R}$ and in addition $\langle \Phi^2 \rangle = -2[\Phi]$ according to (13.3) taking into account $\langle \Phi \rangle = 0$ follows explicitly from (13.6) for the trace and the second invariant of R

$$\langle R \rangle = [R] = 1 + 2 \cos(\sqrt{[\Phi]}). \quad (13.8)$$

The equality $\langle R \rangle = [R]$ of first and second invariant of R is due to $\langle R^T \rangle = \langle \bar{R} \rangle$ combined with the general identity $\langle \bar{A} \rangle = [A]$ for general three-dimensional operators A .

Our next problem is the transition from the antisymmetric operator Φ in the above formulae to a representation by the vector parameter φ by means of the formulae (13.2). First we make the transition of Φ^2 to a representation by φ

$$\Phi_{ik} \Phi_{km} = \varepsilon_{ijk} \varepsilon_{klm} \varphi_j \varphi_l = (\delta_{il} \delta_{jm} - \delta_{im} \delta_{jl}) \varphi_j \varphi_l = \varphi_i \varphi_m - |\varphi|^2 \delta_{im}, \quad (13.9)$$

which in coordinate-invariant representation is

$$\Phi^2 = \varphi \cdot \varphi - |\varphi|^2 I, \Rightarrow \langle \Phi^2 \rangle = -2|\varphi|^2 = -2[\Phi], \quad [\Phi] = |\varphi|^2. \quad (13.10)$$

Inserting (13.10) into (13.6) for R follows

$$R = \mathbf{n} \cdot \mathbf{n} + \sin(|\varphi|)[\mathbf{n}] + \cos(|\varphi|)(I - \mathbf{n} \cdot \mathbf{n}), \quad \left(\mathbf{n} \equiv \frac{\varphi}{|\varphi|}, n^2 = 1 \right). \quad (13.11)$$

The abbreviation $\mathbf{n}$ is a unit vector in direction of the rotation axis and $|\varphi|$ the rotation angle counter-clockwise taken. The application of R onto an arbitrary vector $\mathbf{x}$ leads to

$$R\mathbf{x} = (\mathbf{n}\mathbf{x})\mathbf{n} + \sin(|\varphi|)[\mathbf{n}, \mathbf{x}] + \cos(|\varphi|)[[\mathbf{n}, \mathbf{x}], \mathbf{n}]. \quad (13.12)$$

By comparison of $R = e^\Phi$ with the general form $e^{iJ\varphi}$ of group elements in representations of $SU(2)$ in (5.2) which in specialization to the vector basis becomes $R_{kl} = e^{\Phi_{kl}} = e^{\varepsilon_{klj}\varphi_j} = e^{-i(J_j)_{kl}\varphi_j}$ we find that the operators $J_j, (j=1,2,3)$ are represented in the three-dimensional regular representation by the following matrices closely related to the Levi-Civita symbol

$$J_j \rightarrow (J_j)_{kl} = i\varepsilon_{kjl} = -i\varepsilon_{jkl}. \quad (13.13)$$

One may check that this three-dimensional matrix representation of the operators $J_j, (j=1,2,3)$ satisfies the commutation relations (3.10) and may be taken as alternative starting point for the construction of the three-dimensional repre-

sensation of $SU(2)$.

14. Cayley-Gibbs-Fyodorov Parametrization of Rotation Group $SO(3, \mathbb{R})$ by Vector Parameter $\mathbf{c}$ in Coordinate-Invariant Description

There is yet another very interesting parametrization of the three-dimensional rotation group obtained by specialization from the Cayley representation (e.g., [7]) of proper orthogonal operators R of arbitrary dimension (*i.e.*, operators satisfying $R^{-1} = R^T$ with determinant $|R| = +1$) by antisymmetric operators C which is possible in general n -dimensional case as follows⁶

$$R \equiv \frac{I+C}{I-C} = \exp(2\text{Arth}(C)), \quad C = \frac{R-I}{R+I} = \text{th}\left(\frac{1}{2}\log(R)\right) = -C^T, \quad (14.1)$$

where the second relation of antisymmetry of C follows from

$$C^T = \frac{R^T - I}{R^T + I} = \frac{R^{-1} - I}{R^{-1} + I} = \frac{I - R}{I + R} = -C. \quad (14.2)$$

We now consider specific properties of the transformations in three-dimensional case. After expansion of R in (14.1) in a Taylor series of powers of C according to

$$R = I + 2(C + C^2 + C^3 + \dots), \quad C = -I + 2(R - R^2 + R^3 - \dots), \quad (C^T)^2 = C^2, \quad (14.3)$$

and reduction of powers of C higher than or equal 3 by the Hamilton-Cayley identity (6) (third equation) to powers lower than 3 taking into account the antisymmetry of C with the consequence $C^3 = -[C]C$, we find the following reduced relations between R and C

$$R = I + 2\frac{C + C^2}{1 + [C]}, \quad R^{-1} = R^T = I - 2\frac{C - C^2}{1 + [C]}, \quad C = \frac{R - R^T}{1 + \langle R \rangle} = -C^T, \quad (14.4)$$

with the invariants

$$\langle R \rangle = [R] = \frac{3 - [C]}{1 + [C]}, \quad |R| = 1, \quad \langle C \rangle = 0, \quad [C] = -\frac{1}{2}\langle C^2 \rangle = \frac{3 - \langle R \rangle}{1 + \langle R \rangle}, \quad |C| = 0. \quad (14.5)$$

They do not contain powers of R and C higher than quadratic ones. Furthermore using (13.1) and (14.1) we have

⁶It is not possible to represent operators of improper orthogonal transformations R' with

$|R'| = -1 = e^{i\pi}$ in this way since from $R' \equiv \frac{I+C'}{I-C'}$ and antisymmetry $C' = -C'^T$ automatically fol-

lows $|R'| = +1$, for example, from $|R'| = \exp(2\langle \text{Arth}(C') \rangle) = e^0 = 1$ according to $|e^A| = e^{\langle A \rangle}$ or from

$|R'| = \frac{|I+C'|}{|I-C'|} = 1$ due to $|I \pm C'| = 1 + [C']$ for antisymmetric matrices C' . However, with modifica-

tions not considered here a coordinate-invariant treatment of the case $|R'| = -1$ is possible. We

mention that the Cayley representation is often used in the alternative form $U = \frac{I + iH}{I - iH}$ or $iH = \frac{U - I}{U + I}$

as one of the possible correspondences between unitary operators U and Hermitean operators H [7].

$$C = \operatorname{th}\left(\frac{\Phi}{2}\right) = \frac{\Phi}{\sqrt{[\Phi]}} \operatorname{tg}\left(\frac{\sqrt{[\Phi]}}{2}\right), \quad \Phi = 2 \operatorname{Arth}(C) = 2 \frac{\operatorname{arctg}(\sqrt{[C]})}{\sqrt{[C]}} C, \quad (14.6)$$

with the invariants

$$[\Phi] = \left(2 \operatorname{arctg}(\sqrt{[C]})\right)^2 = |\boldsymbol{\varphi}|^2, \quad [C] = \operatorname{tg}^2\left(\frac{\sqrt{[\Phi]}}{2}\right) = \operatorname{tg}^2\left(\frac{|\boldsymbol{\varphi}|}{2}\right). \quad (14.7)$$

Independently on the choice of the chosen sign of $\sqrt{[\Phi]}$ and $\sqrt{[C]}$ the relations between C and Φ are true since they are involved only in the combinations

$$2 \frac{\operatorname{tg}\left(\frac{\sqrt{[\Phi]}}{2}\right)}{\sqrt{[\Phi]}} \quad \text{and} \quad \frac{\operatorname{arctg}(\sqrt{[C]})}{\sqrt{[C]}}, \quad \text{respectively.}$$

A further specifics of the three-dimensional case of the Cayley representation is that a real antisymmetric tensor C (or operator in Euclidean space) with its 3 independent components can be mapped $C \leftrightarrow \mathbf{c}$ onto a three-dimensional real vector $\mathbf{c}$ which in representation by vector indices in analogy to (13.1) takes on the form (see also ((10) and (13.1))); $C_{ik} \equiv \varepsilon_{ijk} c_j \leftrightarrow c_j \equiv \frac{1}{2} \varepsilon_{ijk} C_{ik}$

$$C \leftrightarrow C_{ik} \equiv [\mathbf{c}]_{ik} = \varepsilon_{ijk} c_j, \quad \mathbf{c} \leftrightarrow c_j = \frac{1}{2} \varepsilon_{ijk} C_{ik}, \quad \Rightarrow \quad C \equiv [\mathbf{c}] \leftrightarrow \mathbf{c}, \quad (14.8)$$

from which follows, for example

$$\begin{aligned} C\mathbf{x} &= [\mathbf{c}]\mathbf{x} = [\mathbf{c}, \mathbf{x}], \quad C[\mathbf{x}, \mathbf{y}] = [\mathbf{c}][\mathbf{x}, \mathbf{y}] = [\mathbf{c}, [\mathbf{x}, \mathbf{y}]] = (\mathbf{c}\mathbf{y})\mathbf{x} - (\mathbf{c}\mathbf{x})\mathbf{y}, \\ \mathbf{y}C &= \mathbf{y}[\mathbf{c}] = [\mathbf{y}, \mathbf{c}], \quad [\mathbf{x}, \mathbf{y}]C = [\mathbf{x}, \mathbf{y}][\mathbf{c}] = [[\mathbf{x}, \mathbf{y}], \mathbf{c}] = (\mathbf{x}\mathbf{c})\mathbf{y} - (\mathbf{y}\mathbf{c})\mathbf{x}. \end{aligned} \quad (14.9)$$

For the squared operator $C \equiv [\mathbf{c}]$ we find (C^2 is a symmetric operator)

$$C^2 = [\mathbf{c}]^2 = \mathbf{c} \cdot \mathbf{c} - |\mathbf{c}|^2 \mathbf{I}, \quad \Rightarrow \quad \langle C^2 \rangle = -2|\mathbf{c}|^2, \quad [C] = -\frac{1}{2} \langle C^2 \rangle = |\mathbf{c}|^2. \quad (14.10)$$

From this using (14.6) one finds the following relations between the vector parameters $\mathbf{c}$ and $\boldsymbol{\varphi}$

$$\mathbf{c} = \frac{\boldsymbol{\varphi}}{|\boldsymbol{\varphi}|} \operatorname{tg}\left(\frac{|\boldsymbol{\varphi}|}{2}\right), \quad \boldsymbol{\varphi} = \frac{\mathbf{c}}{|\mathbf{c}|} 2 \operatorname{arctg}(|\mathbf{c}|), \quad 0 \leq |\boldsymbol{\varphi}| \leq \pi, \quad 0 \leq |\mathbf{c}| \leq +\infty, \quad (14.11)$$

with the consequence

$$\begin{aligned} |\mathbf{c}| &= \operatorname{tg}\left(\frac{|\boldsymbol{\varphi}|}{2}\right), \quad \frac{\mathbf{c}}{|\mathbf{c}|} = \frac{\boldsymbol{\varphi}}{|\boldsymbol{\varphi}|}, \quad \Rightarrow \quad \mathbf{c} = \frac{\mathbf{c}}{|\mathbf{c}|} |\mathbf{c}| = \frac{\boldsymbol{\varphi}}{|\boldsymbol{\varphi}|} \operatorname{tg}\left(\frac{|\boldsymbol{\varphi}|}{2}\right), \\ \cos\left(\frac{|\boldsymbol{\varphi}|}{2}\right) &= \pm \frac{1}{\sqrt{1+|\mathbf{c}|^2}}, \quad \sin\left(\frac{|\boldsymbol{\varphi}|}{2}\right) = \frac{|\mathbf{c}|}{\sqrt{1+|\mathbf{c}|^2}}, \quad 0 \leq |\boldsymbol{\varphi}| \leq \pi. \end{aligned} \quad (14.12)$$

The direct relation to the trigonometric functions in the rotation matrix R_{kl} given in (12.5) by the vector parameter $\boldsymbol{\varphi}$ is

$$\cos(|\boldsymbol{\varphi}|) = \frac{1-|\mathbf{c}|^2}{1+|\mathbf{c}|^2}, \quad \sin(|\boldsymbol{\varphi}|) = \pm \frac{2|\mathbf{c}|}{1+|\mathbf{c}|^2}, \quad 0 \leq |\boldsymbol{\varphi}| \leq \pi, \quad (14.13)$$

and this rotation matrix R_{kl} is therefore uniquely represented by vector parameter $\mathbf{c}$ as follows

$$R_{kl} = \frac{c_k c_l}{|\mathbf{c}|^2} + \frac{1 - |\mathbf{c}|^2}{1 + |\mathbf{c}|^2} \left(\delta_{kl} - \frac{c_k c_l}{|\mathbf{c}|^2} \right) + \frac{2}{1 + |\mathbf{c}|^2} \varepsilon_{jkl} c_j. \quad (14.14)$$

The modulus $|\mathbf{c}|$ of the vector parameter $\mathbf{c}$ with real components (c_1, c_2, c_3) is stretched in comparison to the modulus $|\boldsymbol{\varphi}|$ of the vector parameter $\boldsymbol{\varphi}$ with real components $(\varphi_1, \varphi_2, \varphi_3)$.

All rotations of $SO(3, \mathbb{R})$ about an angle $|\boldsymbol{\varphi}| = \pi$ correspond to the parameter $|\mathbf{c}| \rightarrow \infty$ and in the whole region $0 \leq |\boldsymbol{\varphi}| \leq \pi$ we have the correspondences of $\frac{\mathbf{c}}{|\mathbf{c}|}$ to $\frac{\boldsymbol{\varphi}}{|\boldsymbol{\varphi}|}$ as the rotation axis. The region to angles $\pi \leq |\boldsymbol{\varphi}| \leq 2\pi$ can be reduced by transition $|\boldsymbol{\varphi}| \rightarrow 2\pi - |\boldsymbol{\varphi}|$ to angles in the region $0 \leq \pi - |\boldsymbol{\varphi}| \leq \pi$ with simultaneous transition to the opposite rotation axis $\frac{\boldsymbol{\varphi}}{|\boldsymbol{\varphi}|} \rightarrow -\frac{\boldsymbol{\varphi}}{|\boldsymbol{\varphi}|}$. This is not possible for the fundamental representation of $SU(2)$ and we cannot find equivalent angles within the region $0 \leq |\boldsymbol{\varphi}| \leq 2\pi$ which are equivalent by changing the direction $\frac{\boldsymbol{\varphi}}{|\boldsymbol{\varphi}|}$ of the rotation axis.

The relation (14.11) between $\boldsymbol{\varphi}$ and $\mathbf{c}$ maps all vectors of the three-dimensional ball of vectors $\boldsymbol{\varphi}$ with $|\boldsymbol{\varphi}| \leq \pi$ onto the three-dimensional Euclidean space of vectors $\mathbf{c}$ where opposite points $\pm \boldsymbol{\varphi}$ of the boundary $|\boldsymbol{\varphi}| = \pi$ correspond to single points of $\mathbf{c}$. For $SO(3, \mathbb{R})$ the boundary corresponds to $|\boldsymbol{\varphi}| = \pi$ where opposite directions have to be identified. For $SU(2)$ the parameter $\mathbf{c}$ does not uniquely determine a transformation S of the form (5.13).

The two-dimensional matrix S of the fundamental representation of $SU(2)$ in (5.13) takes on the following very simple form in representation by the vector parameter $\mathbf{c}$ in components $\mathbf{c} \equiv (c_1, c_2, c_3)$

$$S = \pm \frac{1}{\sqrt{1 + |\mathbf{c}|^2}} \begin{pmatrix} 1 + ic_3 & ic_- \\ ic_+ & 1 - ic_3 \end{pmatrix}, \quad c_{\pm} \equiv c_1 \pm ic_2 = c_{\mp}^*, \quad (14.15)$$

The minus sign in front of $\frac{1}{\sqrt{1 + |\mathbf{c}|^2}}$ appears because $\cos\left(\frac{|\boldsymbol{\varphi}|}{2}\right) = \pm \frac{1}{\sqrt{1 + |\mathbf{c}|^2}}$ changes its sign in the region $0 \leq |\boldsymbol{\varphi}| \leq 2\pi$. The special case $c_3 = 0$ corresponds to the parametrization of the special element g_n of $SU(2)$ by Perelomov [32] (Section 4.1) where the complex parameter ζ there corresponds to our ic_- . Furthermore, by comparison of (5.13) and (14.15) with (11.5) we find that the Cayley-Gibbs-Fyodorov parametrization possesses the following relation to the quaternion representation of $SU(2)$ matrices

$$S \leftrightarrow (r_0, \mathbf{r}) = \left(\cos\left(\frac{|\boldsymbol{\varphi}|}{2}\right), -\frac{\boldsymbol{\varphi}}{|\boldsymbol{\varphi}|} \sin\left(\frac{|\boldsymbol{\varphi}|}{2}\right) \right) = \left(\pm \frac{1}{\sqrt{1 + |\mathbf{c}|^2}}, -\frac{\mathbf{c}}{\sqrt{1 + |\mathbf{c}|^2}} \right), \quad \mathbf{c} = -\frac{\mathbf{r}}{r_0}. \quad (14.16)$$

This means that the Cayley-Gibbs-Fyodorov parametrization and the quaternion representation of $SU(2)$ are connected by a simple relation but the parametrization of the group $SU(2)$ by the parameter c is not unique.

The parametrization of elements of $SU(2)$ and of the three-dimensional rotation group $SO(3, \mathbb{R})$ by the three-dimensional vector parameter c which is dual to the antisymmetric operator C of the Cayley representation was introduced by Gibbs (according to [6]) and was used by Fyodorov as the fundament of his approach to the representation theory of this group in coordinate-invariant representation and, furthermore, was extended by him to the Lorentz group [6]. This parametrization is advantageous, in particular, for the coordinate-invariant representation of the composition law of two, in general, non-commuting operators in these groups. We consider this in next Section.

15. Composition Law for the Vector Parameters φ and c Corresponding to Products of Rotations of $SO(3, \mathbb{R})$

As in the case of $SU(2)$ (Section 9) we derive now the composition law of the two vector parameters φ_1 and φ_2 to a new vector parameter φ for the product of two rotations R_1 and R_2 according to

$$R = R_1 R_2 \leftrightarrow \varphi, \quad R_1 \leftrightarrow \varphi_1, \quad R_2 \leftrightarrow \varphi_2. \quad (15.1)$$

This composition law can be obtained from the representation (13.11) of the rotation operator R and possesses the explicit form

$$\begin{aligned} \cos(|\varphi|) &= \cos(|\varphi_1|) \cos(|\varphi_2|) - \frac{\varphi_1 \varphi_2}{|\varphi_1| |\varphi_2|} \sin(|\varphi_1|) \sin(|\varphi_2|), \\ \frac{\varphi}{|\varphi|} \sin(|\varphi|) &= \frac{1}{2} \left\{ \frac{\varphi_1}{|\varphi_1|} \left(\sin(|\varphi_1|) (1 + \cos(|\varphi_2|)) - \frac{\varphi_1 \varphi_2}{|\varphi_1| |\varphi_2|} (1 - \cos(|\varphi_1|)) \sin(|\varphi_2|) \right) \right. \\ &\quad + \frac{\varphi_2}{|\varphi_2|} \left((1 + \cos(|\varphi_1|)) \sin(|\varphi_2|) - \frac{\varphi_1 \varphi_2}{|\varphi_1| |\varphi_2|} \sin(|\varphi_1|) (1 - \cos(|\varphi_2|)) \right) \\ &\quad \left. - \frac{[\varphi_1, \varphi_2]}{|\varphi_1| |\varphi_2|} \left(\sin(|\varphi_1|) \sin(|\varphi_2|) - \frac{\varphi_1 \varphi_2}{|\varphi_1| |\varphi_2|} (1 - \cos(|\varphi_1|)) (1 - \cos(|\varphi_2|)) \right) \right\}. \end{aligned} \quad (15.2)$$

It seems to be also possible to find it from the composition law for $SU(2)$ using the formulae (9.2) and (9.3).

We calculate now the antisymmetric operator C for the product of two three-dimensional rotations $R = R_1 R_2$ (i.e., $R_1^{-1} = R_1^T$, $R_2^{-1} = R_2^T$). If we substitute $R \rightarrow R_1 R_2$ in the formula for C in last of Equations (14.4) and express R_1 and R_2 by the antisymmetric operators C_1 and C_2 we find

$$\begin{aligned} C &= \frac{R_1 R_2 - (R_1 R_2)^{-1}}{1 + \langle R_1 R_2 \rangle} \\ &= \frac{\left(I + 2 \frac{C_1 + C_1^2}{1 + [C_1]} \right) \left(I + 2 \frac{C_2 + C_2^2}{1 + [C_2]} \right) - \left(I - 2 \frac{C_2 - C_2^2}{1 + [C_2]} \right) \left(I - 2 \frac{C_1 - C_1^2}{1 + [C_1]} \right)}{1 + \langle R_1 R_2 \rangle} \end{aligned}$$

$$\begin{aligned}
&= 4 \frac{(1+[C_2])C_1 + (1+[C_1])C_2 + [C_1, C_2] + C_1^2 C_2 + C_2 C_1^2 + C_1 C_2^2 + C_2^2 C_1 + [C_1^2, C_2^2]}{(1+\langle R_1 R_2 \rangle)(1+[C_1])(1+[C_2])} \\
&= 4 \frac{\left(1 + \frac{1}{2}\langle C_1 C_2 \rangle\right)(C_1 + C_2 + [C_1, C_2])}{(1+\langle R_1 R_2 \rangle)(1+[C_1])(1+[C_2])}, \quad (\langle C_1^2 \rangle = -2[C_1], \langle C_2^2 \rangle = -2[C_2]),
\end{aligned} \tag{15.3}$$

where we used the specific three-dimensional identities for general antisymmetric operators C_1 and C_2

$$C_1^2 C_2 + C_2 C_1^2 = \frac{1}{2}(\langle C_1 C_2 \rangle C_1 + \langle C_1^2 \rangle C_2), \quad [C_1^2, C_2^2] = \frac{1}{2}\langle C_1 C_2 \rangle [C_1, C_2], \tag{15.4}$$

which can be directly checked. From the first of these identities and after multiplication of it with C_2 and then forming the traces we find scalar identities for general antisymmetric operators C_1 and C_2 as follows

$$\langle C_1^2 C_2 \rangle = 0, \quad \langle C_1^2 C_2^2 \rangle = \frac{1}{4}(\langle C_1 C_2 \rangle^2 + \langle C_1^2 \rangle \langle C_2^2 \rangle), \quad \langle (C_1 C_2)^2 \rangle = \frac{1}{2}\langle C_1 C_2 \rangle^2. \tag{15.5}$$

Furthermore, starting from first of Equations (14.4) with $R = R_1 R_2$ and using the identities (15.5) we find

$$1 + \langle R_1 R_2 \rangle = \frac{4\left(1 + \frac{1}{2}\langle C_1 C_2 \rangle\right)^2}{(1+[C_1])(1+[C_2])}. \tag{15.6}$$

Inserting the relation (15.6) into (15.3) this provides the composition law for two rotations in the following final operator form

$$R = R_1 R_2, \quad \Leftrightarrow \quad C = \frac{C_1 + C_2 + [C_1, C_2]}{1 + \frac{1}{2}\langle C_1 C_2 \rangle} = -C^\top, \tag{15.7}$$

with the only non-vanishing invariant $[C] = -\frac{1}{2}\langle C^2 \rangle$ of the antisymmetric operator C

$$\begin{aligned}
\langle C^2 \rangle &= \frac{\langle C_1^2 \rangle + \langle C_2^2 \rangle + 2\langle C_1 C_2 \rangle + \frac{1}{2}(\langle C_1 C_2 \rangle^2 - \langle C_1^2 \rangle \langle C_2^2 \rangle)}{\left(1 + \frac{1}{2}\langle C_1 C_2 \rangle\right)^2} = -2[C], \\
\Rightarrow 1 + [C] &= 1 - \frac{1}{2}\langle C^2 \rangle = \frac{\left(1 - \frac{1}{2}\langle C_1^2 \rangle\right)\left(1 - \frac{1}{2}\langle C_2^2 \rangle\right)}{\left(1 + \frac{1}{2}\langle C_1 C_2 \rangle\right)^2} = \frac{(1+[C_1])(1+[C_2])}{\left(1 + \frac{1}{2}\langle C_1 C_2 \rangle\right)^2}.
\end{aligned} \tag{15.8}$$

From (15.7) using the definition (14.8) of the vector parameter $\mathbf{c}$ by the antisymmetric operator C follows the composition law expressed by the vector parameter $\mathbf{c}$ in agreement with [6] [37]⁷

⁷Fyodorov introduces in formula (2.4) in [6] c_j by $C_{kl} \equiv c_{kl}^* = \varepsilon_{kjl} c_j$ but its inversion has to be then $c_j = \frac{1}{2}\varepsilon_{kjl} C_{kl}$ that by the sign is not identical with our (14.8) and is only a mistake since it is correct in earlier [5]. Pars [37] (Equation (7.9.2), without citation) has a minus sign in front of the vector product but this substitution changes only the direction of the rotation.

$$\mathbf{c} \equiv \langle \mathbf{c}_1, \mathbf{c}_2 \rangle = \frac{\mathbf{c}_1 + \mathbf{c}_2 + [\mathbf{c}_1, \mathbf{c}_2]}{1 - \mathbf{c}_1 \mathbf{c}_2}, \quad (15.9)$$

and is near to a formula of Rodrigues cited, e.g., in [33] and [35] (p. 17). The appearance of the vector product $[\mathbf{c}_1, \mathbf{c}_2]$ in this relation is the consequence that the order of two rotations $R_1 R_2$ does not commute. A linearization is obtained by stretching $R\mathbf{c} \rightarrow \mathbf{x}$ (Appendix C). From (15.9) follows for the modulus $|\mathbf{c}|$

$$|\mathbf{c}| \equiv \sqrt{\mathbf{c}^2} = \pm \frac{\sqrt{(\mathbf{c}_1 + \mathbf{c}_2)^2 + [\mathbf{c}_1, \mathbf{c}_2]^2}}{1 - \mathbf{c}_1 \mathbf{c}_2}, \Rightarrow 1 + \mathbf{c}^2 = \frac{(1 + \mathbf{c}_1^2)(1 + \mathbf{c}_2^2)}{(1 - \mathbf{c}_1 \mathbf{c}_2)^2}. \quad (15.10)$$

Our derivation of the composition law (15.9) of two vector parameters $\mathbf{c}_1$ and $\mathbf{c}_2$ distinguishes from the more directly obtained of Fyodorov by using specific three-dimensional identities for antisymmetric operators. The composition of two vector parameters $\mathbf{c}_1$ and $\mathbf{c}_2$ to a new vector parameter $\mathbf{c}$ for the product of two rotations $R_1 R_2 = R$ can be obtained from (12.5) but this is also possible using (9.2)-(9.4). The composition law (9.3) for the vector parameter $\boldsymbol{\varphi}$ is more complicated than that for the vector parameter $\mathbf{c}$ and can be also obtained from (15.9) using (14.8) and (14.6). This is an advantage of using the vector parameter $\mathbf{c}$ which, however, is nonlinear in $\boldsymbol{\varphi}$.

Following Fyodorov ([6]) we introduced in (9.5) and used in (15.9) the convenient symbol $\langle \mathbf{c}_1, \mathbf{c}_2 \rangle$ for the composition of two vector parameters $\mathbf{c}_1$ and $\mathbf{c}_2$ corresponding to the product $R_1 R_2$ or $S_1 S_2$. It allows to formulate some composition rules in a simple way. In general, due to non-commutativity but associativity of the group multiplication one has

$$\langle \mathbf{c}_1, \mathbf{c}_2 \rangle \neq \langle \mathbf{c}_2, \mathbf{c}_1 \rangle, \quad \langle \langle \mathbf{c}_1, \mathbf{c}_2 \rangle, \mathbf{c}_3 \rangle = \langle \mathbf{c}_1, \langle \mathbf{c}_2, \mathbf{c}_3 \rangle \rangle \equiv \langle \mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3 \rangle, \quad (15.11)$$

meaning that one may omit the inner brackets in the composition of three transformations. Explicitly, we calculate in considered case ($[\mathbf{a}, \mathbf{b}, \mathbf{c}]$ denotes the volume product of three vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}$)

$$\langle \mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3 \rangle = \frac{(1 - \mathbf{c}_2 \mathbf{c}_3) \mathbf{c}_1 + (1 + \mathbf{c}_1 \mathbf{c}_3) \mathbf{c}_2 + (1 - \mathbf{c}_1 \mathbf{c}_2) \mathbf{c}_3 + [\mathbf{c}_1, \mathbf{c}_2] + [\mathbf{c}_1, \mathbf{c}_3] + [\mathbf{c}_2, \mathbf{c}_3]}{1 - \mathbf{c}_1 \mathbf{c}_2 - \mathbf{c}_1 \mathbf{c}_3 - \mathbf{c}_2 \mathbf{c}_3 - [\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3]}. \quad (15.12)$$

An important special case of (15.12) corresponds to $\mathbf{R}' = \mathbf{R}_0 \mathbf{R} \mathbf{R}_0^{-1}$ of $SO(3, \mathbb{R})$ where $\mathbf{R}$ and $\mathbf{R}_0$ which provide all conjugate elements $\mathbf{R}'$. Since in both cases the transition to the inverse element means the transition $\boldsymbol{\varphi} \leftrightarrow -\boldsymbol{\varphi}$ or $\mathbf{c} \leftrightarrow -\mathbf{c}$, respectively, we have to calculate $\mathbf{c}' = \langle \mathbf{c}_0, \mathbf{c}, -\mathbf{c}_0 \rangle$ for which we find from (15.12)

$$\begin{aligned} \mathbf{c}' = \langle \mathbf{c}_0, \mathbf{c}, -\mathbf{c}_0 \rangle &\equiv \frac{(1 - |\mathbf{c}_0|^2) \mathbf{c} + 2(\mathbf{c}_0 \mathbf{c}) \mathbf{c}_0 - 2[\mathbf{c}_0, \mathbf{c}]}{1 + |\mathbf{c}_0|^2}, \\ \Rightarrow |\mathbf{c}'|^2 &= |\mathbf{c}|^2, \quad \mathbf{c}' \mathbf{c}_0 = \mathbf{c} \mathbf{c}_0, \end{aligned} \quad (15.13)$$

This means that the vector parameters $\mathbf{c}$ to conjugate elements of the groups $SO(3, \mathbb{R})$ possess the same modulus (length) $|\mathbf{c}|$ and therefore correspond to

rotations about the same angle but, in general, to different rotation axes

$\mathbf{n} \equiv \frac{\mathbf{c}}{|\mathbf{c}|} = \frac{\boldsymbol{\varphi}}{|\boldsymbol{\varphi}|}$ and $\mathbf{n}' \equiv \frac{\mathbf{c}'}{|\mathbf{c}'|} = \frac{\boldsymbol{\varphi}'}{|\boldsymbol{\varphi}'|}$. If the rotation axes of $\mathbf{c}$ and of $\mathbf{c}_0$ are parallel that means $[\mathbf{c}_0, \mathbf{c}] = \mathbf{0}$ then one finds

$$[\mathbf{c}_0, \mathbf{c}] = \mathbf{0}, \Rightarrow \mathbf{c}' = \langle \mathbf{c}_0, \mathbf{c}, -\mathbf{c}_0 \rangle \equiv \frac{(1 - |\mathbf{c}_0|^2)\mathbf{c} + 2(|\mathbf{c}_0|^2)\mathbf{c}}{1 + |\mathbf{c}_0|^2} = \mathbf{c}. \quad (15.14)$$

The formulae (15.13) and (15.14) show that rotations around arbitrary axes but with the same rotation angle belong to the same class of conjugate elements and that with different rotation angles to different classes of conjugate elements.

16. Invariant Integration over Group $SO(3, \mathbb{R})$ and $SU(2)$

The s -dimensional irreducible representations $D_k^i(g)$ of a finite group G with N elements $g \in G$ possess for any fixed $h \in G$ the following orthogonality relations in index notation for operators, e.g., [20] [26]

$$\frac{1}{N} \sum_{g \in G} D_k^i(g) D_l^j(g^{-1}) = \frac{1}{N} \sum_{g \in G} D_k^i(gh) D_l^j((gh)^{-1}) = \frac{1}{s} \delta_k^j \delta_l^i. \quad (16.1)$$

In transition to a Lie group one has to substitute $g \in G$ by the chosen parameter, in our case of $SO(3, \mathbb{R})$, by the vector parameter $\mathbf{c}$ or by the vector parameter $\boldsymbol{\varphi}$ and the summation over the discrete elements g by integration over the chosen parameter, where g^{-1} must be substituted by the corresponding parameter for the inverse element, in our case by $-\mathbf{c}$ that means in our first considered case and h by a fixed parameter $\mathbf{c}_0$

$$\begin{aligned} & \frac{1}{V_G} \int_G d^3 c \mu(\mathbf{c}) D_k^i(\mathbf{c}) D_l^j(-\mathbf{c}) \\ &= \frac{1}{V_G} \int_G d^3 c \mu(\mathbf{c}) D_k^i(\langle \mathbf{c}, \mathbf{c}_0 \rangle) D_l^j(-\langle \mathbf{c}, \mathbf{c}_0 \rangle) = \frac{1}{s} \delta_k^j \delta_l^i, \end{aligned} \quad (16.2)$$

with V_G the group volume in chosen parametrization defined by⁸

$$V_G \equiv \int_G d^3 c \mu(\mathbf{c}) = \int_G d\mathbf{c}_1 \wedge d\mathbf{c}_2 \wedge d\mathbf{c}_3 \mu(\mathbf{c}). \quad (16.3)$$

The function $\frac{\mu(\mathbf{c})}{V_G}$ corresponding to $\frac{1}{N}$ in (16.1) is an “invariant” measure which gives the weight of every group element to parameter $\mathbf{c}$ in the chosen parametrization and depends on it.

In this Section we derive the invariance of a certain measure of the group $SO(3, \mathbb{R})$ and by parameter transformation of a measure of $SU(2)$. For this purpose we make the substitutions $\mathbf{c} \rightarrow \mathbf{c}'$, $\mathbf{c}_1 \rightarrow \mathbf{c}$, $\mathbf{c}_2 \rightarrow d\mathbf{c}'$ in the composition formula (15.9), leading to

$$\mathbf{c}' = \frac{\mathbf{c} + \mathbf{c}_0 + |\mathbf{c}, \mathbf{c}_0|}{1 - \mathbf{c}\mathbf{c}_0}, \Rightarrow 1 + \mathbf{c}'^2 = \frac{(1 + \mathbf{c}^2)(1 + \mathbf{c}_0^2)}{(1 - \mathbf{c}\mathbf{c}_0)^2}, \quad (16.4)$$

⁸In our case of vector parameter $\mathbf{c}$ the weight $\mu(\mathbf{c})$ depends only on the modulus $|\mathbf{c}|$.

where the fixed $\mathbf{c}_0$ transforms an arbitrary vector $\mathbf{c}$ into a certain vector $\mathbf{c}'$. The general composition law $\mathbf{c} \rightarrow \mathbf{c}'$ is nonlinear. We calculate now the differential change of $\mathbf{c} + d\mathbf{c}'$ for a differential change of $\mathbf{c} + d\mathbf{c}$ with fixed $\mathbf{x}_0$

$$dc'_i = \frac{\partial c'_i}{\partial c_l} dc_l = \left\{ \frac{\delta_{il} - \varepsilon_{ikl} c_{0,k}}{1 - \mathbf{c}\mathbf{c}_0} + \frac{(c_i + c_{0,i} + \varepsilon_{ijk} c_j c_{0,k}) c_{0,l}}{(1 - \mathbf{c}\mathbf{c}_0)^2} \right\} dc_l. \quad (16.5)$$

or in coordinate invariant notation

$$d\mathbf{c}' = \frac{(1 - \mathbf{c}\mathbf{c}_0)(\mathbf{I} - [\mathbf{c}_0]) + (\mathbf{c} + \mathbf{c}_0 + [\mathbf{c}, \mathbf{c}_0]) \cdot \mathbf{c}_0}{(1 - \mathbf{c}\mathbf{c}_0)^2} d\mathbf{c}. \quad (16.6)$$

With substitutions $\alpha = \gamma \rightarrow 1 - \mathbf{c}\mathbf{c}_0$, $\beta \rightarrow 1$ and $\mathbf{b} \rightarrow \mathbf{c} + \mathbf{c}_0 + [\mathbf{c}, \mathbf{c}_0]$, $\tilde{\mathbf{b}} = \mathbf{c} \rightarrow \mathbf{c}_0$ we find from (D.2) in **Appendix D** for the determinant in numerator of (16.6) the astonishingly simple result

$$|(1 - \mathbf{c}\mathbf{c}_0)(\mathbf{I} - [\mathbf{c}_0]) + (\mathbf{c} + \mathbf{c}_0 + [\mathbf{c}, \mathbf{c}_0]) \cdot \mathbf{c}_0| = (1 - \mathbf{c}\mathbf{c}_0)^2 (1 + \mathbf{c}_0^2)^2. \quad (16.7)$$

Now we calculate from (16.5) using the determinant (16.7) and the identity in (16.4) the changes of the volumes

$$dc'_1 \wedge dc'_2 \wedge dc'_3 = \frac{(1 + \mathbf{c}_0^2)^2}{(1 - \mathbf{c}\mathbf{c}_0)^4} dc_1 \wedge dc_2 \wedge dc_3 = \frac{(1 + \mathbf{c}'^2)^2}{(1 + \mathbf{c}^2)^2} dc_1 \wedge dc_2 \wedge dc_3 \quad (16.8)$$

This is not an invariant volume element for the group but if we divide both sides by $(1 + \mathbf{c}'^2)^2$ we find

$$dV_G = \frac{dc'_1 \wedge dc'_2 \wedge dc'_3}{(1 + \mathbf{c}'^2)^2} = \frac{dc_1 \wedge dc_2 \wedge dc_3}{(1 + \mathbf{c}^2)^2}. \quad (16.9)$$

However, this is invariant after the transformation $\mathbf{c} \rightarrow \mathbf{c}'$ and does not depend on the chosen fixed parameter $\mathbf{c}_0$. For the group volume we calculate

$$V_G = \int_{-\infty}^{+\infty} dc_1 \int_{-\infty}^{+\infty} dc_2 \int_{-\infty}^{+\infty} dc_3 \frac{1}{(1 + (c_1^2 + c_2^2 + c_3^2))^2} = \int_0^\infty \frac{4\pi d|\mathbf{c}||\mathbf{c}|^2}{(1 + |\mathbf{c}|^2)^2} = \pi^2. \quad (16.10)$$

Thus the invariant measure of $SO(3, \mathbb{R})$ with integration of $\mathbf{c}$ over the whole space $\mathbf{c} \in \mathbb{R}^3$ is

$$\frac{1}{\pi^2} \frac{dc_1 \wedge dc_2 \wedge dc_3}{(1 + \mathbf{c}^2)^2}, \quad \mu(\mathbf{c}) \equiv \frac{1}{(1 + \mathbf{c}^2)^2}, \quad (16.11)$$

for certain translation-invariant functions (e.g., representations of the group, in particular, the identity representation). This result agrees with Fyodorov [6] (§. 6) who give a modified derivation, likely, the only known such result.

From relation (14.11) between the vector parameters $\mathbf{c}$ and $\boldsymbol{\varphi}$ follows

$$dc_i = \frac{\partial c_i}{\partial \varphi_k} d\varphi_k = \left\{ \frac{\operatorname{tg}\left(\frac{|\boldsymbol{\varphi}|}{2}\right)}{|\boldsymbol{\varphi}|} \left(\delta_{ik} - \frac{\varphi_i \varphi_k}{|\boldsymbol{\varphi}|^2} \right) + \frac{1}{2 \cos^2\left(\frac{|\boldsymbol{\varphi}|}{2}\right)} \frac{\varphi_i \varphi_k}{|\boldsymbol{\varphi}|^2} \right\} d\varphi_k, \quad (16.12)$$

and using the determinant of the transformation matrix between dc_i and $d\varphi_k$ we find

$$d^3c = \frac{\operatorname{tg}^2\left(\frac{|\varphi|}{2}\right)}{2|\varphi|^2 \cos^2\left(\frac{|\varphi|}{2}\right)} d^3\varphi, \quad d^3c \equiv dc_1 \wedge dc_2 \wedge dc_3, \quad d^3\varphi \equiv d\varphi_1 \wedge d\varphi_2 \wedge d\varphi_3. \quad (16.13)$$

Due to (14.11) and taking into account $\frac{1}{1+|c|^2} = \cos^2\left(\frac{|\varphi|}{2}\right)$ (see (14.8)), we may

choose the volume element dV_G of the invariant integration over the group $SO(3, \mathbb{R})$ as follows

$$dV_G = \frac{1}{8} \left(\frac{\sin\left(\frac{|\varphi|}{2}\right)}{\frac{|\varphi|}{2}} \right)^2 d^3\varphi = \frac{1}{2} \sin^2\left(\frac{|\varphi|}{2}\right) d|\varphi| \wedge d^2n, \quad n \equiv \frac{\varphi}{|\varphi|}, \quad (16.14)$$

where d^2n denotes the area element of the surface $\mathbb{S}^2$ of unit vectors $|n|=1$ of a three-dimensional ball $\mathbb{B}$ and the integration of $|\varphi|$ goes from $|\varphi|=0$ up to $|\varphi|=\pi$ for $SO(3, \mathbb{R})$ and from $|\varphi|=0$ up to $|\varphi|=2\pi$ for $SU(2)$ with weight $d|\varphi||\varphi|^2 \rightarrow d|\varphi| \frac{1}{2} \sin^2\left(\frac{|\varphi|}{2}\right)$

$$\begin{aligned} V_G &= 4\pi \int_0^\pi d|\varphi| \frac{1}{2} \sin^2\left(\frac{|\varphi|}{2}\right) = \pi^2, \quad (SO(3, \mathbb{R})) \\ V_G &= 4\pi \int_0^{2\pi} d|\varphi| \frac{1}{2} \sin^2\left(\frac{|\varphi|}{2}\right) = 2\pi^2, \quad (SU(2)). \end{aligned} \quad (16.15)$$

This suggests that the invariant measure of $SO(3, \mathbb{R})$ using the vector parameter φ is

$$\frac{1}{8\pi^2} \left(\frac{\sin\left(\frac{|\varphi|}{2}\right)}{\frac{|\varphi|}{2}} \right)^2 d^3\varphi, \quad \mu(\varphi) \equiv \frac{1}{8} \left(\frac{\sin\left(\frac{|\varphi|}{2}\right)}{\frac{|\varphi|}{2}} \right)^2, \quad (SO(3, \mathbb{R})), \quad (16.16)$$

with integration of appropriate functions over a ball of radius $|\varphi|=\pi$ and the invariant measure of $SU(2)$ is

$$\frac{1}{16\pi^2} \left(\frac{\sin\left(\frac{|\varphi|}{2}\right)}{\frac{|\varphi|}{2}} \right)^2 d^3\varphi, \quad \mu(\varphi) \equiv \frac{1}{8} \left(\frac{\sin\left(\frac{|\varphi|}{2}\right)}{\frac{|\varphi|}{2}} \right)^2, \quad (SU(2)), \quad (16.17)$$

with integration of appropriate functions over a ball of radius $|\varphi|=2\pi$. Result (16.16) is similar to that for mean value and invariant measure in the parametrization by Euler angles (Lyubarskij [20], §. 16).

17. Stereographic Projection of Unit Sphere from North Pole to an Extended Complex Equator Plane and Its Connection to 3D Rotations

We now deal with a relation between rotations of the surface $\mathbb{S}^2$ of a unit ball $\mathbb{B}^3$ in 3-dimensional space (Riemann sphere) and fractional linear transformation (Möbius transformations) of an extended complex plane $\bar{\mathbb{C}} \equiv \mathbb{C} \cup \infty$ by stereographic projection of the unit sphere onto the complex plane.

We consider a plane $\mathbb{R}^2 \rightarrow \mathbb{C}$ with the Cartesian coordinates (u, v) embedded as Equatorial plane $z = 0$ into a 3-dimensional Euclidean space $\mathbb{R}^3$ with the Cartesian coordinates (x, y, z) . In the 3-dimensional space $\mathbb{R}^3$ we consider the surface of a sphere $\mathbb{S}^2$ with unit radius $R = 1$ described by

$$x^2 + y^2 + z^2 = 1, \quad (17.1)$$

and make the stereographic projection of this spherical surface from the North Pole $(x, y, z) = (0, 0, 1)$ onto the plane with the coordinates $(x = u, y = v)$. The relation of the coordinates of the unit sphere (x, y, z) with their projection points (u, v) onto the Equatorial plane is given by the nonlinear mapping (see **Figure 6**)

$$(x, y, z) \rightarrow (u, v) = \left(\frac{x}{1-z}, \frac{y}{1-z} \right), \quad \left(z = \pm \sqrt{1-x^2-y^2} \equiv \pm \sqrt{1-r^2} \right), \quad (17.2)$$

$$s \equiv \sqrt{u^2+v^2} = \frac{\sqrt{x^2+y^2}}{1-z} \equiv \frac{r}{1-z}, \quad \left(\frac{s}{R} = \frac{s-r}{z}, \quad r \equiv \sqrt{x^2+y^2}, \quad R=1 \right),$$

with the unique inversion $(0 \leq r \leq 1, 0 \leq s \leq \infty)$

$$(u, v) \rightarrow (x, y, z) = \left(\frac{2u}{u^2+v^2+1}, \frac{2v}{u^2+v^2+1}, \frac{u^2+v^2-1}{u^2+v^2+1} \right), \quad \Rightarrow \quad x^2 + y^2 + z^2 = 1, \quad (17.3)$$

$$s = \frac{r}{1-z} = \frac{r}{1 \mp \sqrt{1-r^2}} = \frac{1 \pm \sqrt{1-r^2}}{r}, \quad \Rightarrow \quad 0 \leq (1-rs)^2 = 1-r^2 \leq 1.$$

The Northern part of the spherical surface $\mathbb{S}^2$ is mapped onto the outer part $u^2 + v^2 > 1$ of the unit disc $(\mathbb{D} : u^2 + v^2 \leq 1)$ and the Southern part of the sphere to the inner part $u^2 + v^2 < 1$ of this unit disc $\mathbb{D}$. The Equator E of the sphere is mapped onto the border circle $u^2 + v^2 = 1$ of the unit disc $\mathbb{D}$, the South pole to the center $(u, v) = (0, 0)$ and the North Pole to the infinity point of the extended complex plane $\bar{\mathbb{C}}$.

Instead of the real coordinates (u, v) and (x, y) we introduce now pairs of complex conjugate coordinates (w, w^*) and (z_+, z_-) in the following way

$$w \equiv u + iv, \quad w^* \equiv u - iv, \quad u \equiv \frac{w + w^*}{2}, \quad v \equiv -i \frac{w - w^*}{2}, \quad (17.4)$$

$$z_+ \equiv x + iy, \quad z_- \equiv x - iy, \quad x \equiv \frac{z_+ + z_-}{2}, \quad y \equiv -i \frac{z_+ - z_-}{2}.$$

The Equation (17.1) for the unit sphere is then

$$z_+ z_- + z^2 = 1. \quad (17.5)$$

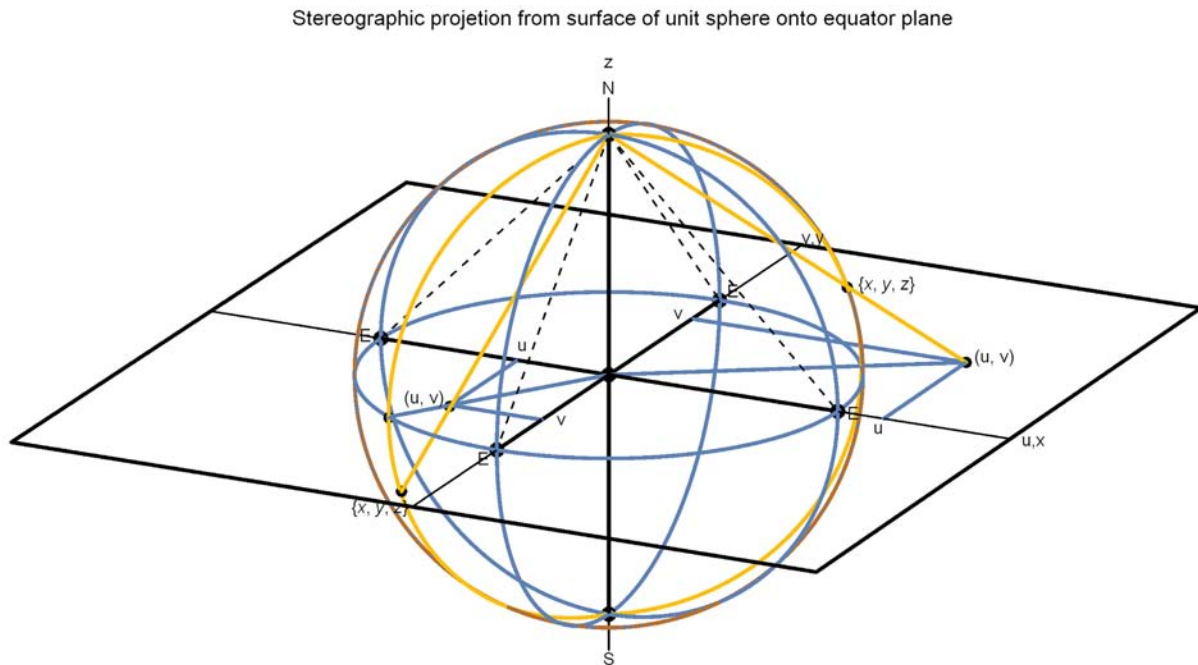


Figure 6. Stereographic projection of the surface of the unit sphere from North Pole onto the Equator plane. There are shown the projections of two points from the surface of the sphere with coordinates (x, y, z) onto the Equator plane with coordinates (u, v) , one with $z > 0$ outside to the unit disc and the second with $z < 0$ inside to the unit disc in the Equator plane. A 3D-rotation of the surface of the sphere makes a unique fractional (or Möbius) transformation of the Equator plane described by formulae in the text.

The mapping (17.2) takes on the form

$$(z_+, z_-, z) \rightarrow (w, w^*) = \left(\frac{z_+}{1-z}, \frac{z_-}{1-z} \right), \quad z = \pm \sqrt{1 - z_+ z_-}, \quad (17.6)$$

with the unique inversion

$$(w, w^*) \rightarrow (z_+, z_-, z) = \left(\frac{2w}{ww^* + 1}, \frac{2w^*}{ww^* + 1}, \frac{ww^* - 1}{ww^* + 1} \right). \quad (17.7)$$

Since in the following we use multiplication of complex numbers we can identify $\mathbb{R}^2$ with the complex plane $\mathbb{C}$ which we extend by a single point ∞ to the extended complex plane $\bar{\mathbb{C}}$. The North Pole $(z_+, z_-, z) = (0, 0, 1)$ is now mapped by the stereographic projection onto this unique new element ∞ of $\bar{\mathbb{C}}$.

We now consider a fractional linear transformation $(w, w^*) \rightarrow (w', w'^*)$ of the extended complex plane $\bar{\mathbb{C}}$ as follows

$$w' = \frac{\kappa w + \mu}{\lambda w + \nu}, \quad w'^* = \frac{\kappa^* w^* + \mu^*}{\lambda^* w^* + \nu^*}, \quad \left| \pm \begin{pmatrix} \kappa & \lambda \\ \mu & \nu \end{pmatrix} \right| = \kappa \nu - \lambda \mu = 1, \quad (17.8)$$

where $(\kappa, \lambda, \mu, \nu)$ are 3 independent, in general, complex numbers. The determinant of the coefficient matrix is equal to +1 that means unimodularity. The fractional transformations of this kind as is well known and can easily be checked form a group which as group of matrices $\begin{pmatrix} \kappa & \lambda \\ \mu & \nu \end{pmatrix}$ is the special linear group

$SL(2, \mathbb{C})$. For the transformed coordinates (z'_+, z'_-, z') using the connection (17.7) for the transformed (primed) coordinates and the transformation (17.8) we find

$$\begin{aligned} z'_+ &= \frac{2w'}{w'w'^* + 1} = \frac{2(\kappa w + \mu)(\lambda^* w^* + \nu^*)}{(\kappa w + \mu)(\kappa^* w^* + \mu^*) + (\lambda w + \nu)(\lambda^* w^* + \nu^*)}, \quad z'_- = z'^*_+, \\ z' &= \frac{w'w'^* - 1}{w'w'^* + 1} = \frac{(\kappa w + \mu)(\kappa^* w^* + \mu^*) - (\lambda w + \nu)(\lambda^* w^* + \nu^*)}{(\kappa w + \mu)(\kappa^* w^* + \mu^*) + (\lambda w + \nu)(\lambda^* w^* + \nu^*)}. \end{aligned} \quad (17.9)$$

If we now specialize the fractional linear transformation supposing

$$\nu = \kappa^*, \quad \mu = -\lambda^*, \quad \Leftrightarrow \quad \kappa = \nu^*, \quad \lambda = -\mu^*, \quad \Rightarrow \quad \begin{pmatrix} \kappa & \lambda \\ \mu & \nu \end{pmatrix} = \begin{pmatrix} \kappa & \lambda \\ -\lambda^* & \kappa^* \end{pmatrix}, \quad (17.10)$$

which specializes the transformation from a general $SL(2, \mathbb{C})$ to a $SU(2)$

matrix $\begin{pmatrix} \kappa & \lambda \\ \mu & \nu \end{pmatrix} = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}$ the relation (17.9) can be expressed

$$\begin{aligned} z'_+ &= \frac{2(-\kappa\mu(w w^* - 1) + \kappa^2 w - \mu^2 w^*)}{\underbrace{(\kappa\nu - \lambda\mu)}_{=1}(w w^* + 1)} = \kappa^2 z_+ - \mu^2 z_- - 2\kappa\mu z, \quad z'_- = z'^*_+, \\ z' &= \frac{(\kappa\nu + \lambda\mu)(w w^* - 1) - 2\kappa\lambda w + 2\mu\nu w^*}{\underbrace{(\kappa\nu - \lambda\mu)}_{=1}(w w^* + 1)} = -\kappa\lambda z_+ + \mu\nu z_- + (\kappa\nu + \lambda\mu)z. \end{aligned} \quad (17.11)$$

and using (17.7) as the following linear transformation $(z_+, z, z_-) \leftrightarrow (z'_+, z', z'_-)$

$$(z'_+, z', z'_-) = (z_+, z, z_-)R, \quad \Leftrightarrow \quad \begin{pmatrix} z'_+ \\ z' \\ z'_- \end{pmatrix} = R^T \begin{pmatrix} z_+ \\ z \\ z_- \end{pmatrix}, \quad (17.12)$$

with the (complex) rotation matrix in representation with respect to the (complex) basis (z_+, z, z_-)

$$\begin{aligned} R &\equiv \begin{pmatrix} \kappa^2 & -\kappa\lambda & -\lambda^2 \\ -2\kappa\mu & \kappa\nu + \lambda\mu & 2\lambda\nu \\ -\mu^2 & \mu\nu & \nu^2 \end{pmatrix} = \begin{pmatrix} \kappa^2 & -\kappa\lambda & -\lambda^2 \\ 2\kappa\lambda^* & \kappa\kappa^* - \lambda\lambda^* & 2\kappa^*\lambda \\ -\lambda^{*2} & -\kappa^*\lambda^* & \kappa^{*2} \end{pmatrix}, \\ |R| &= (\kappa\nu - \lambda\mu)^3 = 1, \quad \langle R \rangle = [R] = (\kappa + \nu)^2 - 1. \end{aligned} \quad (17.13)$$

with the two possible identifications distinguished by sign

$$\begin{pmatrix} \kappa & \lambda \\ \mu & \nu \end{pmatrix} = \pm \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} \equiv \pm S, \quad \nu = \kappa^*, \quad \mu = -\lambda^*, \quad (17.14)$$

this leads to the same rotation matrix R in (12.1) explicitly given in (12.2). In representation with respect to the real basis (x, y, z) we find⁹

$$(x', y', z') = (x, y, z)R, \quad \Leftrightarrow \quad \begin{pmatrix} x' \\ y' \\ z' \end{pmatrix} = R^T \begin{pmatrix} x \\ y \\ z \end{pmatrix}, \quad (17.15)$$

⁹We consider R as operator and do not distinguish the notation for R in representations with respect to different basis vectors; see also footnote before (12.3).

with the explicit form of the real matrix R and the same invariant as in (17.13)

$$R = \begin{pmatrix} \frac{1}{2}(\kappa^2 - \lambda^2 - \mu^2 + \nu^2) & -\frac{i}{2}(\kappa^2 + \lambda^2 - \mu^2 - \nu^2) & -\kappa\lambda + \mu\nu \\ \frac{i}{2}(\kappa^2 - \lambda^2 + \mu^2 - \nu^2) & \frac{1}{2}(\kappa^2 + \lambda^2 + \mu^2 + \nu^2) & -i(\kappa\lambda + \mu\nu) \\ -\kappa\mu + \lambda\nu & i(\kappa\mu + \lambda\nu) & \kappa\nu + \lambda\mu \end{pmatrix}, \quad (17.16)$$

$$|R| = (\kappa\nu - \lambda\mu)^3 = 1, \quad \langle R \rangle = [R] = (\kappa + \nu)^2 - 1.$$

Thus a 3D rotation of the surface of a unit sphere S^3 corresponds in unique way to a fractional linear (Möbius) transformations of the extended complex plane $\bar{\mathbb{C}}$ of the form (17.8) equivalent to two different (by sign) unitary unimodular transformations of $SU(2)$. Another special linear fractional transformation is mentioned by Gürsey [21] (around and after Equation (4.27) there).

The mappings in the above considerations are made according to the following scheme

$$\begin{array}{ccc} (x, y, z), \quad (x^2 + y^2 + z^2 = 1) & \leftrightarrow & (u, v), \quad (w, w^*) \equiv (u + iv, u - iv) \\ \uparrow & \text{stereogr. projection} & \uparrow \\ \text{3D rotation (17.15) with (17.13)} & & \text{fract. linear transf. (17.8) with (17.10) (17.17)} \\ \downarrow & \text{stereogr. projection} & \downarrow \\ (x', y', z'), \quad (x'^2 + y'^2 + z'^2 = 1) & \leftrightarrow & (u', v'), \quad (w', w'^*) \equiv (u' + iv', u' - iv'). \end{array}$$

The general Möbius transformations studied in the complex analysis are much richer in content than touched here (e.g., Needham [38] with a great number of wonderful illustrations).

18. Conclusion

This article collects some basic mathematical aspects of the Lie group $SU(2)$ embedded in higher Lie groups such as $Sp(4, \mathbb{R})$ and inhomogeneous Lie group $ISU(2)$ and starts from the middle of the nineteens with collecting material. It was after working in applications of the group $SU(1,1)$ in quantum optics and with the aim to treat such problems as light polarization of beams in reflection and refraction and to discuss the forms of polarized and unpolarized light. It works mainly with coordinate-invariant methods using with some advantage, the vector parameter φ instead of Euler angles (α, β, γ) and with advantage, the vector parameter c for $SO(3, \mathbb{R})$ of F.I. Fyodorov from Minsk. In this way some difficulties were met, for example, of the root diagrams for inhomogeneous groups or in the result for the invariant measures of $SO(3, \mathbb{R})$ which I tried to overcome or to represent as problem (**Appendix C**). A continuation of the article with more aspects of physical applications to polarization of light beams is possible and it is hoped to become realized.

Conflicts of Interest

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

References

- [1] Rosenfel'd, B.A. (1966) *Mnogomernyye prostranstva* (Multi-Dimensional Spaces). Nauka, Moskva.
- [2] Landau L.D. and Lifshitz, E.M. (1988) *Klassicheskaya teoriya polja* (Classical Field Theory). 7th Edition, Nauka, Moskva.
- [3] Landau, L.D. and Lifshitz, E.M. (1987) *Quantum Mechanics*. Pergamon Press, New York.
- [4] Fyodorov, F.I. (1958) *Optika anizotropnykh sred* (Optics of Anisotropic Media). Isdatel'stvo AN BSSR, Minsk.
- [5] Fyodorov, F.I. (1976) *Teoria girotropii* (Theory of Gyrotropy). Isdatel'stvo Nauka i tekhnika, Minsk.
- [6] Fyodorov, F.I. (1979) *Gruppa Lorentsa*. Nauka, Moskva.
- [7] Gantmacher, F.R. (1959) *Matritzenrechnung*, Band I und II. Deutscher Verlag der Wissenschaften, Berlin.
- [8] Lagally, M. (1956) *Vorlesungen über Vektorrechnung*. (5. Auflage). Akademische Verlagsgesellschaft Geest & Portig, Leipzig.
- [9] Jordan, P. (1935) *Zeitschrift für Physik*, **94**, 531.
<https://doi.org/10.1007/BF01330618>
- [10] Schwinger, J. (1952) On Angular Momentum. U.S. Atomic Energy Commission, NYO-3071. (Internally Published), Reproduced in Biedenharn, and Van Dam.
- [11] Biedenharn, L.C. and Van Dam, H. (1965) *Quantum Theory of Angular Momentum* (a Collection of Reprints and Original Papers). Academic Press, New York.
- [12] Jauch, J.M. and Rohrlich, F. (1955) *The Theory of Photons and Electrons*. Springer-Verlag, Berlin. https://doi.org/10.1007/978-3-642-80951-4_1
- [13] Peřina, J. (1971) *Coherence of Light*. Van Nostrand Reinhold Company, New York.
- [14] Prakash, H. and Chandra, N. (1971) *Physical Review A*, **4**, 796.
<https://doi.org/10.1103/PhysRevA.4.796>
- [15] Agarwal, G.S. (1971) *Lettere Al Nuovo Cimento*, **1**, 53.
<https://doi.org/10.1007/BF02774060>
- [16] Luis, A. and Sánchez-Soto, L.L. (2000) Quantum Phase Difference, Phase Measurements and Stokes Operators. In: Wolf, E., Ed., *Progress in Optics*, Vol. 41, Elsevier, Amsterdam, 421-481. [https://doi.org/10.1016/S0079-6638\(00\)80021-9](https://doi.org/10.1016/S0079-6638(00)80021-9)
- [17] Peřinová, V., Lukš, A. and Peřina, J. (1988) *Phase in Optics*. World Scientific, Singapore. <https://doi.org/10.1142/3541>
- [18] Biedenharn, L.C. and Louck, J.D. (1981) *Angular Momentum in Quantum Physics*. Addison Wesley, Reading.
- [19] Messiah, A. (1962) *Quantum Mechanics*, Vol. II. North-Holland Publishing, Amsterdam.
- [20] Lyubarskij, G.Ya. (1958) *Teoriya grupp i ee primeneniya v fizike*. Fizmatgiz, Moskva.
- [21] Gürsey, F. (1964) Introduction to Group Theory. In: DeWitt, C. and DeWitt, B., Eds., *Relativity, Groups and Topology*, Gordon and Breach Science Publishers, New York, 89-161.
- [22] Behrends, R.E., Dreitlein, J., Fronsdal, C. and Lee, W. (1962) *Reviews of Modern Physics*, **34**, 1. <https://doi.org/10.1103/RevModPhys.34.1>
- [23] Wybourne, B. (1974) *Classical Groups for Physicists*. John Wiley, New York.

- [24] Gilmore, R. (1974) Lie Groups, Lie Algebras, and Some of Their Applications. John Wiley and Sons, New York. <https://doi.org/10.1063/1.3128987>
- [25] Barut, A.O. and Rączka, R. (1977) Theory of Group Representations and Applications. PWN-Polish Scientific Publishers, Warszawa.
- [26] Hamermesh, M. (1964) Group Theory and Its Application to Physical Problems. Addison-Wesley, Reading.
- [27] Wünsche, A. (2001) *Journal of Optics B: Quantum and Semiclassical Optics*, **3**, 6-15. <https://doi.org/10.1088/1464-4266/3/1/302>
- [28] Wünsche, A. (2003) Squeezed States. In: Dodonov, V.V. and Man'ko, V.I., Eds., *Theory of Nonclassical States of Light*, Taylor and Francis, London, 95-152.
- [29] Ban, M. (1993) *Journal of the Optical Society of America B*, **10**, 1347-1359. <https://doi.org/10.1364/JOSAB.10.001347>
- [30] Wei, J. and Norman, E. (1963) *Journal of Mathematical Physics*, **4**, 575. <https://doi.org/10.1063/1.1703993>
- [31] Dattoli, G., Richetta, M. and Torre, A. (1988) *Physical Review A*, **37**, 2007. <https://doi.org/10.1103/PhysRevA.37.2007>
- [32] Perelomov, A.M. (1986) Generalized Coherent States and Their Application. Springer-Verlag, Berlin. <https://doi.org/10.1007/978-3-642-61629-7>
- [33] Altmann, S.L. (1986) Rotations, Quaternions, and Double Groups. Clarendon Press, Oxford.
- [34] Koecher, M. and Remmert, R. (1991) Hamilton's Quaternions. In: Ebbinghaus, H.-D., et al., Eds., *Numbers*, Springer, New York, 189-220. https://doi.org/10.1007/978-1-4612-1005-4_10
- [35] Gürlebeck, K. and Sprössing, W. (1997) Quaternionic and Clifford Calculus for Physicists and Engineers. John Wiley, New York.
- [36] Bogoliubov, N.N., Logunov, A.A. and Todorov, I.T. (1969) Introduction to Axiomatic Field Theory. Benjamin/Cummings, Reading.
- [37] Pars, L.A. (1965) A Treatise on Analytical Dynamics. John Wiley & Sons, New York.
- [38] Needham, T. (1997) Visual Complex Analysis. Clarendon Press, Oxford.

Appendix A. Relation between Vector Parameter $\boldsymbol{\varphi}$ and Euler Angles (α, β, γ) for Group $SU(2)$

Using the vector parameter $\boldsymbol{\varphi}$, we obtained in (5.13) the matrix of the fundamental representation of $SU(2)$ in the form

$$S = \begin{pmatrix} \cos\left(\frac{|\boldsymbol{\varphi}|}{2}\right) + i\frac{\varphi_3}{|\boldsymbol{\varphi}|}\sin\left(\frac{|\boldsymbol{\varphi}|}{2}\right) & i\frac{\varphi_-}{|\boldsymbol{\varphi}|}\sin\left(\frac{|\boldsymbol{\varphi}|}{2}\right) \\ i\frac{\varphi_+}{|\boldsymbol{\varphi}|}\sin\left(\frac{|\boldsymbol{\varphi}|}{2}\right) & \cos\left(\frac{|\boldsymbol{\varphi}|}{2}\right) - i\frac{\varphi_3}{|\boldsymbol{\varphi}|}\sin\left(\frac{|\boldsymbol{\varphi}|}{2}\right) \end{pmatrix} \quad (\text{A.1})$$

$$\equiv \begin{pmatrix} \sqrt{T}e^{i\theta} & \sqrt{R}e^{-i\chi} \\ -\sqrt{R}e^{i\chi} & \sqrt{T}e^{-i\theta} \end{pmatrix}, \quad |\boldsymbol{\varphi}| \equiv \sqrt{\varphi_1^2 + \varphi_2^2 + \varphi_3^2} = \sqrt{\varphi_- \varphi_+ + \varphi_3^2},$$

and with abbreviations (T for transmission and R for reflection in application to reflection problems)

$$T \equiv \cos^2\left(\frac{|\boldsymbol{\varphi}|}{2}\right) + \frac{\varphi_3^2}{|\boldsymbol{\varphi}|^2} \sin^2\left(\frac{|\boldsymbol{\varphi}|}{2}\right) = 1 - \frac{\varphi_- \varphi_+}{|\boldsymbol{\varphi}|^2} \sin^2\left(\frac{|\boldsymbol{\varphi}|}{2}\right), \quad (\text{A.2})$$

$$R \equiv \frac{\varphi_- \varphi_+}{|\boldsymbol{\varphi}|^2} \sin^2\left(\frac{|\boldsymbol{\varphi}|}{2}\right) = \left(1 - \frac{\varphi_3^2}{|\boldsymbol{\varphi}|^2}\right) \sin^2\left(\frac{|\boldsymbol{\varphi}|}{2}\right), \quad T + R = 1.$$

On the other side, using the Euler angles (α, β, γ) , one obtains representations of the form [3] [26]

$$S = \begin{pmatrix} \exp\left(i\frac{\alpha+\gamma}{2}\right)\cos\left(\frac{\beta}{2}\right) & e^{i\delta_0}\exp\left(-i\frac{\alpha-\gamma}{2}\right)\sin\left(\frac{\beta}{2}\right) \\ -e^{-i\delta_0}\exp\left(i\frac{\alpha-\gamma}{2}\right)\sin\left(\frac{\beta}{2}\right) & \exp\left(-i\frac{\alpha+\gamma}{2}\right)\cos\left(\frac{\beta}{2}\right) \end{pmatrix} \quad (\text{A.3})$$

$$= \begin{pmatrix} \exp\left(i\frac{\gamma}{2}\right) & 0 \\ 0 & \exp\left(-i\frac{\gamma}{2}\right) \end{pmatrix} \begin{pmatrix} \cos\left(\frac{\beta}{2}\right) & e^{i\delta_0}\sin\left(\frac{\beta}{2}\right) \\ -e^{-i\delta_0}\sin\left(\frac{\beta}{2}\right) & \cos\left(\frac{\beta}{2}\right) \end{pmatrix}.$$

The fixed phase δ_0 which can be chosen arbitrarily is introduced here to bring in correspondence representations by different authors, for example, $\delta_0 = 0$ by Biedenharn and Louck [18] (here $(\alpha, \beta, \gamma) \rightarrow (-\alpha, -\beta, -\gamma)$), by Landau and Lifshits [3] (Equation (58.6)), by Hamermesh [26] (Equation (9.65)) and $\delta_0 = -\frac{\pi}{2}$ by Lyubarski [20] (Equation (53.3)).

With $\varphi_{\pm} \equiv \varphi_1 \pm i\varphi_2$, we find from (A.1) the inversion

$$\varphi_1 = -i(S_{12} + S_{21})\frac{\arcsin(\theta)}{\theta}, \quad \varphi_2 = (S_{12} - S_{21})\frac{\arcsin(\theta)}{\theta}, \quad (\text{A.4})$$

$$\varphi_3 = -i(S_{11} - S_{22})\frac{\arcsin(\theta)}{\theta}, \quad \theta \equiv \sqrt{1 - \left(\frac{S_{11} + S_{22}}{2}\right)^2} \equiv \sin\left(\frac{|\boldsymbol{\varphi}|}{2}\right).$$

Inserting the matrix elements according to (A.1), we get the following relation of $\boldsymbol{\varphi} = (\varphi_1, \varphi_2, \varphi_3)$ to the Euler angles (α, β, γ)

$$\begin{aligned}
\varphi_1 &= -2 \sin\left(\frac{\alpha - \gamma - \delta_0}{2}\right) \sin\left(\frac{\beta}{2}\right) \frac{\arcsin(\theta)}{\theta}, \\
\varphi_2 &= -2 \cos\left(\frac{\alpha - \gamma - \delta_0}{2}\right) \sin\left(\frac{\beta}{2}\right) \frac{\arcsin(\theta)}{\theta}, \\
\varphi_3 &= 2 \sin\left(\frac{\alpha + \gamma}{2}\right) \cos\left(\frac{\beta}{2}\right) \frac{\arcsin(\theta)}{\theta}, \\
|\boldsymbol{\varphi}| &= 2 \left| \arccos \left\{ \cos\left(\frac{\alpha + \gamma}{2}\right) \cos\left(\frac{\beta}{2}\right) \right\} \right|,
\end{aligned} \tag{A.5}$$

with θ expressed by the Euler angles

$$\theta \equiv \sqrt{1 - \cos^2\left(\frac{\alpha + \gamma}{2}\right) \cos^2\left(\frac{\beta}{2}\right)}. \tag{A.6}$$

For the inversion of these relations, we first make the inversion of (A.1) that provides

$$\exp(i\alpha) = \sqrt{-e^{i2\delta_0} \frac{S_{11}S_{21}}{S_{22}S_{12}}}, \quad \sin\left(\frac{\beta}{2}\right) = \sqrt{-S_{12}S_{21}}, \quad \exp(i\gamma) = \sqrt{-e^{-i2\delta_0} \frac{S_{11}S_{12}}{S_{22}S_{21}}}. \tag{A.7}$$

Inserting here the matrix elements (A.1), we find the relations

$$\begin{aligned}
\alpha &= \arctg\left(\frac{\varphi_3}{|\boldsymbol{\varphi}|} \operatorname{tg}\left(\frac{|\boldsymbol{\varphi}|}{2}\right)\right) + \arctg\left(\frac{\varphi_2}{\varphi_1}\right) + \delta_0 \mp \frac{\pi}{2}, \\
\beta &= 2 \arcsin\left(\frac{\sqrt{\varphi_1^2 + \varphi_2^2}}{|\boldsymbol{\varphi}|} \sin\left(\frac{|\boldsymbol{\varphi}|}{2}\right)\right), \\
\gamma &= \arctg\left(\frac{\varphi_3}{|\boldsymbol{\varphi}|} \operatorname{tg}\left(\frac{|\boldsymbol{\varphi}|}{2}\right)\right) - \arctg\left(\frac{\varphi_2}{\varphi_1}\right) - \delta_0 \pm \frac{\pi}{2}.
\end{aligned} \tag{A.8}$$

This shows that the relations between the vector parameter $\boldsymbol{\varphi} = (\varphi_1, \varphi_2, \varphi_3)$ and the Euler angles (α, β, γ) are not very simple and by no means evident and well suited for the calculation of the invariant measure in the vector parameter $\boldsymbol{\varphi}$.

The main disadvantage of the Euler angles (α, β, γ) in comparison to the vector parameter $(\varphi_1, \varphi_2, \varphi_3)$ is that they are not well appropriate for coordinate-invariant considerations and, therefore, for emphasizing invariance properties of the system. An advantage of the Euler angles in comparison to the discussed vector parameter is that the transformation matrices are easily factorized in these variables, whereas the disentanglement of the representation matrices with regard to different components of the vector parameter $(\varphi_1, \varphi_2, \varphi_3)$ leads to more complicated but important relations (see Section 10).

Appendix B. A Curious Problem with Divergence of Traces of Commutators in Lie Algebras

In the Lie algebra the commutator of two operators $[X, Y] \equiv XY - YX$ takes on the role of the product of operators. The trace of an arbitrary operator Z we denote by $\langle Z \rangle$. Then we have

$$[X, Y] \equiv XY - YX, \Rightarrow \langle [X, Y] \rangle = \langle XY \rangle - \langle YX \rangle = 0. \quad (\text{B.1})$$

However, in the special case of a pair of annihilation and creation operator $X = a, Y = a^\dagger$ we find a non-vanishing, even infinite trace

$$[a, a^\dagger] = I, \Rightarrow \langle [a, a^\dagger] \rangle = \langle I \rangle \rightarrow +\infty. \quad (\text{B.2})$$

One may think that a possible reason for this is that $\langle AB \rangle = \langle BA \rangle$ is no more correct for the case of infinite matrices or operators due to problems of divergences.

The identity operator I can be resolved to the infinite sum over number-state operator $|n\rangle\langle n|$ with

$$I = \sum_{n=0}^{\infty} |n\rangle\langle n|, \quad \langle m|n\rangle = \delta_{m,n}, \quad a|n\rangle = \sqrt{n}|n-1\rangle, \quad a^\dagger|n\rangle = \sqrt{n+1}|n+1\rangle. \quad (\text{B.3})$$

Using this we find for the trace of $[a, a^\dagger]$

$$\langle [a, a^\dagger] \rangle = \langle I \rangle = \sum_{n=0}^{\infty} \langle n|I|n\rangle = \sum_{n=0}^{\infty} 1 \rightarrow +\infty. \quad (\text{B.4})$$

On the other side forming the trace over $aa^\dagger$ and $a^\dagger a$ separately

$$\begin{aligned} \langle aa^\dagger \rangle &= \sum_{n=0}^{\infty} \langle n|aa^\dagger|n\rangle = \sum_{n=0}^{\infty} (n+1)\langle n+1|n+1\rangle = \sum_{n=0}^{\infty} (n+1) \rightarrow +\infty, \\ \langle a^\dagger a \rangle &= \sum_{n=0}^{\infty} \langle n|a^\dagger a|n\rangle = \sum_{n=1}^{\infty} n\langle n-1|n-1\rangle = \sum_{n=1}^{\infty} n = \sum_{n=0}^{\infty} (n+1) \rightarrow +\infty, \end{aligned} \quad (\text{B.5})$$

Thus we obtained in this calculation formally the same trace over $aa^\dagger$ and $a^\dagger a$

$$\langle aa^\dagger \rangle = \langle a^\dagger a \rangle, \Rightarrow \langle [a, a^\dagger] \rangle = 0. \quad (\text{B.6})$$

This is the result which we would expect since for finite operators A and B we have in every case $\langle AB \rangle = \langle BA \rangle$.

It is clear that the contradiction results from calculations with divergent sums and by reordering the positive and negative sum terms in different way we may find different results. There remains the problem which of the calculations for expectation values of commutators is the correct one if needed and why. Due to $\langle [A, B] \rangle = 0$ in “general case” we tend to $\langle [a, a^\dagger] \rangle = 0$ in “most” cases.

Appendix C. Transformation of Parameter c for $SO(3, \mathbb{R})$ and Limiting Case

If we make in (16.4) the substitutions

$$c \equiv \frac{x}{R}, \quad c_0 \equiv \frac{x_0}{R}, \quad c' \equiv \frac{x'}{R}, \quad (\text{C.1})$$

then we find

$$\frac{x'}{R} = c' = \frac{c + c_0 + [c, c_0]}{1 - cc_0} = \frac{\frac{x}{R} + \frac{x_0}{R} + \frac{[x, x_0]}{R^2}}{1 - \frac{xx_0}{R^2}}, \Rightarrow x' = \frac{x + x_0 + \frac{[x, x_0]}{R}}{1 - \frac{xx_0}{R^2}}. \quad (\text{C.2})$$

In the limit $R \rightarrow \infty$

$$R \rightarrow \infty: \quad \mathbf{x}' \rightarrow \mathbf{x} + \mathbf{x}_0, \quad (\text{C.3})$$

it makes the transition to a translation of an arbitrary vector $\mathbf{x}$ to a new vector $\mathbf{x}' = \mathbf{x} + \mathbf{x}_0$ with fixed $\mathbf{x}_0$.

Appendix D. Invariants and Complementary Operator to a Special Three-Dimensional Operator $\mathbf{A} = \alpha \mathbf{I} + \beta \mathbf{b} \cdot \tilde{\mathbf{b}} + \gamma [\mathbf{c}]$

We consider the following special three-dimensional operator $\mathbf{A}$ in three-dimensional Euclidean space ($g_{ij} = \delta_{ij}$) which is a little more generally as needed in Section 16 but suited also for many other applications

$$\mathbf{A} = \alpha \mathbf{I} + \beta \mathbf{b} \cdot \tilde{\mathbf{b}} + \gamma [\mathbf{c}], \quad (\text{D.1})$$

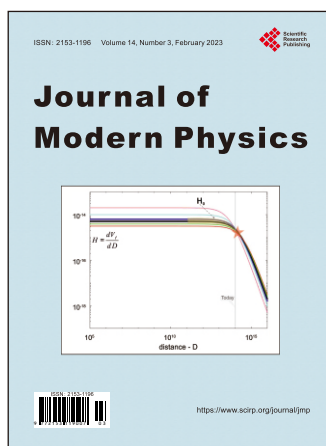
where $[\mathbf{c}]$ is an antisymmetric operator to vector $\mathbf{c}$ (see 10) and calculate its invariants and the complementary operator. According to (7) we have first to calculate its powers $\mathbf{A}^2$ and $\mathbf{A}^3$ and then to form their traces that for length we do not write down. Then with (7) we can find for their invariants

$$\begin{aligned} \langle \mathbf{A} \rangle &= 3\alpha + \beta \tilde{\mathbf{b}} \mathbf{b}, \\ [\mathbf{A}] &= 3\alpha^2 + 2\alpha\beta \tilde{\mathbf{b}} \mathbf{b} + \beta\gamma [\tilde{\mathbf{b}}, \mathbf{b}, \mathbf{c}] + \gamma^2 \mathbf{c}^2, \\ |\mathbf{A}| &= \alpha^2 (\alpha + \beta (\tilde{\mathbf{b}} \mathbf{b})) + \alpha\beta\gamma [\tilde{\mathbf{b}}, \mathbf{b}, \mathbf{c}] + \alpha\gamma^2 \mathbf{c}^2 + \beta\gamma^2 (\tilde{\mathbf{b}} \mathbf{c})(\mathbf{c} \mathbf{b}) \\ &= (\alpha + \beta \tilde{\mathbf{b}} \mathbf{b})(\alpha^2 + \gamma^2 \mathbf{c}^2) + \alpha\beta\gamma [\tilde{\mathbf{b}}, \mathbf{b}, \mathbf{c}] + \beta\gamma^2 [\tilde{\mathbf{b}}, \mathbf{c}][\mathbf{c}, \mathbf{b}]. \end{aligned} \quad (\text{D.2})$$

The complementary operator $\bar{\mathbf{A}} \equiv [\mathbf{A}]\mathbf{I} - \langle \mathbf{A} \rangle \mathbf{A} + \mathbf{A}^2$ to $\mathbf{A}$ possesses the following representation

$$\begin{aligned} \bar{\mathbf{A}} &= \left\{ \alpha (\alpha + \beta \tilde{\mathbf{b}} \mathbf{b}) + \beta\gamma [\tilde{\mathbf{b}}, \mathbf{b}, \mathbf{c}] \right\} \mathbf{I} - \alpha\beta \mathbf{b} \cdot \tilde{\mathbf{b}} + \gamma^2 \mathbf{c} \cdot \mathbf{c} - (\alpha + \beta \tilde{\mathbf{b}} \mathbf{b})\gamma [\mathbf{c}] \\ &\quad + \beta\gamma (\mathbf{b} \cdot [\tilde{\mathbf{b}}, \mathbf{c}] + [\mathbf{c}, \mathbf{b}] \cdot \tilde{\mathbf{b}}), \quad \langle \bar{\mathbf{A}} \rangle = [\mathbf{A}], \quad \mathbf{A}^{-1} = \frac{\bar{\mathbf{A}}}{|\mathbf{A}|}. \end{aligned} \quad (\text{D.3})$$

One may consider many special cases of these formulae.



Call for Papers

Journal of Modern Physics

ISSN: 2153-1196 (Print) ISSN: 2153-120X (Online)
<https://www.scirp.org/journal/jmp>

Journal of Modern Physics (JMP) is an international journal dedicated to the latest advancement of modern physics. The goal of this journal is to provide a platform for scientists and academicians all over the world to promote, share, and discuss various new issues and developments in different areas of modern physics.

Subject Coverage

Journal of Modern Physics publishes original papers including but not limited to the following fields:

Biophysics and Medical Physics
 Complex Systems Physics
 Computational Physics
 Condensed Matter Physics
 Cosmology and Early Universe
 Earth and Planetary Sciences
 General Relativity
 High Energy Astrophysics
 High Energy/Accelerator Physics
 Instrumentation and Measurement
 Interdisciplinary Physics
 Materials Sciences and Technology
 Mathematical Physics
 Mechanical Response of Solids and Structures

New Materials: Micro and Nano-Mechanics and Homogeneization
 Non-Equilibrium Thermodynamics and Statistical Mechanics
 Nuclear Science and Engineering
 Optics
 Physics of Nanostructures
 Plasma Physics
 Quantum Mechanical Developments
 Quantum Theory
 Relativistic Astrophysics
 String Theory
 Superconducting Physics
 Theoretical High Energy Physics
 Thermology

We are also interested in: 1) Short Reports—2-5 page papers where an author can either present an idea with theoretical background but has not yet completed the research needed for a complete paper or preliminary data; 2) Book Reviews—Comments and critiques.

Notes for Intending Authors

Submitted papers should not have been previously published nor be currently under consideration for publication elsewhere. Paper submission will be handled electronically through the website. All papers are refereed through a peer review process. For more details about the submissions, please access the website.

Website and E-Mail

<https://www.scirp.org/journal/jmp> E-mail: jmp@scirp.org

What is SCIRP?

Scientific Research Publishing (SCIRP) is one of the largest Open Access journal publishers. It is currently publishing more than 200 open access, online, peer-reviewed journals covering a wide range of academic disciplines. SCIRP serves the worldwide academic communities and contributes to the progress and application of science with its publication.

What is Open Access?

All original research papers published by SCIRP are made freely and permanently accessible online immediately upon publication. To be able to provide open access journals, SCIRP defrays operation costs from authors and subscription charges only for its printed version. Open access publishing allows an immediate, worldwide, barrier-free, open access to the full text of research papers, which is in the best interests of the scientific community.

- High visibility for maximum global exposure with open access publishing model
- Rigorous peer review of research papers
- Prompt faster publication with less cost
- Guaranteed targeted, multidisciplinary audience



**Scientific
Research
Publishing**

Website: <https://www.scirp.org>
Subscription: sub@scirp.org
Advertisement: service@scirp.org